

Companion Proceeding of the 9th
International Conference on
Learning Analytics & Knowledge
(LAK'19)



March 4-8, 2019, Tempe, Arizona, USA

<https://lak19.solaresearch.org>



This work is published under the terms of the Creative Commons Attribution- Noncommercial-ShareAlike 3.0 Australia Licence. Under this Licence you are free to:



Share — copy and redistribute the material in any medium or format
The licensor cannot revoke these freedoms as long as you follow the license terms.

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Non-Commercial — You may not use the material for commercial purposes.

NoDerivatives — If you remix, transform, or build upon the material, you may not distribute the modified material.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Program Committee

Ani Aghababayan
Cecilia Aguerrebere
Stephen Aguilar
Vincent Aleven

Laura Allen
Kimberly Arnold
Roger Azevedo
Ryan Baker
Aneesha Bakharia
Tiffany Barnes
Alan Berg
Yoav Bergner
Marie Bienkowski
Robert Bodily
Anthony F. Botelho
Christopher Brooks
Michael Brown
Simon Buckingham Shum
C Sean Burns
Carol Calvert
Sven Charleer
Mohamed Amine Chatti
George Chen
Guanliang Chen
Weiqin Chen
Katherine Chiluiza

Yong-Sang Cho
Irene Chounta
Cassandra Colvin
Adam Cooper
Linda Corrin
Scott Crossley
Yi Cui
Mutlu Cukurova
Sidney D'Mello
Dan Davis
Daniel Davis
Shane Dawson
Paula de Barba
Carrie Demmans Epp
Matt Demonbrun
Stefan Dietze
Pierre Dillenbourg
Yannis Dimitriadis
Nia Dowell
Hendrik Drachsler
Michael Eagle
Martin Ebner
Vanessa Echeverria

Tanya Elias
Alfred Essa
Stephen Fancsali
Rebecca Ferguson

McGraw-Hill Education
Duke University
University of Southern California, Rossier School of Education
Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA
Mississippi State University
University of Wisconsin-Madison
Simon Fraser University
University of Pennsylvania
The University of Queensland
North Carolina State University
University of Amsterdam
New York University
SRI International
Brigham Young University
Worcester Polytechnic Institute
School of Information, University of Michigan
Iowa State University
University of Technology, Sydney
University of Kentucky

Katholieke Universiteit Leuven
University of Duisburg-Essen
UNE Business School, University of New England
Delft University of Technology
Oslo and Akershus University College of Applied Sciences
Escuela Superior Politécnica del Litoral / Information Technology Center
i-Scream edu
University of Tartu
Charles Sturt University
Tribal
The University of Melbourne
Georgia State University
Simon Fraser University
University College London
University of Colorado Boulder
Delft University of Technology
Delft University of Technology
University of South Australia
The University of Melbourne
University of Alberta
University of Michigan
GESIS - Leibniz Institute for the Social Sciences
Ecole Polytechnique Fédérale de Lausanne
University of Valladolid
Institute for Intelligent Systems, The University of Memphis
The Open University
Carnegie Mellon University
Graz University of Technology
Connected Intelligence Centre, University of Technology Sydney, Australia
Athabasca University
McGraw Hill Education
Carnegie Learning, Inc.
The Open University

James Folkestad	Colorado State University
Albrecht Fortenbacher	HTW Berlin
Joshua Gardner	University of Michigan
Dragan Gasevic	Monash University
Michail Giannakos	Norwegian University of Science and Technology
Andrew Gibson	University of Technology, Sydney
Sean Goggins	University of Missouri
Sabine Graf	Athabasca University
Wolfgang Greller	Vienna University of Education
David Griffiths	Institute for Educational Cybernetics, University of Bolton
Dennis Groth	Indiana University Bloomington
Gahgene Gweon	Seoul National University
Andreas Harrer	University of Applied Sciences and Arts Dortmund
Scott Harrison	German Institute for International Educational Research
Marek Hatala	Simon Fraser University
Claudia Hauff	Delft University of Technology
Caroline Haythornthwaite	Syracuse University
Tobias Hecking	University of Duisburg-Essen
Neil Heffernan	Worcester Polytechnic Institute
Eelco Herder	Radboud University
Davinia Hernandez-Leo	Universitat Pompeu Fabra, Barcelona
Christothea Herodotou	The Open University
Daniel Hickey	Indiana University Bloomington
Martin Hlosta	The Open University
Tore Hoel	Høgskolen i Oslo og Akershus
Caitlin Holman	University of Michigan
Kenneth Holstein	Carnegie Mellon University
Adrian Holzer	Ecole Polytechnique Fédérale de Lausanne
Helen Hong	Nanyang Technological University, National Institute of Education
Ulrich Hoppe	University of Duisburg-Essen
Geert-Jan Houben	Delft University of Technology
Sharon Hsiao	Arizona State University
Xiao Hu	The University of Hong Kong
Stian Håklev	University of Toronto
Yang Jiang	Columbia University
John Johnston	University of Michigan
Srecko Joksimovic	Teaching Innovation Unit and School of Education, University of South Australia
Jelena Jovanovic	University of Belgrade
Sanna Järvelä	University of Oulu
Hassan Khosravi	The University of Queensland
Gábor Kismihók	Leibniz Information Centre for Science and Technology
Kirsty Kitto	University of Technology, Sydney
Rene Kizilcec	Cornell University
Ralf Klamma	RWTH Aachen University
Simon Knight	University of Technology, Sydney
Vitomir Kovanovic	University of South Australia
Andy Krumm	Digital Promise
Tanja Käser	AAA Lab, Graduate School of Education, Stanford University
Charles Lang	Columbia University
Eitel Lauria	Marist College
Nguyen-Thinh Le	Humboldt Universität zu Berlin
Stella Lee	Paradox Learning, Inc.
Chi-Un Lei	The University of Hong Kong
James Lester	North Carolina State University
Tobias Ley	Tallinn University
Warren Li	University of Michigan
Ming Liu	University of Technology, Sydney

Lori Lockyer	University of Technology, Sydney
Jason Lodge	The University of Queensland
Steven Lonn	University of Michigan
Rose Luckin	University College London
Vanda Luengo	Laboratoire d'informatique de Paris, LIP6, Sorbonne Université
Leah Macfadyen	The University of British Columbia
Katherine Maillet	Institut Mines-Télécom, Télécom Ecole de Management
Jocelyn Manderveld	SURFnet
Roberto Martinez-Maldonado	University of Technology, Sydney
Alejandra Martínez-Monés	Universidad de Valladolid, Spain
Timothy McKay	University of Michigan
Bruce McLaren	Carnegie Mellon University
Agathe Merceron	Beuth University of Applied Sciences Berlin
Negin Mirriahi	University of South Australia
Inge Molenaar	Radboud University
Inger Molenaar	Radboud University
Yishay Mor	Levinsky College of Education
Kousuke Mouri	Tokyo University of Agriculture and Technology
Pablo Munguia	RMIT University
Pedro J. Muñoz-Merino	Universidad Carlos III de Madrid
Le Quan Nguyen	The Open University
Katja Niemann	XING SE
Rachel Niemer	University of Michigan
Xavier Ochoa	Escuela Superior Politécnica del Litoral
Hiroaki Ogata	Kyoto University
Fumiya Okubo	Takachiho University
Jun Oshima	Shizuoka University
Elizabeth Owen	University of Wisconsin-Madison
Zacharoula Papamitsiou	Norwegian University of Science and Technology
Luc Paquette	University of Illinois at Urbana-Champaign
Abelardo Pardo	University of South Australia
Zach Pardos	University of California, Berkeley
Drew Paulin	University of California, Berkeley
Radek Pelánek	Masaryk University Brno
Niels Pinkwart	Humboldt-Universität zu Berlin
Matthew Pistilli	Iowa State University
Oleksandra Poquet	University of South Australia
Luis P. Prieto	School of Educational Sciences, Tallinn University (Estonia)
Paul Prinsloo	University of South Africa
Rebecca Quintana	University of Michigan
George Rehrey	IUB
Justin Reich	Massachusetts Institute of Technology
Christoph Rensing	TU Darmstadt
Bart Rienties	The Open University
Ulla Ringtved	UCN, University College of Northern Denmark
Steven Ritter	Carnegie Learning, Inc.
María Jesús Rodríguez-Triana	Tallinn University
Tim Rogers	University of South Australia
Ido Roll	The University of British Columbia
Cristobal Romero	Department of Computer Sciences and Numerical Analysis
Carolyn Rose	Carnegie Mellon University
Demetrios Sampson	Curtin University
Maria Ofelia San Pedro	ACT, Inc.
Sweet San Pedro	ACT Inc.
Agnes Sandor	Naver Labs Europe
Anna Saranti	Graz University of Technology
Maren Scheffel	Open Universiteit
Bertrand Schneider	Harvard University

Ulrik Schroeder	RWTH Aachen University
Michael Sharkey	Data & Graphs
Kshitij Sharma	Norwegian University of Science and Technology
Linda Shepard	Indiana University Bloomington
Antonette Shibani	University of Technology, Sydney
Atsushi Shimada	Kyushu University
Sanam Shirazi Beheshtiha	The University of British Columbia
Mina Shirvani Boroujeni	École polytechnique fédérale de Lausanne (EPFL)
George Siemens	UT Arlington
Sharon Slade	The Open University
Erica Snow	Imbellus
Marcus Specht	Open University of the Netherlands
Daniel Spikol	Malmö University
John Stamper	Carnegie Mellon University
Tamara Sumner	University of Colorado Boulder
Daniel Suthers	University of Hawaii
Davide Taibi	Tampere University of Technology
Kairit Tammets	Tallinna Ülikool
Yuta Taniguchi	Kyushu University
Michelle Taub	University of Central Florida
Stephanie Teasley	School of Information, University of Michigan
Chris Teplov	University of Michigan
Stefaan Ternier	Open University of the Netherlands
Amrita Thakur	D2L
Craig Thompson	University of Saskatchewan
Stefan Trausan-Matu	University Politehnica of Bucharest
Yi-Shan Tsai	The University of Edinburgh
Katrien Verbert	Katholieke Universiteit Leuven
Josine Verhagen	Kidaptive
Andrii Vozniuk	Ecole Polytechnique Fédérale de Lausanne
Elle Wang	Arizona State University
Elle Yuan Wang	Arizona State University
Barbara Wasson	University of Bergen
Professor Denise Whitelock	The Open University
John Whitmer	ACT, Inc.
Phil Winne	Simon Fraser University
Alyssa Wise	New York University (NYU)
Annika Wolff	Lappeenranta University of Technology
Marcelo Worsley	Northwestern University
Wanli Xing	Texas Tech University
Zhenhua Xu	University of Toronto
Zdenek Zdrahal	The Open University
Mengxiao Zhu	Educational Testing Service
John Zilvinskis	Binghamton University

Table of Contents

1 Practitioner Presentations

Understanding Teaching and Learning Practices of An Online Adaptive Mathematics Tutoring Platform	14
<i>Jean Yin Chiun Phua, Evan Min-Yang Yeo and Shannalyn Jia Yun Ng</i>	
Reflections of Visual Form Learning Analytics: Spaced Retrieval Practice Activity	20
<i>Kelly Mckenna, James Folkestad and Marcia Cristina Moraes</i>	
Designing a Digital Jigsaw Game based Measurement of Collaborative Problem-Solving Skills	26
<i>Pravin Chopade, David Edwards and Saad M. Khan</i>	
Evaluating a Learning Analytics Research Community: A Framework to Advance Cultural Change	32
<i>Linda Shepard, George Rehrey, Dennis Groth and Amberly Reynolds</i>	
Evaluating Preparatory Writing and Writing Placement at a Large Public University	38
<i>Sattik Ghosh, Emily Watkins and Meryl Motika</i>	
Developing an English Learner Corpus for Materials Creation and Evaluation	44
<i>Amanda Hilliard</i>	
Empowering Tutors with Big-data Learning Analytics	50
<i>Uma Vijn, Josine Verhagen, Webb Phillips and Ji An</i>	
Implementing Learning Analytics: Instructor Perspectives	56
<i>Julie Wei, Fred Cutler, Leah Macfadyen and Sanam Shirazi</i>	
Improving Dashboard Usability: A Case Study	62
<i>Bradley Coverdale and Matthew Hendrickson</i>	
Identification of sample comparability issues during the iterative design of game-based cognitive assessments	68
<i>Rebecca Kantar, David Laing, Matthew Emery, Sonia Doshi, Yao Xiong and Erica Snow</i>	
LAView: Learning Analytics Dashboard Towards Evidence-based Education	74
<i>Rwitaajit Majumdar, Arzu Akçapınar, Gökhan Akçapınar, Brendan Flanagan and Hiroaki Ogata</i>	

2 Doctoral Consortium

Towards Enhancing Conceptual Knowledge in Algebra through Diagrammatic Self-explanation in an Intelligent Tutoring System	80
<i>Tomohiro Nagashima</i>	
Investigating the Effectiveness of Online Learning Environments for Complex Learning	86
<i>Charlotte Larmuseau, Piet Desmet, Luigi Lancieri and Fien Depaepe</i>	

The Influence of Geo-Cultural Background on MOOC Learning Trajectories	92
<i>Saman Rizvi</i>	
The Effects of Discussion Strategies and Learner Interactions on Performance in Online Mathematics Courses: An Application of Learning Analytics	99
<i>Ji Eun Lee</i>	
The Use of Learning Analytics in a Blended Learning Context	107
<i>Elise Ameloot and Tammy Schellens</i>	
Analytics for the Measurement of Process Dimensions of Self-Regulated Learning and Feedback Impact	114
<i>John Saint</i>	
Innovation to Improve Learning: A Study of Content Modalities in Universal Design for Learning	120
<i>Catherine Manly</i>	
Designing a Teacher Dashboard for Interactive Simulations	127
<i>Diana López Tavares</i>	
Improving research students writing with writing analytics	135
<i>Sophie Abel</i>	
Trace Data: How to Improve a Method to Measure Self-regulated Learning in Online Courses	143
<i>Heeryung Choi</i>	

3 Posters and Demonstrations

IntVisRep: An Interactive Social Learning Analytics Tool	149
<i>Fan Ouyang</i>	
A stage-based matrix model of student progression	151
<i>Amelia Brennan and Pablo Munguia</i>	
Learning analytics adoption – approaches and maturity	153
<i>Yi-Shan Tsai, Vitomir Kovanović and Dragan Gašević</i>	
What Can We Learn About Learner Interaction When One Course is Hosted on Two MOOC Platforms?	155
<i>Yuanru Tan and Rebecca Quintana</i>	
Dynamic Feedback System supporting self-regulation, adaptive teaching and program level curricular development	157
<i>Ville Kivimäki and Joonas Pesonen</i>	
Comparing Interaction Activity Patterns of Different Achievement Learner Groups in MPOCs	159
<i>Di Sun, Gang Cheng, Pengfei Xu and Qinhua Zheng</i>	
Academic Quality Data Landscape: Establishing a Sustainable Process to Measure Learner Performance	162
<i>Mamta Saxena and Melanie Kasparian</i>	
Finding the At-Risk Online Learners: Development of the Online REadiness Screener (ORES)	165
<i>Oi-Man Kwok, Yu-Chen Yeh, Hsiang-Yu Chien, Noelle Wall Sweany, Eunkyeng Baek and William McIntosh</i>	

Examining the Effects of Adaptive Task Selection on Students' Motivation in an Intelligent Tutoring System	167
<i>Micah Watanabe, Kathryn Mccarthy and Danielle McNamara</i>	
Know Your Students: Empowering Educators to Improve Learning Using Data	169
<i>Matthew Steinwachs and Marco Molinaro</i>	
Examining Procrastination Behavior in Academic Settings Using a Mobile App	171
<i>Semih Bursali, Majed Ali and Reza Feyzi Behnagh</i>	
MOOC Effort Dashboard: An Interactive Web Dashboard Built in R	173
<i>Jason Baik, John Stamper and Huzefa Rangwala</i>	
Development of a Real Time Viewing Status Feedback System and Its Impact	174
<i>Yasuhiro Mori, Komei Sakamoto and Takahiko Mendori</i>	
Multimodal Learning Analytics: Society 5.0 Project in Japan	176
<i>Shizuka Shirai, Noriko Takemura, Yuta Nakashima, Hajime Nagahara and Haruo Takemura</i>	
Exploring Medical Education Learning Analytics from the Use of Electronic Health Record Systems	178
<i>Yancy Vance Paredes, Sarada Panchanathan, Pamela Carol Garcia-Filion and I-Han Hsiao</i>	
Development of a Visualization Tool for Student Characterization using Mobility Data	181
<i>Hieu Nguyen, Hyoungjoon Lim, Sung Bum Yun and Joon Heo</i>	
Learning Activity Analytics across Courses	183
<i>Atsushi Shimada, Takuro Owatari, Tsubasa Minematsu and Rin-Ichiro Taniguchi</i>	
Determining reading comprehension of domain texts	185
<i>David Quigley, Donna Caccamise, Peter Foltz, Eileen Kintsch, John Weatherley and Holly Kurtz</i>	
A Platform for Image Recommendation in Foreign Word Learning	187
<i>Mohammad Nehal Hasnine, Brendan Flanagan, Masatoshi Ishikawa, Hiroaki Ogata, Kousuke Mouri and Keiichi Kaneko</i>	
Tigris: An Online Workflow Tool for Sharing Educational Data and Analytic Methods	189
<i>John Stamper, Paulo Carvalho, Steven Moore and Kenneth Koedinger</i>	
Deciphering Dr. Discovery: Data Analytics for Interpreting Museum Visitor Demographics and Engagement with Exhibit Content	190
<i>Luis Perez Cortes, Brian Nelson, Catherine Bowman, Judd Bowman, Brooke Owen, Jeff Danas, Eleanor Dhuyvetter, Edgar Escalante, Kyle Rodgers, Abigail Weibel and Jesse Ha</i>	
Exploring Persistence and Regularity Behavioral Analytics in Online Self-Assessments	193
<i>Cheng-Yu Chung and Sharon Hsiao</i>	
Augmenting Authentic Data Science Environments for Learning Analytics	195
<i>Anant Mittal and Christopher Brooks</i>	
Critical Thinking Training and Formative Measurement Using Student Questions	197
<i>Turner Bohlen, Carolyn Bickers, Linda Elkins-Tanton, James Tanton, Calvin Dunwoody and Megan Allen</i>	
The Role of Learning Analytics in Redefining Nursing Skills for Artificial Intelligence and Robotization in the Healthcare	199
<i>Gábor Kismihók and Maritina Hasseler</i>	
Accessible Learning Analytics	202
<i>Mohammed Ibrahim, Daniel McSweeney and Geraldine Gray</i>	

Towards a Process to Integrate Learning Analytics and Evidence-Centered Design for Game-based Assessment	204
<i>Yoon Jeon Kim, José A. Ruipérez Valiente, Philip Tan, Louisa Rosenheck and Eric Klopfer</i>	
Elements of Success: Supporting At-risk Student Resilience through Learning Analytics	206
<i>Jae-Eun Russell and Anna Smith</i>	
Promoting college readiness in math with ALEKS: How restudy and learning behaviors relate to enrollment, achievement, and retention	208
<i>Matthew Bernacki, Kat Campise, Megan Bavaro Romero, William Speer and Diane Chase</i>	
(Un)Readiness for College Algebra: Using Learning Analytics to Design Interventions for Student Success	210
<i>Goutam Sarker, Lisa Berry, Shanna Banda and George Siemens</i>	
CoderBot: AI Chatbot to Support Adaptive Feedback for Programming Courses	212
<i>David Azcona, Enric Moreu, Sharon Hsiao and Alan Smeaton</i>	
Exploring Writing Analytics and Postsecondary Success Indicators	213
<i>Jill Burstein, Daniel McCaffrey, Beata Beigman Klebanov, Guangming Ling and Steven Holtzman</i>	
Growing an Institutional Data Lake into a Community Good	215
<i>Kevin Hartman</i>	
Multimodal Tutor Builder Kit	217
<i>Jan Schneider, Daniele Di Mitri and Hendrik Drachsler</i>	
Preparing Successful Facilitation: Designing A Teacher Dashboard to Support PBL Classroom Orchestration in A Game-based Learning Environment	218
<i>Yuxin Chen, Asmalina Saleh, Cindy Hmelo-Silver, Krista Glazewski and James Lester</i>	
Bttn: A Simple Data Collection App for Learning Analytics	221
<i>Charles Lang, Xiaoting Kuang and Sai Raj Reddy</i>	
Toward More Meaningful Analytics: Refining Social Presence Within the Community of Inquiry Model	222
<i>Valerie Barbaro</i>	
Log-based learning analytics in vector space	224
<i>Di Sun, Pengfei Xu, Junlei Du, Qinhua Zheng and Jingjing Zhang</i>	
Using Natural Language Processing to Assess Explanation Quality in Retrieval Practice Tasks	227
<i>Kathryn McCarthy and Scott Hinze</i>	
Data analysis and visualization for supporting academic writing and its instruction – the example of Thesis Writer	229
<i>Christian Rapp, Jakob Ott and Peter Kauf</i>	
Determining Learning Pathway Choices Utilizing Process Mining Analysis on Click-stream Data in a Traditional College Course	231
<i>Matt Crosslin, Justin T. Dellinger, Nikola Milikic, Igor Jovic and Kim Breuer</i>	
Page-wise Difficulty Level Estimation using e-Book Operation Logs	233
<i>Tetsuya Shiino, Atsushi Shimada, Tsubasa Minematsu, Kohei Hatano, Yuta Taniguchi, Shinichi Konomi and Rinichiro Taniguchi</i>	
Analytics of Time Management Strategies in a Flipped Classroom	235
<i>Nora Ayu Ahmad Uzir, Dragan Gasevic, Abelardo Pardo, Jelena Jovanovic and Wannisa Matcha</i>	

Examining Science Learning by At-Risk Middle School Students in a Multimedia-Enriched Problem-Based Learning Environment	237
<i>Sa Liu, Min Liu, Zilong Pan, Wenting Zou and Chenglu Li</i>	
Examining gameplay of high score achieving students: comparison of replaying after a failed gameplay	240
<i>Jihyun Rho and Gahgene Gweon</i>	
Predicting Graduation at a Public R1 University	242
<i>Henry Anderson, Afshan Boodhwani and Ryan Baker</i>	
Person-Oriented Approach to Profiling Learners' self-regulation in STEM learning	245
<i>Juan Zheng, Wanli Xing, Gaoxia Zhu, Guanhua Chen, Henglv Zhao and Xudong Huang</i>	
Understanding the factors contributing to persistence among undergraduate engineering students in online courses	247
<i>Samantha Brunhaver, Jennifer Bekki, Eunsil Lee and Javeed Kittur</i>	
CanoPy: Using Python Scripts to Promote Teacher-Driven Learning Analytics	249
<i>Charles Lang and Detra Price-Dennis</i>	

4 Workshops and Tutorials

Advances in Writing Analytics: Mapping the state of the field	251
<i>Antonette Shibani, Ming Liu, Christian Rapp and Simon Knight</i>	
Analyzing learners' online behaviour for student success and course enhancement: Case-studies from Blackboard	273
<i>Christine Armatas, Ada Tse and Chun Sang Chan</i>	
Learning Analytics Deployment Tactics: A meta-workshop	277
<i>Pablo Munguia</i>	
Supporting Feedback Processes at Scale with OnTask. A Hands-on Tutorial	285
<i>Abelardo Pardo, Shane Dawson, Dragan Gasevic and George Siemens</i>	
2nd Educational Data Mining in Computer Science Education (CSEDM) Workshop	289
<i>David Azcona, Yancy Vance Paredes, Sharon Hsiao and Thomas Price</i>	
Sharing and Reusing Data and Analytic Methods with LearnSphere	334
<i>Kenneth Koedinger, John Stamper and Paulo Carvalho</i>	
Python Bootcamp for Learning Analytics Practitioners	359
<i>Alfred Essa, Shirin Mojarad and Neil Zimmerman</i>	
Developing A Learning Analytics Community for Ethical Discourse	362
<i>James Folkestad, George Rehrey, Linda Shepard, Dennis Groth and Matthew Hickey</i>	
Workshop on Social-Emotional Learning (SEL): Assessment toward Diversity and Inclusion	368
<i>Elle Yuan Wang, Maria Ofelia San Pedro, Srecko Joksimovic and Jason Way</i>	
Predicting Performance Based on the Analysis of Reading Behavior: A Data Challenge	398
<i>Brendan Flanagan, Atsushi Shimada, Stephen Yang, Bae-Ling Chen, Yang-Chia Shih and Hiroaki Ogata</i>	
Beyond Identifying Areas for Improvement in Schools: Using the NILS™ Online Platform to Accelerate Improvement Work	492
<i>Ouajdi Manai, Hiroyuki Yamada and Susan Haynes</i>	

Interdisciplinary Learning Analytics: What to Know, Who to Talk To, and How It's Done	496
<i>Danielle Hagood and Robert Bodily</i>	
Fairness and Equity in Learning Analytics Systems (FairLAK)	500
<i>Kenneth Holstein and Shayan Doroudi</i>	
Analytics as a Team Sport: Using Cloud-Based Tools to Support Data-Intensive Research-Practice Partnerships	528
<i>Andrew Krumm, Jeremy Roschelle and Patricia Schank</i>	
3rd CrossMMLA: Multimodal Learning Analytics Across Physical and Digital Spaces	532
<i>Daniel Spikol, Daniele Di Mitri, Vanessa Echeverria, Roberto Martinez-Maldonado, Mutlu Cukurova, Luis P. Prieto, María Jesús Rodríguez-Triana, Xavier Ochoa, Marcelo Worsley and Michail Giannakos</i>	
Innovative problem solving assessment with learning analytics	580
<i>Lishan Zhang, Baoping Li, Yigal Rosen, Kristin Stoeffler and Shengquan Yu</i>	
How to Generate Actionable Predictions on Student Engagement: Hands-on Tutorial with Python Scikit-Learn	597
<i>Erkan Er</i>	
International Workshop on Technology-Enhanced and Evidence-Based Education and Learning	601
<i>Rwitajit Majumdar, Ivica Boticki and Hiroaki Ogata</i>	
Workshop on Educational Data Visualization	685
<i>Nirmal Patel, Derek Lomas and Collin Sellman</i>	
3rd Annual Workshop of the Methodology in Learning Analytics Bloc (LAKMLA19)	715
<i>Yoav Bergner, Geraldine Gray and Charles Lang</i>	
The Fifth LAK Hackathon: Trusted and Inclusive Learning Analytics Across Spaces with New Tools, Standards and Infrastructures	719
<i>Daniele Di Mitri, Adam Cooper, Kirsty Kitto, Gábor Kismihók, Stefan T. Mol, Niall Sclater, Jan Schneider and Alan Berg</i>	
Diving in to Educational Experiments: Process, Evaluation, and Reasoning in Support of Learning (DEEPER Support of Learning)	728
<i>Christopher Brooks, Dan Davis, Nia Dowell, Joshua Gardner, Timothy Necamp, Oleksandra Poquet, Rene Kizilcec and Joseph Williams</i>	
Connectivism: Using learning analytics to operationalize a research agenda	732
<i>Srecko Joksimovic, George Siemens, Shane Dawson and Vitomir Kovanovic</i>	
VISLA: Visual Approaches to Learning Analytics	750
<i>Katrien Verbert, Robin De Croon, Tinne De Laet, Tom Broos, Xavier Ochoa, Robert Bodily, Judy Kay, Hendrik Drachler and Cristina Conati</i>	
AutoTutor Tutorial: Conversational Intelligent Systems and Learning Analytics	802
<i>Bor-Chen Kuo, Chen-Huei Liao, Kai-Chih Pai, Chia-Hua Lin, Xiangen Hu, Zhiqiang Cai and Arthur C. Graesser</i>	
2nd Personalising feedback at scale Workshop: Focusing on Approaches and Students	806
<i>Lorenzo Vigentini, Danny Y.T. Liu and Lisa Lim</i>	
Exploiting data intelligence in education from three levels: Practice, challenges and expectations	836
<i>Gu Xiaoping</i>	

Scalability and Sustainability of Learning Analytics Solutions

888

Tom Broos, Dragan Gašević, Abelardo Pardo, Hendrik Drachsler, Rafael Ferreira, Katrien Verbert and Tinne De Laet

Workshop on Learning Analytic Services to Support Personalized Learning and Assessment at Scale

956

Alina Von Davier, Michael Yudelson, Kenneth Koedinger, Steven Ritter and Peter Brusilovsky

Understanding Teaching and Learning Practices of Online Adaptive Mathematics Tutoring Platform

Jean Yin Chiun Phua

Senior Specialist, Technologies for Learning Branch, Educational Technology Division
Jean_PHUA@moe.gov.sg

Evan Min-Yang Yeo

Educational Technology Officer, Technologies for Learning Branch, Educational Technology Division
Evan_YEO@moe.gov.sg

Shannalyn Jia Yun Ng

Educational Technology Officer, Technologies for Learning Branch, Educational Technology Division
Shannalyn_NG@moe.gov.sg

ABSTRACT: Presentation. Online platforms offer the promise, through artificial intelligence, of providing optimal course pacing and content to fit each student's needs, thereby improving educational learning. The latest "intelligent" tutoring systems, not only assess students' current weaknesses, but also diagnose why students make the specific errors. These systems then adjust instructional materials to meet students' needs. In our context, schools prevalently administer online learning in mathematics for students from Grades 7 to 9, with some claims that these platforms are adaptive. This on-going action research study documents the journey of a group of practitioners who sets out in a two-phase process to understand teachers' current use of an adaptive learning platform and to generate teaching and learning insights of an adaptive mathematics platform. In the first phase, the team seeks to understand teachers' teaching and learning practices, their beliefs about adaptive systems and their use of the adaptive feature of existing mathematics tutoring platforms through a survey with 53 teachers from 19 schools. Phase two involves a quasi-experimental evaluation of the implementation of a selected adaptive mathematics platform with two classes of students from two different schools with 40 Grade 7 students, through four to five sessions in deriving the student-tool-teacher interactions.

Keywords: Adaptive Learning, Personalised Learning, Intelligent Tutoring, Mathematics

1 INTRODUCTION

Online learning systems have been and will be increasingly leveraged to support classroom instruction. In our context, many local schools administer some form of online mathematics learning platforms from Grades 3 to 9. However, it is observed that there are no studies or findings on how these systems are used to improve student learning. Teachers also claim to have used the adaptive features on these platforms to provide optimal course pacing and content to fit each student's needs but it is not often understood if these benefitted students. We aim to understand how teachers' existing teaching practices of current platforms, their beliefs on the potential of adaptive platforms and the use of the adaptive features of existing online mathematics platforms. The team surveyed and evaluated existing platforms and have conclusively decided that teachers have not utilised existing platforms to provide the level of adaptive nature expected of an "intelligent" tutoring or adaptive learning system. The team hopes to gain deeper insights into students' experience of an adaptive learning mathematics platform as well as working knowledge of the affordances of such platforms.

2 LITERATURE REVIEW

For the purposes of this paper, Adaptive Learning (AL) is defined as education technologies that can respond to a student's interactions in real-time by automatically providing the student with individual support (EdSurge, 2016). Such adaptive learning tools collect specific information about a student's behaviours by tracking the student's responses. The tools then respond to each student by changing the learning experience to better suit that student's needs, based on the unique and specific behaviours and answers provided. Non-examples of AL include providing all students with the same response, or marking students' responses and providing them with the same learning pathway. Where real-time data is not collected, or if data is collected through a single assessment with a prescribed path of learning, that tool does not support adaptive learning.

Research studies conducted by companies owning the AL tools and learning institutions deploying such tools seem to indicate several positive findings. These include **less time needed to master the learning of topics, higher passing rates** and **higher achievement gains** (Johanes & Lagerstrom, 2015). A meta-analytic study by VanLehn (2011) in comparing the effectiveness of human tutoring, intelligent tutoring and no tutoring concluded that intelligent tutoring systems are "just as effective" as human one-on-one tutoring for increasing learning gains in STEM (Science, Technology, Engineering and Mathematics) topics. In particular, he found an effect size of 0.79 for human one-on-one tutoring as compared to no tutoring, and an almost identical effect size of 0.76 for computer-based tutoring. It can be inferred that **AL providing one-on-one tutoring is as effective as human one-on-one tutoring**. These promises are convincing for the team to study how an adaptive learning platform can benefit student learning in the local context.

3 PHASE 1: CURRENT STATE OF USE

In Phase 1, the team is interested in understanding the current state of use of online mathematics tutoring system and in gathering grounds-up anecdotal feedback on the effectiveness of deploying such systems. The three key questions we want to present for this paper are:

1. How are teachers leveraging existing online mathematics tutoring platform?
2. Are teachers using the adaptive feature? If not, why? If yes, how?
3. What are their beliefs about adaptive learning platforms?

3.1 Data Collection

This segment details the methodology taken and discusses the findings obtained.

3.1.1 Instrumentation

The survey items comprise of a combination of multiple-choice, open-ended response and Likert scale questions that aimed to understand how teachers use mathematics online platforms in their classrooms. Some of the sample questions in the survey are shown in Table 1.

3.1.2 Participants

An email invite was sent to a total of 19 schools teaching students from Grades 7 to 10. The research team also approached teacher participants at mathematics workshops to respond to the online survey. A total of 53 mathematics teachers teaching Grades 7 to 10 with 1 to 35 years of teaching experience responded. The years of teaching experience that our respondents have approximates a

normal distribution with skewness of 0.89 and kurtosis of 3.12. Thus, we have a sample that is representative and contains teachers with a wide range of teaching experience.

Table 1: Sample survey questions

Q1	What are some of the features that you found most useful for students? (You can select more than 1 option.)
	<input type="checkbox"/> Question bank of exam papers for Grades 7 to 10; organized school-wise and topic-wise <input type="checkbox"/> Interactive tools and resources such as Virtual Manipulatives, Exploratory Activities, etc. <input type="checkbox"/> Mathematics Exam Revision Kit with practice questions, including answer keys, thinking process and working steps <input type="checkbox"/> Topic specific games and multi-player games
Q2	We want to hear your beliefs about the adaptive capability of the platform; please feel free to choose the option that you identify with the most. (There is no right or wrong answer)
	<input type="radio"/> The adaptive feature helps choose questions so that it tailors to each individual student. <input type="radio"/> The adaptive feature provides another question bank from which students can attempt more assessment questions. <input type="radio"/> The adaptive feature allows me to be less focused on keeping track of students' progress and let them be in charge of their own learning progress. <input type="radio"/> The adaptive feature helps choose questions so that teachers can spend time on other teaching activities instead of choosing assessment questions for their students.

3.2 Findings

We found that teachers envisioned the potential for using adaptive platforms for personalised learning and self-directed learning. Despite the overwhelming optimism in their beliefs, however, there remained a dissonance between their beliefs and how they currently use mathematics platforms in their own classrooms. The teachers in our sample answered that they mostly use these platforms for examination revision, interactive resources and to support e-learning programmes.

34% of the respondents answered that they have used some sort of adaptive feature. They used it to bridge learning gaps for weaker students and for drill-and-practice purposes. The remaining 66% of the sample cited reasons such as an unfamiliarity with the platforms, preferring to stick to tried-and-tested methods and students' lack of discipline for not attempting to use the adaptive features.

4 PHASE 2: UNDERSTANDING ADAPTIVE LEARNING

4.1 Design of Implementation

4.1.1 Selecting Adaptive Learning Platform

A US based adaptive learning mathematics platform was selected, based on the following set of criteria: (1) the product is research-based, (2) students experience a cycle of assessment and learning, (3) each student is provided with a different start state based on pre-test performance, (4) practice worksheets individualised for each student's knowledge state is provided, (5) immediate feedback for each question with in-built explanation for each question, (6) remediation practices based on student's knowledge state and (7) progress monitoring. The only short-coming of the product is the lack of fit to local content, for language use and the metric unit of measurement.

4.1.2 Setting Up Adaptive Learning Platform

An attempt to map the US-based content to the local syllabus, resulted in the choice of 350 topics (refer to Table 2) within High School Algebra to be administered to local Grade 7 students.

Table 2: Exemplar list of Topics within High School Algebra

Main Topics	Examples of Sub-Topics	Number of Topics
1. Arithmetic Readiness	a. Factors, Multiples, and Equivalent Fractions b. Addition and Subtraction with Fractions	116
2. Real Numbers	a. Operations with Signed Numbers b. Exponents and Order of Operations	104
3. Linear Equations	a. Multi-Step Linear Equations b. Applications of Linear Equations	82
4. Linear Inequalities	a. Writing and Graphing Inequalities	6
5. Functions and Lines	a. Tables and Graphs of Lines b. Introduction to Functions c. Arithmetic Sequences	13

4.1.3 School Participants

A total of two schools responded to an invitation to participate. Complete data of the administered online pre-post quizzes were obtained from 35 students. Each school provided consent to schedule one hour of face-to-face contact with the students in the computer lab to carry out the online knowledge quizzes as well as learning time, over four or five sessions every week or every two weeks depending on the school curriculum schedule. These sessions were scheduled outside of curriculum time.

4.1.4 Implementation

The purpose of this exploratory study was to gain an understanding of the affordances of an adaptive learning mathematics platform, how it works and the student-tool-teacher interactions involved.

Each school committed a fixed number of scheduled one-hour session, offering the use of the computer laboratory, providing each student with access to a computer laptop. For every student's first encounter, a knowledge quiz must be administered to gauge the prior knowledge level of the student. This knowledge quiz is essential to establish what the students know, and what the students need to learn to gain mastery in High School Algebra. After the knowledge quiz, the research team then offered an explanation on how students could access and use the platform to help them with the mastery of learning of the concepts. Students were informed that they were revising and learning and that the platform served as their personalized mathematics tutor. All students were encouraged to access the platform for at least half an hour, and for up to twice a week, before the next face-to-face computer laboratory session.

Between the first and the last scheduled knowledge quiz session, for each face-to-face session, the research team coached the struggling students while they interacted with the platform. At the start of every session, a ledger board of the students who had spent time interacting with the platform would be flashed on the screen, to encourage active self-learning without teacher supervision. The research team motivated the students with an extrinsic gift for the top 10 students for each school, who clocked the most amount of time learning on their own.

To gain deeper insights on students' experience, selected students were interviewed to understand: (1) What they like or dislike about the platform, (2) What they think could motivate or discourage them to access the platform on their own, without teacher supervision.

4.2 Observed Teacher-Tool-Student Interactions

The theory of instrumental genesis (Lonchamp, 2012; Rabardel and Beguin, 2005) explains how learners appropriate technological tools and accomplish tasks while interacting with these tools. Instrumental genesis is the complex integrative and dynamic processes where learners are able to incorporate and appropriate the potentialities and emerging possible use of an artefact (adaptive learning platform) adapting it into their individual and group activities (Trouche, 2004 and Drivers et al. 2010). In this segment, we think about the features of the adaptive learning platform, how they can be used by teachers and students in the classroom, and how the features of the ICT tool influence the way teaching and learning is done. Figure 1 presents a summary of the student-tool-teacher interactions derived through iterative team discussion amongst the research team members. The key tenets are that the (1) Student- feedback provided to the learner is clear, constructive and immediate through the tool and teacher presence, (2) Tool- difficulty level of the achievement task is appropriate to the progress level of the learner and (3) Teacher- adequate scaffolds and timely support is given to learners through motivating students to complete their task.

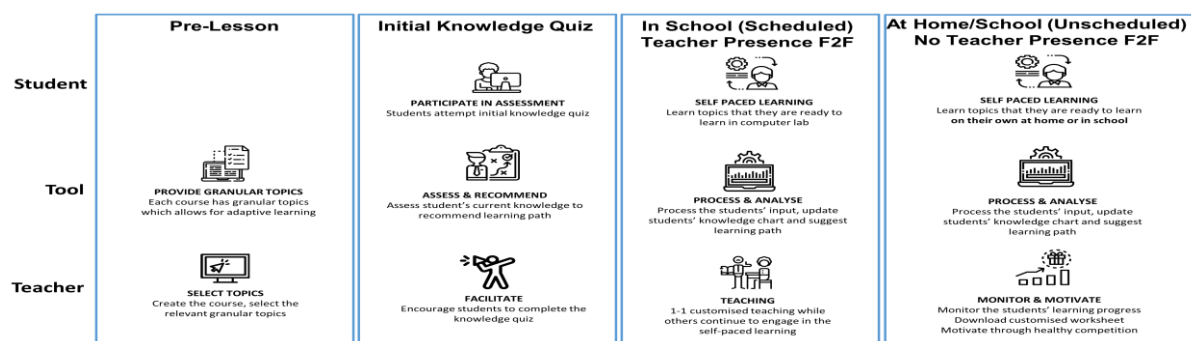


Figure 2: Student-Tool-Teacher Interactions

4.3 Findings

From the 35 students who participated in the sessions, all of them gained mastery. The level of progress mastery corresponded to the amount of time spent in the system. Mastery Progress is defined as the difference in mastery between initial and latest quiz which reflects the growth in mastery after accounting for initial knowledge. There was a strong positive correlation between mastery progress ($M = 6.1\%$, $SD = 2.5\%$) and time spent learning ($M = 228.9$, $SD = 99.8$) in minutes. Pearson's correlation, $r = 0.75$, which was significant at the 1% significance level. ($n = 35$)

A selection of 10 students were interviewed. Majority of the students liked the user-friendly system which afforded a lot of practice with many topics. Several students echoed that the targeted feedback with similar questions surfaced helped them to improve and enhance their confidence level. For example, a student articulated that *"The explanation provides an overview of what went wrong. After reading it, I can attempt the next question, which is similar to the previous one."*

Intrinsic and extrinsic factors for engaging them on the platforms are present, it was generally the extrinsic factor that played a greater role; with quotes like *"Give more prizes"*, *"We want to improve"* and *"Friends are also doing math practice"*. Most times, in the local context, students were overloaded with other homework and this study was packaged as an additional after-school activity, and hence most students were not motivated to access the system beyond curriculum time citing reasons such as *"No time"*, *"Forget password"*, *"Lots of projects to do"* and *"No computer at home"*.

5 PHASE 2: NEXT STEPS

The opportunity to participate in this research has enabled the team to acclimatise to and deepen our understanding of the affordances of an adaptive mathematics learning platform. Some of our preliminary learning points can be summarised as follows:

1. Detailed curriculum mapping is necessary for facilitating adaptive learning. The adopted mathematics learning platform provides insights into the depth of granularity of the topics.
2. Re-imaging how to teach. With AL, teachers have to re-imagine how to teach. The adaptive platform has to be deployed in a learning environment that supports students working at their own pace, on different content and on different skills at different levels. This could be rather disruptive to teachers who are more used to planning and conducting their lessons based on a pre-determined scheme of work. For the pilot trial, the participating schools did not engage the AL systems as part of their classroom teaching practice. They were more inclined to deploy AL as a supplement to classroom teaching beyond curriculum time or during remedial sessions, and this practice did not encourage active student participation after the face-to-face sessions.

The team will be revising some of its implementation strategies in the next phase of the research in re-imaging how to teach with AL. Instead of positioning the AL as an after-school remediation, the AL could be infused into curriculum time through flipped learning approach, with students having to review the content before actual lessons, changing actual classroom teaching practice to focus on other aspects. The system may still be deployed as a system to supplement teaching. It is observed that the skills taught within the AL system are procedural, and hence the teachers might wish to spend time to expand on the conceptual understanding of selected content within curriculum time.

REFERENCES

- Drijvers, P., Doorman, M., Boon, P., Reed, H., & Gravemeijer, K (2010). The teacher and the tool: instrumental orchestrations in the technology-rich mathematics classroom. *Educational Studies in Mathematics*, 75(2), 213-234. *Educational Studies in Mathematics*.
- Johanes, M. P., & Lagerstrom, L (2015). Adaptive Learning: The Premise, Promise, and Pitfalls. Retrieved from <https://peer.asee.org/27538.pdf>
- Lonchamp, J. (2012). An instrumental perspective on CSCL systems. *International Journal of Computer-Supported Collaborative Learning*, 7(2), 211–237.
- Moskal, P., Carter, D. and Johnson D. (2017). 7 Things You Should Learn About Adaptive Learning. Retrieved from <https://library.educause.edu/resources/2017/1/7-things-you-should-know-about-adaptive-learning>
- Rabardel, P., & Beguin, P. (2005). Instrument mediated activity: from subject development to anthropocentric design. *Theoretical Issues in Ergonomics Science*, 6(5), 429–461.
- Trouche, L. (2004). Managing complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning*, 9, 281–307.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197-221.

Reflections of Visual Form Learning Analytics: Spaced Retrieval Practice Activity

Kelly McKenna

Colorado State University
kelly.mckenna@colostate.edu

James E. Folkestad

Colorado State University
james.folkestad@colostate.edu

Marcia Cristina Moraes

Colorado State University
marciacristina.moraes@gmail.com

ABSTRACT: This presentation highlights a study utilizing student learning analytics accessible to course instructors and administrators through learning management systems. Students completed self-regulated retrieval practice activities, quiz-based learning opportunities, which were then presented to them as visual-form learning analytics. Visual-form learning analytics create opportunities for feedback and critical reflection and improve student learning. Learners were prompted to reflect on these personalized visual-form learning analytics. Findings from the reflections and visual-form learning analytics include students' understanding of high impact learning practices, the realization of intended study behaviors versus engrained behaviors, high score orientation, and a focus on comparisons.

Keywords: learning analytics, visualizations, high impact learning practices

1 INTRODUCTION

In learning environments that rely on self-regulation, such as online classes or online class activities, “making effective choices and adaptation of learning strategies in response to the emerging needs from the learning environment are critical features of effective self-regulated learning” (Gasevic, Jovanovic, Pardo, & Dawson, 2017, p. 115). LA, specifically those presented in the visual-form, can provide information that supports learners' reflection and guides them to the necessary changes that lead to successful self-regulated learning. This study used a photo-elicitation research method to create visual-form learning analytics (LA) as a tool for reflection of students' self-regulated learning behaviors. Visual-form LA consist of the process of representing learner usage data as a visualization. Student reflections and the visual-form LA were analyzed by the researchers to identify applications related to the use of visual-form LA in online higher education. The findings from this study ascertained students' application of learning, perfunctory behaviors regarding study habits, assessment emphasis, and affinity for comparisons.

2 LITERATURE REVIEW

2.1 High Impact Learning Practices

“Learning is an acquired skill, and the most effective strategies are often counter intuitive” (Brown, Roediger, & McDaniel, 2014, p. 2). One category of effective strategies for learning is high-impact learning practices (HILPs). These practices give students studying and learning strategies varied from most commonly used techniques such as re-reading or rote memorization. Some HILPs include reflection, spaced retrieval, interleaving, elaboration, and mental models. The practices help learners identify for themselves what they don't know as compared to what they do know in memorization of what is provided.

2.2 Quizzes as a Learning Strategy

Weinstein, Husman, and Dierking (2000) define learning strategies as “any thoughts, behaviors, beliefs or emotions that facilitate the acquisition, understanding or later transfer of new knowledge and skills” (p.227). One example of a learning strategy is retrieval practice testing, or low-stakes quizzes that account for little or nothing towards a student’s grade in the course (Roediger, Agarwal, McDaniel, & McDermott, 2011). They can be a pedagogical choice for the purpose of creating a learning strategy rather than simply as an assessment. Testing enhances learning and retention of the material, as well as the metacognitive use of tests which informs learners regarding what they do and do not know, allowing them to concentrate study efforts on topics they do not know. Several studies have argued that effective online quizzes can enhance learner engagement and have a positive impact in student’s learning outcomes (Gikandi, Morrow, & Davis, 2011; McDaniel, Wildman, & Anderson, 2012; Balter, Enstrom, & Klingenberg, 2013; Berrais, 2015; Cohen & Sasson, 2016; O'Dowd, 2017).

2.3 Learning Analytics

In technology-aided classrooms, data about student’s work exists in massive quantities (Gasevic, Jovanovic, Pardo, & Dawson, 2017). Utilizing this data for student and instructor knowledge transforms the data into LA with specific goals aiding students and faculty. There are two primary principles regarding the use of LA: “to understand and to optimize learning and learning environments in which learning occurs” (Gasevic, Jovanovic, Pardo, & Dawson, 2017, p. 113). Some insights can be directed to class-level where learners immediately benefit and others to course-level where learners benefit more in the future planning and design.

2.4 Visual Data for Learning

Images are “unique sources of evidence” that create opportunities for dialogue (Rose, 2007, p. 238) and making meaning from the data can come from the creation of a visualization based on the data. Materials such as photographs, videos, drawings, collages, maps, graphs, and diagrams are just some forms of potential visual data (Harper, 2002; Wagner, 2006). Visualizations of user actions, or visual-form LA, can be used in technology enhanced learning to support the learning process, increase awareness for learners and teachers, and to support self-reflection (Kruse & Pongsajapan, 2012; Govaerts, Verbert, Duval, & Pardo, 2012; Beheshitha, Hatala, Gašević, & Joksimović, 2016).

3 RESEARCH QUESTION

How does the integration of visual-form LA contribute to teaching and learning practices?

4 METHODS

Throughout the course students were assigned self-regulated retrieval practice activities (RPAs), quiz-based learning opportunities, in which students were to implement the HILPs learned throughout the course. Following the final RPA, learners were presented with a visual representation of their personalized RPAs in the form of visual-form LA. The visual-form LA were then utilized as a reference tool for prompted reflection questions designed to create an opportunity for students to reflect on their semester-long learning behaviors.

4.1 Setting and Participants

This project was initialized in an online master’s level class, On Demand Learning-Improving Performance in the fall 2017 semester. Twenty-four adult learners were registered for the course with 19 consenting to be included in the research. The objectives of the course included developing connections between supportive learning theories and applying the theories in course assignments.

4.2 Data Collection

There were two phases of data collection in this study. In the first phase usage data was collected and transformed into visual-form LAs (see figure 1). In phase two, the photo-elicitation research method was used in a process of gathering student reflections, written accounts, and interpreted meanings of those visuals. In order to build the visual-form of learners' RPA attempt behavior, we extracted quiz-logfile data from the LMS. A Python-based application (U-Behavior) was developed by the researchers to extract the necessary data from the LMS. From this data, a visual-form LA was generated for each student. The X-axis reflects the RPA submission times and the Y-axis represents the score obtained in each attempt. Each attempt is represented by a colored node and each color signifies one of the RPAs offered during the class.

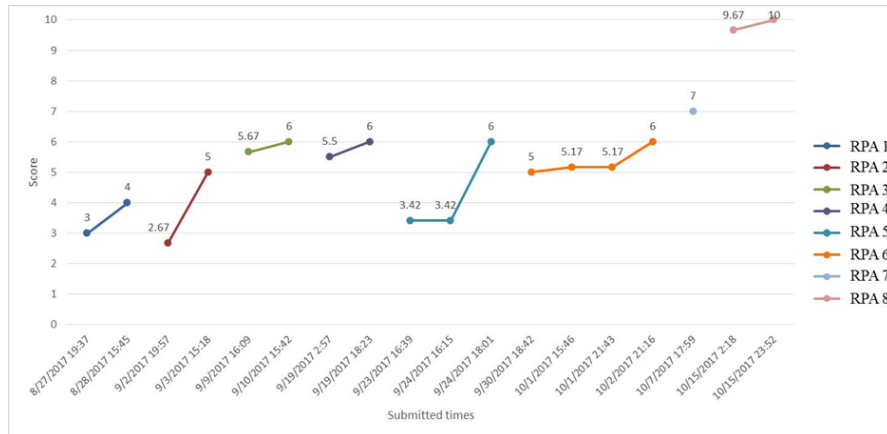


Figure 1: Example of Student's Visual-Form Learning Analytics

Phase two consisted of student reflections. Following creation of the visual-form LA, students were prompted with a series of reflective questions and asked to reflect on the visual representation (visual-form LA) of their learning behaviors. These reflection prompts incorporated questions from two perspectives: personal reflection of the RPAs as the student and from the perspective of an instructor reviewing the RPA data.

4.3 Data Analysis

Students' visual-form learning analytics and photo-elicitation responses were analyzed and coded for the emergence of visual and written themes. Each researcher identified several initial codes which were then presented to the team, which allowed the group to combine codes and collectively define four themes. Discussions among the researchers continued throughout the data review, analysis, and reporting of the combined themes, particularly regarding candid ideas from the students versus prompted themes.

5 FINDINGS AND DISCUSSION

Four themes emerged from analysis of the visual-form LA and reflection activity. The four themes were: HILPs understanding, individual study behaviors-intent vs. engrained practices, high score-orientation, and the use of comparisons to evaluate work and efforts based on the visual-form LA.

5.1 High Impact Learning Practice Understanding

In the reflections, learners described HILPs such as interleaving, elaboration, summarization, reflection, content retrieval, repeated elaboration, and spaced retrieval. The integration of HILPs and lack of utilization were both included in the reflective responses, but so were contradictions to learning. Students completed all RPAs before seeing their data visualized as visual-form LA, but

learners remarked that the visualizations of their scores offered additional learning practices: “If reflection is truly an effective form of learning, this tool really supports that action and personal growth” (Student O). One student specifically identified the reflection opportunity: “It’s nice to see it in retrospect, I think. These reflection questions were also helpful to try to think back on learning and apply the analysis to our learning” (Student M). In comparing the student’s reflective responses to the visualizations each student received, one student appeared to have a misperception of the HILPs. The student’s visual-form LA did not show attempt data that would indicate the use of particular HILPs, but their written reflections stated they used specific techniques. This disconnect between the meaning of the practice as presented in the class and the student’s explanation of a practice not represented in the visual-form LA could be an indication of misunderstanding of the content. This realization of potential lack of understanding might not have been identified without the combination of the visualization and reflective response. Student M provided this reflection: I didn’t re—take any of them [the RPAs] immediately after the first attempt. I took the suggested recommendation and studied a little more, then took it a day or two later—showing interleaving and some desirable difficulty. (Student M)

5.2 Individual Study Behaviors-Intention vs. Engrained Practices

Within the reflections several students commented on their intention to utilize HILPs in their RPA attempts, but found themselves returning to routine or habitual learning practices. “I fell back into old habits, even though I knew spaced practice would be better. I often took the tests really quickly to try and ‘check it off the list’” (Student O). However, the battle between a student’s intention and their engrained practices offered an opportunity regarding their learning strategies: “As the course progressed I found myself taking a quiz not to demonstrate performance but as a review of what I had learned” (Student H). Many of the learners recognized the opportunity to complete the RPAs employing HILPs, but found it challenging to connect the strategies to their study practices.” What is unfortunate is that everything I have learned I find so much value in, but it seems as though shifting that mindset is harder than I anticipated. I did take the quizzes at least a day later, which was new for me, but I found myself having a harder time going back when I did do well on the quizzes” (Student E).

In review of the visual-form LA for each individual, we determined that a total of 10 learners attempted spacing, two attempted interleaving and 18 practiced retaking. Learners’ responses confirmed the division between their intentions of completing the RPAs with various learning strategies and the actual application, however some did implement what they learned in class. Student R stated: “From this course, I have learned the importance of interleaved practice and trying again later. After each quiz, I would review the material that I missed and reflected on why I chose that answer, and why it was incorrect.”

5.3 High Score Oriented

In reviewing the visual-form LA, 14 of the 19 students appeared to have been seeking the high score in the majority of the RPAs they attempted as indicated by attaining the high score and then ceasing their attempts; no students re-attempted an RPA after the high score was realized. Considering the reflective responses, only seven students indicated their high score orientation. For example: “The graph depicts my eagerness to complete the quiz in hopes of receiving a high score. It was my goal to achieve full marks while not exhausting my remaining takes.” (Student A). Another student (C) explained their high score orientation: “The pursuit of ‘maximum point value’ served as the conditioned objective and since the grading rubric for the class incorporated only the highest score, once I achieved that score, I checked it off the mental ‘to-dos’”. Learners’ final grade alone on each RPA would not represent the high score-oriented strategy due the option for multiple attempts and the opportunity to receive the highest score realized, however the visual-form LA clearly presents

high score orientation. Student C, a high score-oriented student, reflected that: “I didn’t use that tool [visual-form LA] in a way that enhanced my true learning from the course. And honestly, it sort of bums me out because I missed an opportunity”. This learner’s statement regarding the visual-form LA offers an honest reflection that they missed the opportunity for learning that was presented as they instead sought a high score. In pursuit of the high score, Student C had the highest number of attempts on a single RPA, 10, which was the maximum number of attempts allowed on the RPAs. The first five attempts had the same score and all ten attempts were completed in quick succession over a period of 14 minutes, which provides basis for the assumption that no HILP was used.

One learner’s reflection showed a change in their intention and shift from the initial motivation of the high-score after experiencing the learning activity: “I didn’t like it at first, because I want to score a perfect 100%. I was frustrated after only getting two questions right on my first RPA. After getting used to taking these, I began to feel more at ease, which I believe also helped improve my performance. I wasn’t stressing on getting the correct answer. Instead I was thinking smarter about which answers were most correct from what I remembered. I found this exercise to be very helpful and it has really improved my retention of the materials” (Student I). Student I retook 100% of the RPAs and achieved a high score on 75% of the activities they completed. The visual-form LA show progressive scores throughout the attempts and paring the visual depictions with the learners’ reflections gave us insights into the learners’ desires to frequently reach the highest score.

5.4 Use of Comparisons

Learners frequently expressed their thought processes using comparisons. This included comparisons of their own visual-form learning analytics and interest in comparing to others. Some compared their own attempts on a single RPA as the grade-driven outcome of the learning objective. “The one overarching thing I could conclude was that between attempts, some studying and learning was being done because the trend is all upward/positive. There were no lower scores than a previous attempt” (Student M). Learners also compared their attempts on different RPAs, “the graph was useful in that it showed me my performance throughout the semester. I forgot my performance in the beginning of the semester to be reminded was helpful. It also brings light to the areas you need to improve” (Student E).

Additionally, while learners were not given the data of other students, several commented on the interest in seeing how and what others were doing when completing the RPAs. “It might be helpful to show a composite result (of a previous class, perhaps) to the current class about mid-way in a course. If this information could compare and contrast grade improvements that were spaced out as opposed to taken in quick succession. This could reinforce the point that spaced learning is effective” (Student O). Even in a personal reflection activity, which should be focused on self, the desire to compare to others and their motivation/approach for the activities was present.

6 IMPLICATIONS

In this study we found evidence that visual-form LA can be used by instructors as a tool for feedback, and when provided to the learners a powerful tool for their own critical reflection. Visual-form LA can be provided directly to students as feedback or as a tool for instructors to gather information for feedback. Instructors can use visual-form LA to better understand students’ learning behavior, identifying their strengths and areas that need to be improved. The use of visual-form LA can also contribute to improved learning strategies as it presents information that is useful for both instructors and students to reflect on. Presenting learners with visual-form LA creates opportunities for them to reflect on and understand their practices; changing and improving their learning strategies when they think it could be beneficial for them.

7 CONCLUSION

The visual-form LA utilized in this research were created by the researchers from raw LMS data of student activity on quiz-like RPAs. The question of understanding the contributions of these visualizations for both learners and instructors were satisfied through analysis of the visual-form LA and the corresponding student reflections. Four primary findings emerged including learners': deeper understanding of the HILPs, returning to engrained learning habits vs. initial intentions, orientation towards earning the highest score possible, and employing comparisons to understand their visualizations. Faculty have the opportunity to use visual-form LA as a form of feedback to deepen learners' understanding and to improve student learning strategies through critical reflection.

REFERENCES

- Balter, O., Enstrom, E., & Klingenberg, B. (2013). The effect of short formative diagnostic web quizzes with minimal feedback. *Computers & Education*, 60(1), 234-242.
- Beheshitha, S. S., Hatala, M., Gašević, D., Joksimović, S. (2016). The role of achievement goal orientations when studying effect of learning analytics visualizations. *LAK '16 Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 54-63. Edinburgh, UK.
- Berrais, A. (2015). Using online Moodle quizzes to support the teaching of mathematics to foundation engineering students. *QScience Proceedings: Engineering Leaders Conference 2014*, 8.
- Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014) Make it stick : The science of successful learning. Cambridge, MA: Harvard University Press.
- Cohen, D. & Sasson, I. (2016). Online quizzes in a virtual learning environment as a tool for formative assessment. *Journal of Technology and Science Education*, 6(3), 188-208.
- Gasevic, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *Journal of Learning Analytics*, 4(2), 113–128. Retrieved from <http://dx.doi.org/10.18608/jla.2017.42.10>
- Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, 57(4), 2333-2351.
- Govaerts, S., Verbert, K., Duval, E. & Pardo, A. (2012). The student activity meter for awareness and self-reflection. *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 869–884.
- Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual Studies*, 17(1), 13-26.
- Kruse, A. & Pongsajapan, R. (2012). Student-centered learning analytics. *CNDLS Thought Papers*, 1–9.
- McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, 1, 18–26.
- O'Dowd, I. (2018). Using learning analytics to improve online formative quiz engagement. *Irish Journal of Technology Enhanced Learning* 3(1). <https://doi.org/10.22554/ijtel.v3i1.25>
- Roediger, H. I., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17(4), 382-395. doi: 10.1037/a0026252
- Rose, G. (2007). Making photographs as part of a research project: Photo-elicitation, photo-documentation and other uses of photos. In G. Rose, *Visual methodologies: An introduction to the interpretation of visual materials* (2nd ed.) (pp. 237-256). Thousand Oaks, CA: Sage.
- Wagner, J. (2006). Visible materials, visualized theory, and images of social research. *Visual Studies*, 21(1), 55-69.
- Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Self-regulation interventions with a focus on learning strategies. In P. R. Pintrich & M. Boekaerts (Eds.), *Handbook on self-regulation* (pp. 727–747). New York: Academic Press.

Designing a Digital Jigsaw Game based Measurement of Collaborative Problem-Solving Skills

Pravin Chopade, David Edwards and Saad M. Khan

ACTNext, ACT Inc, Iowa City, IA, USA

Pravin.Chopade@act.org, David.Edwards@act.org, Saad.Khan@act.org

ABSTRACT: In this paper, we describe a non-invasive, in vivo approach to assessing collaborative problem solving (CPS) skills. Specifically, we focus on digital collaborative environments where behaviors indicative of CPS skills can be captured in multiple modalities including video, audio, and eye tracking recordings and analyzed using machine learning techniques. We use an online CPS game that involves a two-player jigsaw task composed of a series of puzzles. The paper describes our computational framework for evidence extraction and accumulation and presents early stage results from a pilot study.

Keywords: Collaborative Problem Solving, Collaborative learning environment, Skills, Evidence, Machine Learning, Human Behavior.

1 INTRODUCTION

Collaborative problem solving (CPS) has been found to have a direct impact on a variety of educational outcomes. Practicing in collaborative environments has been shown to enhance students' understanding of content in areas such as science (Hao et al. 2015), chemistry (Case et al., 2007) and creative writing (Hillocks, 1984). Participation in collaborative tasks has also been shown to improve regulation of metacognitive skills and increase engagement in knowledge construction (Stahl, 2004).

CPS consists of numerous skills, such as maintaining communication, assimilation of knowledge, and sharing resources. Moreover, these skills can be challenging to assess in any context, much less in collaboration. Efforts to assess collaboration often involve simulations, games and other team-based classroom activities which provide students opportunities to use the necessary skills and provide evidence of CPS. Humans often rate these performances, but such ratings require extensive training for raters applying limited rubrics with the potential for the rater error. Consequently, a scalable system for assessing CPS remains elusive.

Computer-based environments can help for efficient test delivery and data capture. The best environments allow participants to work on complex challenges and show mastery by successive attempts, which make assessment more reliable. Simulations and games provide high levels of engagement as well as rich task designs in pursuit of new models for measuring skills, knowledge, and abilities that are hard to address with more familiar item types.

In this paper, we describe a non-invasive, in vivo approach to assess collaborative problem solving (CPS) skills. Specifically, we focus on digital collaborative environments where behaviors indicative of CPS skills can be captured in multiple modalities including video, audio, and eye tracking recordings

and analyzed using machine learning techniques. We utilize a collaborative computer game where players/students interact through chat, video, and audio channels in a shared workspace to solve jigsaw-like tasks. Our approach uses economical and pervasive sensors such as microphones and webcams to enable real-time capture of rich data streams. This data is then analyzed with a machine learning based computational framework for evidence extraction and scalable inference of CPS skills.

The rest of this paper is organized in the following sections. Section 2 presents a theoretical model for the CPS construct. In section 3, we discuss the collaborative game used in this study and its functional details. The game's experimental design and data collection are discussed in section 4. Mapping CPS skills evidence and analyzing collected data from different sources is described in section 5 along with our machine learning based framework. Section 6 discusses preliminary results obtained using data obtained from summer 2018 pilot study participants. Finally, section 7 concludes the paper with major findings and guidance for proceeding phases and potential studies to follow.

2 CPS CONSTRUCT THEORETICAL MODEL

The Holistic Framework (HF) (Camara et al. 2015) provides a comprehensive mapping of the knowledge and skills needed for education and workplace success. The Holistic Framework includes 4 broad domains: a) Core academic skills, b) Cross-Cutting capabilities, c) Behavioral skills, and d) Education and career navigation skills (Mattern et al. 2016). In the HF, CPS is outlined as the 21st-century skills required to successfully combine communication, problem-solving, and behavioral strategies to solve a problem within a team context effectively. This study explores the assessment of these constructs through a new modality - playing a jigsaw game Crisis in Space (see section 3 for details). The collaborative problem-solving construct is segmented into two components - Team Effectiveness, and Task Effectiveness. These two components are key to the effective collaboration within a group. These two broad components are further broken down into ten functional categories as shown in Table 1, which supports the evaluation of the team's behavior and outputs, and even further into CPS skills and behaviors such as: Perspective Taking; Goodwill; Cooperation; Patience; Helpfulness; Dependability; Persistence; Accepting Differences; Fairness; Modesty (Colbow et al 2017). We extract evidence of these skills from participant discourse, as well as telemetry within the virtual space. This evidence varies based on the category: for example, Clarity is related to the number of clarifying questions that are asked, and Strategy is related to identifying the information necessary to resolve the task.

Table 1: CPS functional attributes.

Team Effectiveness	Task Effectiveness
Inclusiveness	Shared Understanding
Clarity	Strategy
Commitment	Execution
Communication	Monitoring and evaluating
Contextualization	
Goal orientation	

The collection and analysis of this data in real time will require an integrated system that seamlessly

captures the data and updates models of user ability, and this is an objective of future work and is beyond the scope of this paper. The next section describes the CPS task used in this study for collecting and mapping CPS skill evidence.

3 CPS TASK: CRISIS IN SPACE GAME

The CPS task used in this study is a collaborative game called “Crisis in Space” (CIS) and is published by LRNG (Glass Lab). Players work in pairs as shown in Figure 1 to repair the International Space Station (ISS) by solving a series of puzzles. The players are posed with a variety of “modules” each of which must be “repaired” in order to win the challenge. The players take on each of two roles throughout the game: The Operator, or person manning the space station, and the Engineer, the person at mission control with the information required for repairing the broken components. Each dyad attempts a series of five (5) missions, each with a small set of broken components in need of repair, sixteen (16) in total.

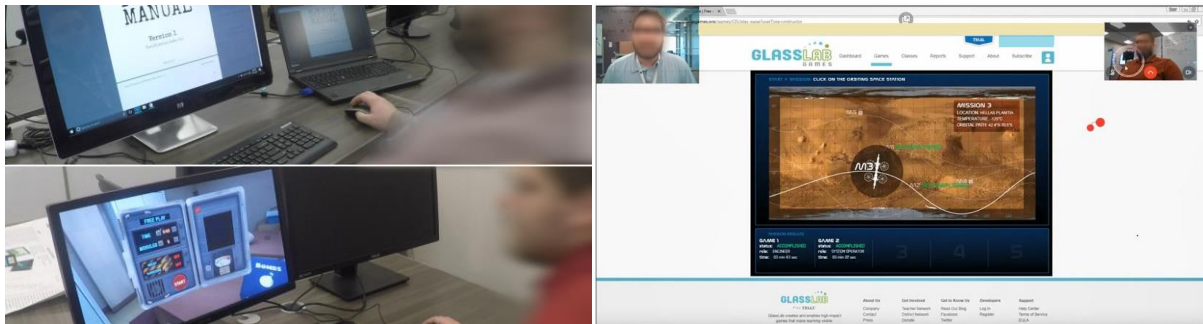


Figure 1: CPS Human-Human (HH) dyadic Crisis in Space gameplay.

Players succeed in a mission by repairing all components (two to four) within the time limit (five to eight minutes), before making three (3) mistakes along the way. The two roles have different user interfaces – the Operator is presented with a control panel for the space station containing some lights, dials, meters and other indicators while the Engineer’s screen is taken up by an instruction booklet with navigation tools. One task (named “circuit board” or “wires”) presents the Operator with a circuit board containing between three and six wires of various colors; the Operator’s booklet provides instructions on which wire needs to be cut depending on information available to the Operator, such as how many wires there are and the order of the colors. For collection and analyzing data, next section layouts game experimental design and data sources.

4 EXPERIMENTAL DESIGN AND DATA SOURCES

The experience of playing Crisis in Space for the first time – as with anything – is significantly different from playing subsequently, since learning what to do in the environment and how to interact effectively with a partner are key elements to success. One dimension of the CPS construct is Contextualization of the collaborative task, because of this, a key to the design must include multiple initial trials for each participant, each with different implications on the skills demonstrated.

4.1 Game Experimental Design

The study participants were each invited to play twice and were assigned different partners in each

case. The participants alternated roles between Operator and Engineer between each of the five missions; each participant that was assigned Operator on their first trial was assigned as the Engineer on their second and vice versa. Participants were seated in separate rooms with a laptop, monitor, keyboard, mouse, webcam and eye tracker. They communicated through audio and video Skype call with the video displayed in a small frame on the monitor. Tobii Studio Pro (Tobii) was used to simultaneously capture the monitor screen, the webcam stream, the audio stream, and eye tracking data as shown in Figure 2. The game was loaded in a Chrome browser.

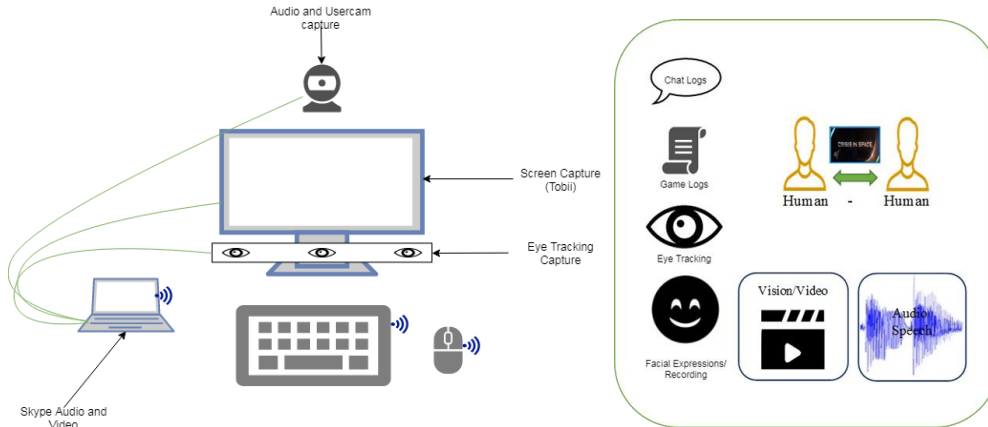


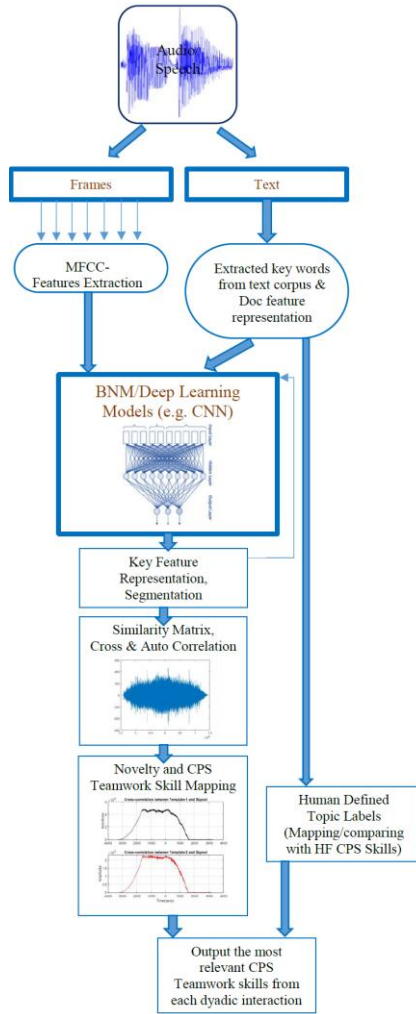
Figure 2: CPS Human-Human (HH) interaction equipment setup and telemetry data collection.

4.2 Data Sources/Data Collection

A Total of 34 participants (Female: 15, Male: 19) participated by playing the game and providing survey data. We collected game log data, user eye tracking, and user portrait video/audio as shown in Figure 2. The face video data stream was transcoded and then processed in Noldus Face Reader (Noldus) to extract frame by frame facial expression probability distributions, such as Happy, Sad and Angry. The audio data stream was submitted to Amazon Transcribe web-service to generate text transcriptions. The game server collects data related to the individual game interactions and labels each as a success or failure. The eye tracking data file also contains other useful interaction data such as mouse motion and mouse interaction. After data has been collected, we begin work on manipulating the data for the various stages of model building, in particular, the development of machine learning models. In the next section, we present our efforts towards the development of multimodal analytics framework.

5 COMPUTATIONAL FRAMEWORK

Our aim here is to design and develop a computational model based on our theoretical knowledge of the CPS construct that may enable automated evidence identification and assessment of the complex skills associated with CPS. In fact, the measurement of these constructs entails the understanding of behaviors such as the interaction between individuals as well as with the task. These multimodal data can be combined to model the user's actions and behaviors indicative of cognitive and non-cognitive processes during the interaction, thereby enabling the evaluation of complex CPS competencies (Chopade et al. 2018).



5.1 Audio Analysis Feature Extraction: Communication Skills Evidence Measurement

In this section, we further explore and present detailed steps in audio data analysis. As shown in Figures 1 and 2, the online audio-visual interface allows participants to interact while playing the CIS game. Audio data from the conversation is captured and processed to extract text, and to distill low-level (machine) features. As shown in Figure 3, we first used MATLAB signal processing and audio analysis toolbox (Mathworks 2018, MATLAB STAT & ML) for low-level audio features such as Rhythm, Timbre, Pitch, and Tonality. These features are used to make inferences about states of engagement and emotional states of the user.

5.2 Audio Analysis: Sentiment Analysis

Using Amazon transcribe we extracted transcripts from the audio files, then performed sentiment analysis by extracting positive and negative words as well as keywords from the text. Some human labeling based on task-specific words was also leveraged. These positive and negative words were then extracted and counted for each dyad to measure positivity and goodwill in the group interaction.

Figure 3: Bayesian network model (BNM)-Deep Learning based Audio data analysis framework.

6 PRELIMINARY RESULTS

In this section, we present and discuss some of the preliminary results based on audio analysis for pilot study participants. Out of 34 CIS game play participants, we ran the audio analysis for 6 groups (12 participants). These were selected based on the number of successfully completed game missions. Using the analytics framework discussed in section 5, we carry out pre-processing, audio feature extraction and Natural Language Processing (NLP) Sentiment analysis for audio text data. We



Figure 4: Audio text NLP analysis for CIS gameplay – positive and negative word cloud.

trained a Latent Dirichlet allocation (LDA) model and Latent semantic analysis (LSA) model (Mathworks 2018), which extracts the semantic meaning from the words. Sentiment analysis from the list of common positive and negative words is shown in Figure 4 as a positive and negative word cloud. These features lay the initial foundation for using a Bayes net to map CPS teamwork skills evidence from dyadic CIS gameplay interactions.

7 CONCLUSIONS AND FUTURE WORK

Developing a framework for identifying evidence of CPS will provide a general, replicable approach for drawing inferences that support cognitive research using educational learning environments and, potentially, allow for the development of valid, reliable measurement of constructs of interest to educational practitioners and administrators that are difficult to measure. Additionally, the development of such a framework would provide the ability to examine the impact of CPS on educational outcomes such as the scientific inquiry skills and argumentation skills measured in the computerized educational environment used in this study. In our future work, we will use these extracted features for mapping and analyzing CPS teamwork skills evidence. We will compare CPS skills mapping obtained from humans along with CPS teamwork skills received from our machine learning process. CPS teamwork skills such as knowledge assimilation, positive communication, and resource sharing will be further investigated. Results obtained in this stage will be significant for proceeding phases and potential studies to follow.

ACKNOWLEDGMENT

The authors are thankful and acknowledge the support from LRNG, Glass Lab for this work. The authors would like to thank Andrew Cantine-Communications and Publications Manager, ACTNext for editing this work. We are also thankful to ACT, Inc. for their ongoing support as this paper took shape.

REFERENCES

- Camara, W., O'Connor, R., Mattern, K., & Hanson, M. (2015). *Beyond academics: A Holistic Framework for enhancing education and workplace success*. ACT Research Report Series, 2015 (4).
- Case, E., Stevens, R., & Cooper, M. (2007). Is collaborative grouping an effective instructional strategy? *Journal of College Science Teaching*, 6, 42-47.
- Chopade, P., Khan, S. M., Edwards, D., & von Davier, A. (2018). *Machine Learning for Efficient Assessment and Prediction of Human Performance in Collaborative Learning Environments*. IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, MA, USA, 1-6.
- Colbow, A., Latino, C.A., Way, J. D., Casillas, A., & McKinnis, T. (2017). *The ACT Behavioral Skills Framework: How does it compare to other behavioral models?* ACT Research Report Series, 2017 (6).
- GlassLab, Inc. (2018). <https://www.glasslabgames.org/>, LRNG <https://www.lrng.org/>
- Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). *Assessing Collaborative Problem Solving with Simulation Based Tasks*. International Conference on Computer Supported Collaborative Learning 2015.
- Hillocks, G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education*, 93(1), 133-170.
- Mathworks Inc. MATLAB, (2018).
- Mattern, K., Allen, J., & Camara, W. (2016, Fall). Thoughts on a multidimensional middle school index of college readiness, *Educational Measurement: Issues and Practice*, 35(3), 30-34.
- Noldus (2018). <https://www.noldus.com/>
- Stahl, G. (2004). Building collaborative knowing: Elements of a social theory of CSCL. In J.-W. Strijbos, P. A. Kirschner, & R. L. Martens (Eds.), *What we know about CSCL and implementing it in higher education*, (pp. 53-85). Boston, MA: Kluwer Academic Publishers.
- Tobii Eye Tracking (2018). <https://www.tobii.com/>

Evaluating a Learning Analytics Research Community: A Framework to Advance Cultural Change

Linda Shepard, George Rehrey, Dennis Groth and Amberly Reynolds

Indiana University
lshepard@indiana.edu

ABSTRACT: Institutions implementing innovative LA tools encounter cultural barriers that limit full adoption. This has led to an emerging focus on models for cultural change that aim to mitigate these known obstacles. Transforming campus culture is an arduous task that may take 7-10 years to fully mature. In this **practitioner paper** we suggest an evaluation process, measuring incremental growth toward this long-term goal. We adopt a well-known logic model to consider short, medium and long-term LA outcomes. We frame our evaluation strategy around a model of institutional change that we define with a) shifts in knowledge b) shifts in behaviors, and c) the final goal of shifts in cultural norms.

Keywords: Network Improvement Communities, Change Theories, Evaluation Framework
Faculty Engagement

1 INTRODUCTION

Institutions of higher learning have rapidly adopted various types of Learning Analytics (LA) tools and techniques to improve educational environments with limited success. Such tools have the potential for significantly altering the ways we understand teaching, learning, and student success in higher education, but institutional wide acceptance of new practices takes significant effort and often meets skeptical resistance from faculty (Tagg, 2012), staff and administrators alike. According to the National Research Council (2012), the actual adoption of innovative practices already proven to enhance undergraduate STEM education (Freeman et al., 2014; Macfayden 2014; Fairweather 2009) remains low, while it has become increasingly apparent that networked approaches provide greater potential for widespread institutional change (Williams et al., 2014).

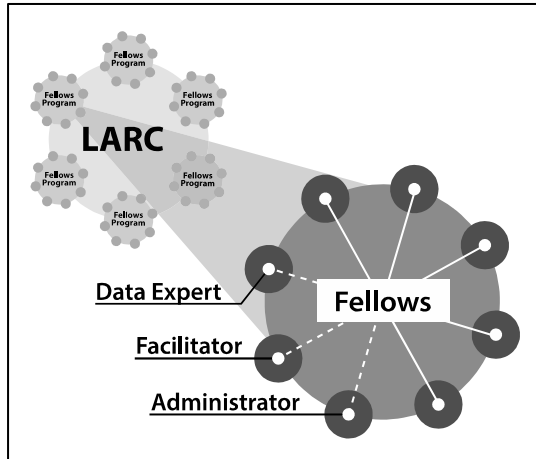
2 OUR CHANGE MODEL

A Network Improvement Community (NIC) (Bryk et al., 2011) is a method for linking institutions in support of the adoption of new innovations. They focus on common problems and test hypotheses for improvement at multiple sites and use a combination of academic and practitioner research to analyze the local context around the issue they are trying to solve. With its origins in the healthcare industry, NICs have also been recognized as effective in community colleges, but as of yet have not been wholeheartedly embraced within higher education, nor are there many examples for evaluating their effectiveness in driving institutional transformations.

A NIC improves the possibility of sustained adoption of new practices, but not without careful planning and ongoing program monitoring. Such formative evaluation should occur early and often, learning from both mistakes and successes, inclusive of all participants, while making the necessary adjustments along the way to ensure program goals are achieved. Such is the case here, where we suggest a framework for planning and monitoring the long-term success of our Learning Analytics Research Community (LARC), which came about through participation in the Bay View Alliance (BVA).

The BVA is comprised of 10 research institutions from Canada and US with an overall goal of fostering change in the teaching and learning cultures within STEM departments.

Informed by the NIC approach, the BVA seeds and supports institutional change efforts by fostering the formation of Research Action Clusters (RAC), where participants across institutions share results,



and build upon individual and collective successes and failures. At the same time, interventions and innovations are tested within the culture of each institution, acknowledging that all change must be sensitive to local context. Thus LARC, which is one of the five different RACs within the BVA, is preparing to test the viability of multiple LA Fellows program aimed at improving STEM education. The collaborating institutions include Indiana University Bloomington (Lead), University of California Davis, University of Kansas, University of British Columbia, University of Saskatchewan, and Queens University (Figure 1).

Figure 1: LARC is a community of Fellows programs situated at 6 institutions

3 OUR LOCAL LA FELLOWS PROGRAM

The first LA Fellows program originated at one of the participating institutions of the BVA (Rehrey et al., 2018) and is a local community of administrators, faculty, and staff (See Figure 2). The program's purpose is to establish an evidence-informed culture in which departments make ongoing use of LA to further student success. We envision a time when LA will guide how departments plan and make decisions that may impact teaching, learning and the student experience, which includes academic advising, course and program design, and institutional-wide resource allocations. We recognize that the successful adoption of new ideas and practices in higher education quite often hinges upon a department's culture, which in turn influences, and is influenced by, faculty beliefs and behaviors. So, we employ a top-down, bottom-up and middle-out change model (AAU, 2017) (Corbo et al., 2016) (Rehrey, et al., 2018), getting faculty involved from the very start of the program. With the full support of top administrators, along with the engagement of research faculty, we anticipate a cultural shift will eventually occur within departments and how they view the role of LA in improving student success.

In the program, faculty LA Fellows engage in their own scholarly research, using institutional data to gain knowledge about students in their courses, curriculum, programs and schools to advance our institution's strategic plan and commitment to institutional improvement. Participating faculty gain skills necessary to use institutional data to make inquiry-informed decisions about their students, generally focused on four broad categories: student choice, demographics, preparation, and performance. Often, these factors overlap and are interrelated within any given Fellows research project. For further discussions about the program and implementation strategies please refer to Pardo, et al. 2018 and Rehrey et al. 2019.

As part of LARC, six institutions are in various stages of creating their own version of the LA Fellows program on their campuses. This allows us to share strategies for engaging faculty, share results from faculty research projects, consider the nuances of local context for scaling up and share strengths and challenges of our implementation strategies. Through this multi-institutional community, we support the work and create sustainable programs on our individual campuses.

4 THE EVALUATION MODEL

Institutions implementing innovations are faced with the unseen, yet ever-present, obstacle of cultural values, beliefs and behaviors. To evaluate the adoption (receptiveness) of these innovations we propose intentionally capturing evidence that reflects the various stages of an evolving culture. This means collecting more data points than the number of log-ins, or measuring time on task or counting the number of clicks, even though that may be completely appropriate for certain initiatives. We suggest collecting evidence that will ultimately lead to measuring the shift in campus norms, thereby indicating how a school values the data-guided continuous improvement model. We offer a framework to guide thoughtful, deliberate dialog concerning measurements for the elusive concept of cultural change. With this model, we seek metrics that can detect changes or provide evidence for a nascent cultural shift. We share our program goals, processes and activities in the spirit of generalizing to a broader dialog about evaluating cultural transformations.

We take advantage of McLaughlin and Jordan's logic model (1999) to describe the short-term, mid-term and long-term outcomes of our efforts, and how they align with program's inputs, strategies, and outputs (Figure 3). Outcomes indicate what should be measured, and sometimes even how to measure them, to best determine if the program is meeting its goals. Over the duration of LA projects, outcomes can be evaluated in phases. During the first phase, short-term outcomes measure the change in the knowledge, skills, and attitudes of our LA Fellows. During the second phase, mid-term outcomes measure the change in our Fellows' behavior. Finally, long-term outcomes measure the change in their norms and the overall impact of the program on the department's culture. In our local programs are currently measuring both short and mid-term outcomes.

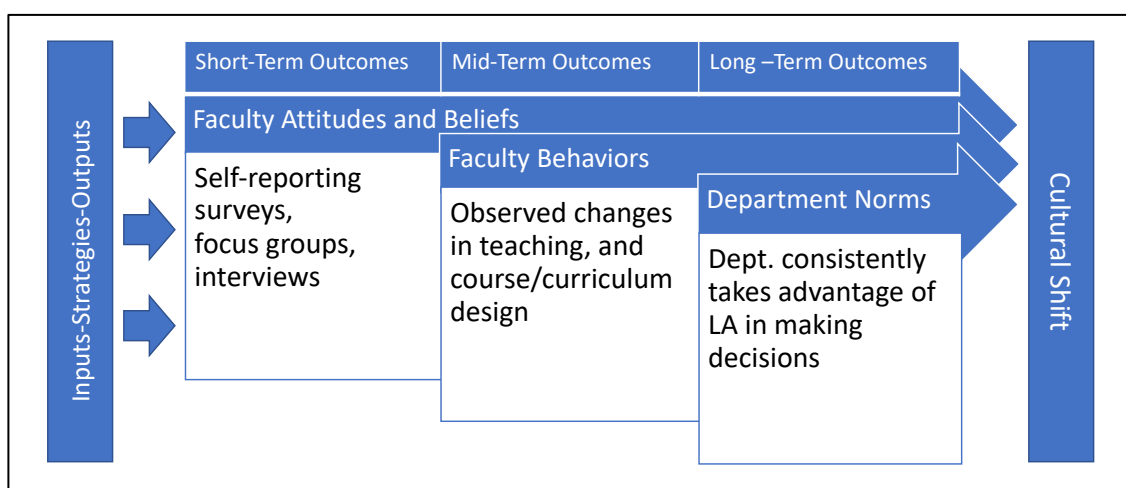


Figure 3: An Evaluation Framework for the Sustained Adoption of Learning Analytics

This evaluation plan shares similar characteristics to the ADKAR model of change (citation) We start by bringing awareness to individuals, before supporting their desire to make changes through community, which leads to changes in their teaching practices. Those changes are then reinforced as the departments use of LA in making decisions and changes.

4.1 Inputs/Strategies/Outputs

All aspects of a logic model align with the primary goal of the program, which is to establish an inclusive data-guided culture at our collaborating institutions. In general, inputs are the resources that support our work, while the strategies are considered the activities that drive the program outputs or products. The **Inputs** of our evaluation model include the 6 research institutions of LARC, as well as all partner campuses of the BVA, a champion of our work. Then, on each individual campus additional inputs include administrators, faculty, and staff from various departments, along with data stewards and some Institutional Research (IR) resources (Figure 1). The **Strategies** employed to shape program outputs include financial incentives the Fellows receive in the form of research stipends, along with the social interactions and rewards that result from belonging to a community of researchers. Providing LA data in a useable form to the Fellows, along with research support and statistical expertise are also considered strategies. Other implementation activities involve faculty dissemination of their research on their own campuses, at BVA annual summits, and at national and international conferences. **Outputs** include the individual project proposals and summary reports submitted by faculty, course and curricular transformations resulting from their projects, and the publication of white papers, and peer-reviewed papers describing these results. Now we move on to describe shared outcomes as they relate to our collective efforts.

4.2 Outcomes:

Short-term outcomes (1-5 years) articulate the knowledge, skills, and attitudes of the Fellows. These outcomes include the added value faculty place on LA, the knowledge they gain in working with the data, and the insights they gain about their students and programs. We also track how this new knowledge is shared locally, in professional organizations and with peers. Attitudes are measured largely by self-reports. During the last reporting period, 79% of the faculty valued the importance of using LA to further student success more since joining the Fellows program.

With **Midterm outcomes** (2-7 years) Fellows make changes to their course, curriculum or program, engage in further exploration for decision-making and encourage their colleagues, both within their program and in other programs to engage in the use of LA. Deans/chairs encourage others to use data. Currently, nearly 90% of the LA Fellows say that they have, or will make, teaching and learning changes. We see a sustainable community emerging, as 15 fellows are repeat participants, and departments continue to increase their representation in the community. One of our departments began with just one faculty member, but now has 3 other faculty conducting LA research, along with the department chair. Survey results also indicate an increase in knowledge about Fellows' projects, as conversations are taking place in departmental meetings and across campus. **A more detailed discussion about outcome results is forthcoming in the special edition of JLA 2019** (invited paper).

Since the goal of the program is to influence departmental culture, indication of **Long-term outcomes** (5 – 7 years) will be achieved when faculty consider student success their responsibility, both

individually and collectively, departments establish a culture where student success is an ongoing concern, and LA is consistently used in planning, implementing and evaluating courses and curricula.

With this framework, we offer the opportunity to evaluate the effectiveness of an implementation strategy to change institutional norms at large or small scale. Each innovation brings a set of resources and activities to encourage adoption; and depending on the project various products are expected. When it comes to outcomes, we initially expect to see awareness and knowledge of new innovations, being mindful of marginalized participants. Finding ways to measure the campus knowledge of those innovations will be unique. For example, if an innovative LMS tool is released to instructors to enhance student performance, we may first want to understand the depth of knowledge faculty have about the tool. The instructor may a) read an email, b) use the tool as a pilot in their course or c) provide feedback to the development team to enhance future use. All represent awareness of the tool. Midterm outcomes speak to behaviors or application of the tool, as faculty pilot or use the tool in their courses. Department norms are realized once the tool is widely accepted and used in a majority of the department's courses.

5 LIMITATIONS

One of the limitations of our models is that we are not evaluating the quality of the individual faculty research projects. Instead we are interested in whether the program is achieving its goal, which is to establish a data-informed culture throughout the institution. In other words, when depts or teachers are attempting solve problems, they turn to the data before making decisions and changes. Another limitation is that we did not collect baseline data prior to starting the Fellows program in 2015. We will address this in future programing by asking all Fellows to complete a pre-test prior to providing access to the LA data. We also realize that this framework may not capture unintended outcomes that may influence the success of the program.

For example, a recent development on one of our campuses is an organically formed community of Fellows who are requesting funding for the formation of an Educational Data Science Program, a new interdisciplinary field of study that would advance this type of work and share knowledge more broadly with relevant disciplinary communities.

6 CONCLUSION

Innovations in LA are rapidly emerging with great promise. We track their effectiveness by measuring usage or other indicators of student success, retention, graduation, and performance of students. We propose that culture change is also a goal shared across many LA activities on many campuses. Small-scale and large-scale implementations alike, share in the challenge of campus adoption and a campus mindset for data-guided planning, actions and decision making. Here, we focus on developing formative and summative evaluation processes that indicate readiness for full LA adoption, a precursor to an environment that seeks equitable evidence-based solutions to problems. By deliberately considering the phases of change, we measure our incremental progress as we work with faculty and departments towards changing institutional norms. We offer a logic model framework, as a guide for thoughtful intentional measurements of cultural change as colleges and universities adopt LA to further student success.

REFERENCES

- Association of American Universities. (2017). *Progress toward achieving systemic change: a five-year report on the AAU undergraduate STEM education initiative*. Retrieved from <https://www.aau.edu/progress-toward-achieving-systemic-change>
- Beach, Andrea L. & C., Milton D. (2009). The impact of faculty learning communities on teaching and learning. *Learning Communities Library*, 1(1), 7–27.
- Corbo, J. C., Reinholz, D. L., Dancy, M. H., Deetz, S., & Finkelstein, N. (2016). Framework for transforming departmental culture to support educational innovation. *Physical Review Physics Education Research*, 12(1), 010113.
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>
- Hiatt, J. (2016). *ADKAR: A model for change in business, government and our community*. Prosci Learning Center Publications.
- Macfadyen, L. P., Dawson, S., Pardo, A., & Gašević, D. (2014). Embracing big data in complex educational systems: the learning analytics imperative and the policy challenge. *Research and Practice in Assessment*, 9(2), 17–28.
- Macfadyen, L. P., Steinwachs, Rehrey, G., Shepard, L., Greer, J., ... Molinaro, M. (2017). Developing institutional learning analytics “communities of transformation” to support student success (pp. 498–499). Presented at the Learning Analytics and Knowledge Conference, Vancouver, BC: ACM Press. <https://doi.org/10.1145/3027385.3029426>
- McLaughlin, J. A., & Jordan, G. B. (1999). Logic models: a tool for telling your programs performance story. *Evaluation and Program Planning*, 22(1), 65–72. [https://doi.org/10.1016/S0149-7189\(98\)00042-1](https://doi.org/10.1016/S0149-7189(98)00042-1)
- Pardo, A., Bartimote, K., Lynch, G., Buckingham Shum, S., Ferguson, R., Merceron, A., & Ochoa, X. (Eds.). *Implementation of a learning analytics program* (2018). Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge. Sydney, Australia: Society for Learning Analytics Research.
- Rehrey, G., Groth, D., Shepard, L., & Hostetter, C. (2019). The scholarship of teaching, learning and student success: Big data and the landscape of new opportunities. In J. Friberg & K. McKinney (Eds.), *Conducting and Applying SoTL Beyond the Individual Classroom Level*. Indiana University Press.
- Singer, S., & Smith, K. A. (2013). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. *Journal of Engineering Education*, 102(4), 468–471.
- Tagg, J. (2012). Why does the faculty resist change? *Change: The Magazine of Higher Learning*, 44(1), 6–15. <https://doi.org/10.1080/00091383.2012.635987>
- Williams, A. L., Verwoord, R., Beery, T. A., Dalton, H., McKinnon, J., Strickland, K., ... Poole, G. (2013). The power of social networks: a model for weaving the scholarship of teaching and learning into institutional culture. *Teaching and Learning Inquiry: The ISSOTL Journal*, 1(2), 49–62.

Evaluating Preparatory Writing and Writing Placement at a Large Public University

Sattik Ghosh

UC Davis Center for Educational Effectiveness
stkghosh@ucdavis.edu

Emily Watkins

UC Davis Center for Educational Effectiveness
emwatkins@ucdavis.edu

Meryl Motika

UC Davis Center for Educational Effectiveness
mimotika@ucdavis.edu

ABSTRACT: This paper describes a quantitative evaluation of preparatory writing at a large American university. We used administrative data to study the validity of a placement exam, the effect of a preparatory course on outcomes in later writing courses, and the effect of placing into the preparatory course on retention in the university. We controlled for selection into the preparatory course through the use of differences in when students took the preparatory course and propensity score weights based on prior academic records. We found that the placement exam was less predictive of later grades in writing-intensive courses than SAT, AP, and high school grades; the preparatory course had a slightly positive but insignificant effect on students' grades in later writing-intensive courses; and placement into the preparatory writing course did not have a measurable effect on attrition from the university. Exploiting the timing of writing-intensive general education courses to identify effects of the preparatory course is shown to be valuable both for evaluation and to validate the use of propensity weights. Propensity weighting is also shown to be useful for evaluating course outcomes in the context of a weak placement exam.

Keywords: preparatory courses; writing; placement; evaluation

1 BACKGROUND

In Fall 2017, faculty overseeing the writing program at a large, public university in the United States asked us to provide statistics relating to their preparatory writing course. In particular, they were concerned about a perception of unfairness in placement and course grades. All first-year students were required to show proficiency in English writing either by passing a placement exam or by earning a sufficiently high score on a standardized exam such as AP English. About 30% typically did not meet this requirement and were therefore required to pass the preparatory course within three quarters of study. As a result, any problems with the course would have affected a substantial portion of the student body.

We found that the average grade given by different instructors in one quarter varied from about 1.1 to 2.9 on a 4-point scale. Scores were somewhat lower for versions of the course intended for English language learners, but variation between instructors for the same course dwarfed the differences between versions. As shown in Figure 1, differences in average grades by class were not substantially

explained by differences in SAT scores. Since SAT scores are typically correlated with academic performance, it seemed unlikely that variation in either preparation or study habits was driving the variation in average grades. Many of these average grades, let alone individual grades, were not sufficient to pass the course.

These observations led to the three questions we address in this study.

1. Did the exam place students appropriately, meaning that students who were required to take preparatory writing lacked writing skills necessary for college?
2. Did students who completed preparatory writing seem to benefit from it in terms of later grades in writing courses?
3. How did placement into preparatory writing affect college completion?

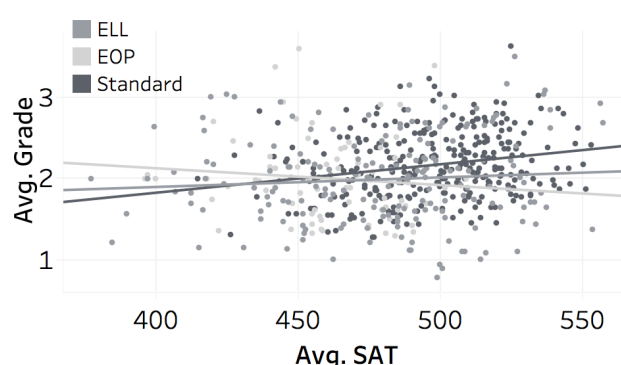


Figure 1: Preparatory writing class average grades by average SAT reading scores

Answering these questions required us to find comparison groups for students placed into the preparatory course. The placement exam was scored in a way that precluded a regression discontinuity design. However, we uncovered two features of the data that made the comparisons possible. First, while the preparatory course was a pre-requisite for the main writing series, it was not required for most of the courses categorized as ‘writing-intensive’ for the university’s general education requirements. Many students enrolled in these writing-intensive courses both before and after taking the preparatory course. Second, we found that the placement system was sufficiently noisy to allow the use of propensity score weights for questions 2 and 3 (Rhodes 2010; Wooldridge 2007).

This study follows a literature on placement exams in which it has become clear that despite variation between high schools, high school grades are better than these types of exams at predicting success in college writing (Barnett et al. 2018, Scott-Clayton 2012, Scott-Clayton et al. 2014, Willett 2013). Quantitative evaluations of preparatory writing courses seem less common. Multiple studies found reducing the number of required courses helped community college students transfer to a four-year school, but the value of individual courses was not separately identified (Bailey, Jeong, and Cho 2010; Rodriguez et al. 2018). Aiken et al. (1998) found no effect of a preparatory course in an experiment with a pre-post test similar to a placement exam. Southard and Clay (2004) found positive effects of a developmental writing course in a subsequent course using a simple comparison. Our study contributes robust methods for managing selection bias in a non-experimental design.

2 DATA AND METHODS

We used registrar data for all students admitted as freshmen between Fall 2010 and Fall 2017 who scored a 6 (highest failing grade) or an 8 (lowest passing grade) on the writing placement exam, and whose first writing-intensive course did not have a pre-requisite. Placement exam data were only available for in-

state students. Background characteristics were self-reported in admissions forms or derived from high school transcripts. We excluded students with missing data for language spoken at home, international status, SAT or ACT score, and high school GPA. Students with missing indicators for low-income status, first-generation status, or under-represented minority status were assumed not to belong to the relevant group. All models of course grades included an indicator for whether the course was taken in the student's first quarter. Controlled regressions included gender, SAT score, high school GPA, language spoken at home, first generation status, low income status, under-represented minority status, AP exams reported and passed, and international student status where relevant.

Methods for Research Question 1: Placement Exam

We assessed the placement exam using the population of students who took a writing-intensive course before beginning the main writing program and before taking the preparatory course. Specifically, we compared the grades of students who just passed versus just failed the placement exam in their first writing-intensive course before taking one of the main writing program courses or the preparatory course. In addition to a simple comparison of these two groups of students, we investigated whether admissions information predicted writing ability as well as or better than the placement exam by including SAT, GPA, and AP scores as control variables.

Methods for Research Question 2: Preparatory Course

We used two different approaches to estimate the impact of the preparatory course. Our first approach relied again on the timing of the course. If it taught useful skills, then students placed into the preparatory course who completed it before a writing-intensive general course could be expected to outperform students who had not yet taken the preparatory course. This approach relied on the assumption that there was no relevant difference between students who chose to register for the preparatory course first and those who registered for another writing-intensive course first.

We also estimated a propensity score-weighted model of grades in writing-intensive courses. The weights were derived from a logistic regression of passing the placement exam on the standard set of controls for background characteristics except AP exam results, which perfectly determined placement. AP result was instead included as a control in the final regression. Assuming the placement exam did not capture important information about students' academic ability independent of background characteristics, this method simulates random assignment into the preparatory course. The assumption seems reasonable given the results described in section 3.1.

Methods for Research Question 3: College Completion

We used the same propensity score weighting method described above to compare attrition rates between students who did versus did not have to take preparatory writing. This model included dummies for admit year in addition to the controls used in the previous model. As above, this method relies on the assumption that placement was random after controlling for background characteristics.

3 RESULTS

3.1 Placement exam

Table 1 shows that absent any controls other than an indicator for classes taken in the student's first quarter of college, passing the writing placement exam was associated with grades about 0.17 points higher in the first writing-intensive course – about half the difference between a B and B-. This was already a fairly limited relationship, and including controls for earning at least a 3 in AP English, SAT writing score, and high school GPA reduced that correlation to effectively zero. In other words, the grade was much better explained by AP English, SAT writing, and high school GPA than by the placement exam.

Table 1: OLS Regressions of grade in first writing GE course on placement exam score.

	Final grade Coefficients	
	I	II
Passed placement exam	0.162*** (0.037)	0.061 (0.037)
Controls for AP taken and score, SAT writing, and high school GPA	No	Yes

N=2,696. Sample: students who took a writing-intensive course before any University writing department course including the preparatory writing course and earned a numeric grade in the course. Control for whether the course was taken in the student's first quarter at the University included in all models. Standard errors in parentheses. *** p< 0.001

3.2 Effect of preparatory writing course on later writing grades

Table 2 shows the results of a regression of outcomes in students' first writing-intensive GE course on whether they took the preparatory course concurrently with or after the course, with the baseline being students who took the preparatory course before their first writing-intensive GE. While the negative coefficients suggest that there was some correlation between taking the preparatory course and a higher grade, the relationship was so noisy that despite such a large sample size only one coefficient is statistically significant at even the 5% level. Including controls for academic and personal background eliminated this result, showing that background factors explained the writing grade better than participation in the preparatory course.

**Table 2:
Regression of grade in first writing GE course on timing of preparatory course enrollment.**

	I	II
Concurrent enrollment in preparatory course	-0.140 (0.099)	-0.102 (0.095)
Preparatory course not taken until after the GE course	-0.238* (0.104)	-0.155 (0.100)
Standard controls	No	Yes

N=2,351. Baseline: students who took the preparatory writing course prior to enrolling in their first writing GE. Sample: students who placed into preparatory course, took a writing-intensive course before taking a University writing department course, and earned a numeric grade in the course. Control for whether the course was taken in the student's first quarter at the University included in all models. Standard errors in parentheses. * p< 0.05.

This result was replicated using the propensity-weighted comparison between students who were versus were not required to take the preparatory course, as shown in Table 3 Model I. The pseudo- R^2 for the logistic regression used to generate weights was 0.14 and no control was significantly different between the matched and the unmatched group in the weighted sample except whether the student reported an AP score – this was included as a control in the weighted regression. As in the previous test, the possible effect of the preparatory writing course could not be distinguished from random variation in grades.

3.3 Preparatory writing and attrition

Table 3 Model II shows the result of regressing attrition on whether the student was exempt from taking the preparatory writing course, using propensity score weighting to create equivalent samples of students who were versus were not required to take the course. The regression used to generate these weights had a pseudo- R^2 of 0.13 and balance tests revealed no significant differences in the weighted samples. We found that placement into the course did not measurably predict attrition.

Table 3: Propensity score weighted regressions

	I	II
	Grade in first writing GE	Attrition
Exempt from preparatory course	0.069 (0.050)	-0.007 (0.426)
N	3,441	7,785

Baseline: students who were required to take the preparatory course. Sample I: students who took a writing-intensive course before any University writing department course. Sample II: all students in the dataset. Controls as described in methods. Linear probability model shown for II; logit results were equivalent. Standard errors in parentheses. No result was significant at the 5% level.

4 DISCUSSION

Our results show that although the placement exam was related to ability to succeed in a writing course, it did not capture substantial new information beyond what was accounted for by SAT, GPA, and AP scores. Further, there was no apparent effect of participation in the preparatory course on grades in later writing-intensive courses. Robustness checks (not shown) of sections 3.1 and 3.2 using only courses in which writing assignments were known to be a large proportion of the grade produced similar results.

Thanks to the randomness in placement, the propensity score weighting model was useful for comparing students who were versus were not placed in preparatory writing to estimate the value of the preparatory course in improving student outcomes and the effect on attrition. The appropriateness of this technique was supported by the evidence of random placement, acceptable predictive power of the selection models, clean balance tests, and similar results between two alternative methods of testing course outcomes. Another propensity model intended to estimate the effect on attrition of failing the preparatory course was excluded from this report due to poor balance between the weighted samples.

For analysts we suggest the following lessons from our experience:

- 1) While many uses of propensity score weighting may fail to resolve or even exacerbate selection bias, in a situation such as this one where selection is based on an observable exam score it may be appropriate. Balance tests remain vital to ensure the weights have the intended effect.
- 2) Variation in timing for taking courses provides an excellent way to estimate the effect of related courses. We were able to estimate the effects of the exam and writing course because some students took the preparatory course after their first writing-intensive general education course.

- Aiken, L., West, S., Schwalm, D., Carroll, J., Hsiung, S. (1998) Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation. *Evaluation Review* 22(2), 207-244. <https://doi.org/10.1177/0193841X9802200203>
- Bailey, T., Jeong, D., & Cho, S. (2010). Referral, Enrollment, and Completion in Developmental Education Sequences in Community Colleges. *Economics of Education Review* 29 (2), 255–70. <https://doi.org/10.1016/j.econedurev.2009.09.002>
- Barnett, E., Bergman, P., Kopko, E., Reddy, V., Belfield, C., Roy, S., Cullinan, D. (2018). *Multiple Measures Placement Using Data Analytics: An implementation and early impacts report*. Center for the Analysis of Postsecondary Readiness. https://www.mdrc.org/sites/default/files/CAPR_Multiple_Measures_Assessment_implementation_report_final.pdf
- Rhodes, W. (2010). Heterogeneous Treatment Effects: What does a regression estimate? *Evaluation Review*, 34(4), 334-361. <https://doi.org/10.1177/0193841X10372890>
- Rodriguez, O., Cuellar Mejia, M., Johnson, H. (2018, August) *Remedial Education Reforms and California's Community Colleges: Early evidence on placement and curricular reforms*. Public Policy Institute of California. <https://www.ppic.org/wp-content/uploads/remedial-education-reforms-at-californias-community-colleges-august-2018.pdf>
- Scott-Clayton, J. (2012). *Do High-Stakes Placement Exams Predict College Success? Working Paper No. 41*. Columbia University Teachers College, Community College Research Center. <https://ccrc.tc.columbia.edu/publications/high-stakes-placement-exams-predict.html>
- Scott-Clayton, J, Crosta, P., & Belfield, C. (2014). Improving the Targeting of Treatment: Evidence from College Remediation. *Educational Evaluation and Policy Analysis*, 36, 371–393. <https://doi.org/10.3102/0162373713517935>
- Southard, A., & Clay, J. (2004) Measuring the Effectiveness of Developmental Writing Courses. *Community College Review* 32(2). 39-50. <https://doi.org/10.1177/009155210403200203>
- Willett, T. (2013). *Student Transcript-Enhanced Placement Study (STEPS) Technical Report*. The Research and Planning Group for California Community Colleges. <https://files.eric.ed.gov/fulltext/ED577267.pdf>
- Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141:1281-301. <https://doi.org/10.1016/j.jeconom.2007.02.002>

Developing an English Learner Corpus for Materials Creation and Evaluation

Amanda Hilliard

Arizona State University

Amanda.D.Hilliard@asu.edu

ABSTRACT: This presentation will describe the development and use of a learner corpus for materials development in an advanced English as a Second Language (ESL) writing course. Using corpus software, student essays were evaluated for frequency and errors of transition words in four separate genres, and student examples were used to create classroom materials to target underused and misused transitions. After using these classroom materials, students more frequently produced the targeted transitions but still made some mistakes. This raises issues for further materials development, showing that a learner corpus can inform a continual cycle of evaluation and implementation for more successful student outcomes.

Keywords: learner corpus, ESL, curriculum development, action research, writing instruction

1 INTRODUCTION AND PROBLEM

A corpus is “a collection of naturally occurring language texts in electronic form, selected according to external criteria to represent as far, as possible, a language or language variety as a source of data for linguistic research” (Sinclair, 1991, p. 171). Over the last few decades, developments in corpus linguistics, which uses computer software to analyze language use in these corpora, have shifted the focus from more theoretical ideas based on researchers’ intuition to analyses of large databases of language for frequency and authentic language use. Yet, practitioners seldom apply this research directly to the language classroom. Indeed, Römer (2010) claims that “The practice of ELT (English Language Teaching) to date . . . seems to be largely unaffected by the advances of corpus research, and comparatively few teachers and learners know about the availability of useful resources” or use corpus tools themselves (p. 18). This is unfortunate as the use of corpora has been shown to have a positive influence on the teaching of vocabulary (Soruc & Tekin, 2017), grammar (O’Donnell, 2012), speaking (Jones, Byrne, & Halenko, 2017; Mukherjee, 2009), and writing (Hasselgard, 2009), as well as the development of classroom materials and assessments (Gilquin, Granger, & Paquot, 2007; Herriman & Aronsson, 2009; Seidlhofer, 2002).

One particular type of corpus, a learner corpus, which is a systematic computerized collection of written or spoken texts produced by language learners, can be a powerful tool for informing curriculum and materials development for language classes. Learner corpora can be used to identify particularly difficult items for students, uncover insights into the language learning process, and make comparisons between students’ language use and that of native speakers (Nesselhauf, 2004). For example, Crosthwaite (2013) examined the Cambridge Learner Corpus to research the development of learners’ language skills at various proficiency levels of the Common European Framework as well as to compare the performances of Korean and Chinese test takers. Jones, Byrne, and Halenko (2017) used a learner corpus to analyze learners’ linguistic, strategic, discourse, and pragmatic competence, providing a number of implications for both researchers and teachers, such as the need to focus on core vocabulary words and to teach chunks for

communicative competence. Finally, O'Donnell (2012) describes the TREACLE project, which built a learner corpus from a collection of student texts and then analysed the texts to redesign English grammar curricula in Spanish university contexts.

In this study, the goal was to develop a collection of learner texts that could then be analyzed to determine how students were applying transition words from their textbooks to writing essays in four separate genres in order to develop more effective teaching materials. This was particularly important as the textbooks lacked information on frequency of each transition word, guidance for teachers on which items were most difficult for students, or practice exercises for student production.

In addition to addressing shortcomings in textbook materials, the learner corpus was also a way to analyze students' writing skills, particularly their use of transition words for developing cohesion. While it is relatively easy for writing instructors to assess individual students' grammar mistakes, understanding how students are building cohesion and analysing frequency of specific words within all the students' texts is much more difficult. Corpus software can address this problem by helping instructors understand aspects of student writing through analysis of large sets of data that would be impossible for teachers to do manually.

Finally, developing academic writing skills can be particularly demanding for international students, yet it is essential for inclusion in the American university system. By analysing and addressing students' shortcomings, the use of corpus tools can help better prepare students for a successful transition from ESL to regular university classes.

2 METHODOLOGY AND IMPLEMENTATION

For this study, a learner corpus was created from student essays submitted for an advanced ESL writing course at a university level Intensive English Program (IEP). The corpus comprised 338 essays from four separate genres and four different IEP instructors, for a total of approximately 220,000 words.

The study addressed the following research questions:

1. How frequently did students use the transition words from their writing textbook in their essays? How does this frequency compare to the frequency of these words in the academic subcorpus of the Corpus of Contemporary American English (COCA)?
2. What kinds of errors did the learners make using the transition words?
3. What were the effects of developing and using teaching materials from the analysis of the learner corpus on subsequent students' use of the transition words in their essays? What were the students' attitudes towards these new classroom materials?

AntConc, a free corpus analysis toolkit created by Laurence Anthony, was used to analyze the students' use of transition words from the writing textbook. The frequency of the transition words in the students' essays was then compared to the frequency of these transition words in the academic subcorpus of the Corpus of Contemporary American English (COCA) to determine which words were overused or underused by the students as well as which words were more frequent in academic English. Next, AntConc was used to analyze students' errors in the use of these transition words. Finally, specific transition words to target in each rhetorical style of writing were selected based on

the data on frequency and student error. Classroom materials including student examples of appropriate use of these transition words, exercises for students to fix mistakes in other student examples with errors, and a section for students to practice using the transition words in sentences were created to help students build cohesion in their academic writing.

To evaluate the effects of the instruction, AntConc was again used to find the frequencies and student errors in the essays of students who had received the new instruction and compare these to the essays of prior students who had not been exposed to the new teaching materials. Students were also surveyed to determine their attitudes towards the new materials. In this way, the use of the learner corpus made it possible to compare essays from different groups of students and continuously evaluate the effectiveness of the teaching material, aiding in the action research cycle by facilitating data collection in the observation stage and analysis in the reflection stage.

3. RESULTS

To answer the first research question, the frequency of the transition words in the learner corpus was compared with the frequency of those words in the academic subsection of the COCA. Although the academic subsection of the COCA contains a mixture of different genres, this free and open source was used to determine a relative, base frequency for the word in academic English in general which could then be compared to the frequencies in the student essays and could also be used to inform curriculum design. In other words, some transition words included in the textbook were actually found to be quite infrequent in the COCA while others occurred very frequently, so just because the students did not use a word very often did not mean that it was underused or that it should be targeted in the course materials.

To answer the second research question, AntConc was again used to search for the transition words in the students' essays and these were analyzed for errors. Both the overall frequency compared with the academic subcorpus of COCA and student errors were taken into account when designing the course materials. That is, the course materials aimed to target underused transition words that appeared frequently in the academic subcorpus of COCA as well as transitions that students frequently misused. General results for both the frequencies and student errors are shown in Table 1 below:

Table 1: Overused/Underused Transitions from the Learner Corpus with Common Mistakes

Essay Genre	Most Overused Transitions	Most underused transitions	Common Mistakes
Cause/Effect	Cause/caused by, factor, because of, due to, lead to, this means	Consequence/ consequently/as a consequence, influence, result, thus	"consequence" with positive effects, mistakes with commas, fragments, or run-on sentences using the transitions, not using a gerund after the preposition "to" in "due to/lead to," wrong structure with "attribute to"
Compare/Contrast	Difference/ different, similarity, however, but, also	Likewise, whereas, yet, though, on the contrary,	Missing "the" with "the same as," "in the other hand(s)" instead of "on the other hand," mistakes with commas, fragments, or run-on sentences using transitions
Problem/Solution	Issue, problem, solve, solution	Burden, complication, alleviate, cope with, ease,	Wrong meaning (i.e. using a problem work for a solution or vice versa), using "with" after "tackle"
Opinion	Also, but, however, while, for example	Similarly, in contrast, in spite of, nonetheless, some people might say,	Using "in" or leaving out "the" for "on the other hand," incorrect punctuation, wrong structure with "in spite of"

After the students received class instruction using the course materials developed from the learner corpus, AntConc was again used to analyze their essays. It was found that the students generally improved the overall frequency of the targeted transition words, but their essays still exhibited some mistakes (see example in Table 2 below from the compare/contrast essays). This evaluation can then inform revisions to the original teaching materials so that they can be further improved for the next semester. In this way, the learner corpus can be used to continuously evaluate and improve teaching materials for the advanced English class.

Table 2: Results of Using Classroom Materials for Compare/Contrast Essays

Transition Word	Frequency (per million) in original learner corpus	Errors from Learner Corpus	Frequency (per million) after using classroom materials	Errors after using classroom materials
On the contrary	175	None	394	Capitalization
On the other hand	979	In other hands, on the other hands, incorrect punctuation (3x)	787	Spelling, on the other hands
Whereas	35	None	394	Punctuation, fragment (2x)
Though	105	Punctuation, structure	98	None
Compared to	140	Punctuation, structure	394	Structure (2x)
The same as	140	No article "the" (3x)	197	No article "the"
Similar to	105	No gerund after "to"	197	Wrong preposition (for instead of to)
Yet	140	None	197	None
But	3,566	Starting the sentence (15x)	4,574	Starting the sentence (12x)
However	2,552	punctuation	1,918	Capitalization, punctuation (2x)

Finally, the students were surveyed to determine their attitude towards the course materials. As can be seen in table 3 below, students had a positive attitude towards the course materials and appreciated the inclusion of prior students' examples. The students preferred the exercises in which they had to fix other students' mistakes, but also had favorable attitudes towards seeing model student examples as well as writing their own example sentences.

Table 3: Students' Attitudes Towards the New Course Materials

Survey Question	It was helpful	It was not helpful
In class, we used Cause/Effect, Problem/Solution, Argument, and Compare/Contrast transition worksheets. How did you feel about seeing the good student example sentences?	91% (21/23 students)	9% (2/23 students)
In class, we used Cause/Effect, Problem/Solution, Argument, and Compare/Contrast transition worksheets. How did you feel about practicing fixing the incorrect student example sentences?	96% (22/23 students)	4% (1/23 students)
In class, we used Cause/Effect, Problem/Solution, Argument, and Compare/Contrast transition worksheets. How did you feel about writing your own example sentences using the transitions?	91% (21/23 students)	9% (2/23 students)

Overall, it can be seen that developing and analyzing the learner corpus allowed instructors to better understand the students' writing and address underused and misused transitions through materials created using student examples, leading to more successful classroom outcomes for the students.

4. FURTHER RESEARCH AND NEXT STEPS

The next step is to continue using the learner corpus to update and evaluate the class materials targeting transition words in different writing genres. So far, the course materials have only been implemented in the Fall 2018 semester. Evaluating the student essays written during this semester and considering the student comments from the surveys, the materials can be adjusted and improved for the next semester. In this way, the learner corpus contributes to the action research cycle through continuous observation, evaluation, and implementation of newer, more effective materials based on previous student performance.

In addition, the methods and course materials described in this proposal is just one of many ways that this learner corpus can be used to improve course curriculum and student outcomes in this advanced English course. Using a corpus linguistics approach, the student essays can be evaluated for other characteristics, such as other vocabulary words, grammatical errors, and writing style. Moreover, as this advanced course is the highest level offered in the IEP, the student essays could be used to evaluate program objectives as a whole in order to further develop and improve student outcomes. This analysis can then inform curriculum development for the program and individual courses, which can then be further evaluated using the corpus software for continuous reflection on and improvement of the curriculum and teaching materials.

5. CONCLUSION

This presentation highlights a practitioner application of computer tools to analyze student performance and develop class materials. Moreover, it is an example of how the action research process can enable instructors to better understand their students, reflect on their own teaching, and improve course outcomes for their students.

The corpus software used in this research can be applied to many other contexts as well. Developing a learner corpus can aid in any course that requires written or spoken discourse, for example foreign language courses and composition courses. Corpus tools can also be used to analyze student writing in specific genres, such as dissertations, case reports, legal documents, experimental reports, or other discipline-specific writings. Thus, learner corpora and corpus software has a wide range of application in various learning contexts.

It is hoped that the presentation will inspire others to consider ways of using the action research process to help practitioners improve student learning. This presentation should also encourage audience members to consider how focusing on ways to improve curriculum development, evaluating student progress, and supplementing textbook materials can improve student outcomes in their own contexts, leading to greater student success and inclusion.

REFERENCES

- Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software>
- Crosthwaite, P. (2013). An error analysis of L2 English discourse reference through learner corpora analysis. *언어연구*, 30(2), 163-193.
- Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at <https://corpus.byu.edu/coca/>.
- Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-335.
- Hasselgard, H. (2009). Thematic choice and expression of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.), *Corpora and language teaching* (Studies in corpus linguistics, v. 33) (pp. 121 – 140). Amsterdam: John Benjamins Publishing Company.
- Herriman, J., & Aronsson, M. (2009). Themes in Swedish advanced learners' writing in English. In K. Aijmer (Ed.), *Corpora and language teaching* (Studies in corpus linguistics, v. 33) (pp. 101 – 120). Amsterdam: John Benjamins Publishing Company.
- Jones, C., Byrne, S., Halenko, N. (2017). *Successful Spoken English: Findings from Learner Corpora* (1st ed.). New York, NY: Routledge.
- Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and language teaching* (Studies in corpus linguistics, v. 33) (pp. 203 - 230). Amsterdam: John Benjamins Publishing Company.
- Nesselhauf, N. (2004). "Learner corpora and their potential for language teaching." In Sinclair, J. (ed.) *How to Use Corpora in Language Teaching* (pp. 125 – 152). Amsterdam: John Benjamins Publishing Company.
- O'Donnell, M. (2012). Using learner corpora to redesign university-level EFL grammar education. *Revista Española De Lingüística Aplicada*, (1), 145-160.
- Römer, U. (2010). Using general and specialized corpora in English language teaching: past, present, and future. In M.C. Campoy-Cubillo, B. Belles-Fortuno, & M. Gea-Valor (Eds.) *Corpus-based Approaches to English Language Teaching*, (pp. 18 – 37). Bloomsburg Publishing PLC.
- Seidlhofer, B. (2002). Pedagogy and local learner corpora: working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching* (pp. 213 – 234). Amsterdam: John Benjamins Publishing Company.
- Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press
- Soruç, A., & Tekin, B. (2017). Vocabulary Learning through Data-driven Learning in an English as a Second Language Setting. *Kuram Ve Uygulamada Egitim Bilimleri*, 17(6), 1811-1832.

Empowering Tutors with Big-data Learning Analytics

Author(s): Uma P. Vijh

Kidaptive Inc.
uma.vijh@kidaptive.com

Author(s): Josine Verhagen

Kidaptive Inc.
josine.verhagen@kidaptive.com

Author(s): Webb Phillips

Converz Analytics B.V.
webb.phillips@gmail.com

Author(s): Ji An

Kidaptive Inc.
ji.an@kidaptive.com

ABSTRACT: Presentation. Online education has been growing over the past few years, and massive amounts of learning data are being generated. We are reporting on our efforts to use learning analytics to empower teachers to help all learners reach their full potential. We provide teachers with insights about student behavior and achievement on a weekly basis and supplement these with summary monthly reports about student study patterns and trends. This paper provides more detailed descriptions of these reports and also includes preliminary efficacy study results that show positive effects on mean student test scores.

Keywords: study pattern analytics, teacher insights, informal learning environments, predictive modeling, bayesian item response theory

1. INTRODUCTION

In recent years, learning analytics has emerged as a powerful learning tool for teachers who participate in online learning programs. Big-data learning analytics deciphers massive amounts of data generated in different learning contexts. It can help to assess students' academic progress, predict their future performance, and identify potential problems (Johnson, Adams & Cummins, 2012). For teachers, learning analytics can be used to carry out a more in-depth analysis of the teaching process to provide more targeted teaching interventions for students (Chen, Heritage & Lee, 2005).

2. BACKGROUND

In this paper, we describe our learning analytics efforts to support teachers helping K–6 learners. The data are event data as learners interact with curricular content from a Korean partner's tablet-based educational system. The system supports over 200,000 learners in Math, Korean, Social Studies and Science, following the Korean national curriculum. Students in the program mostly work at home and are visited by a teacher once a week. The

content is arranged in weekly topics and further broken down into small content blocks containing lectures and practice questions. Each week ends with a test. As the learners progress through the curriculum, they watch lectures, answer 50–100 practice questions, and complete a test with 10–20 questions. Our technology provides teachers with weekly reports that are updated continuously, as well as monthly reports to track the learners' progress over time. These reports (described in subsequent sections) contain more information than just the correctness/incorrectness of student answers. Our cloud-based analytics engine processes millions of events streaming in, using regularly calibrated psychometric models to produce hundreds of distinct personalized metrics and insights. These insights are dynamically prioritized, with the most important passed along to teachers to help all learners reach their full potential.

3. METHODS

3.1. Description of the report:

For every weekly curricular unit attempted by a learner, we produce a report for that learner's teacher. In this weekly report, we provide general behavioral insights, specific question-level insights, and one overall message about the learner's behavior and achievement during the week.

The behaviors analyzed are: skipping questions, answering too quickly or slowly, guessing, leaving parts of the question blank, skipping a question after getting the previous one wrong, retrying or not retrying incorrect questions, watching or not watching all lectures, and checking or not checking hints after getting a question wrong. In addition to these behavior metrics, the reports also include question insights based on personalized speed and ability estimates and performance on the weekly test. These details empower the teacher to quickly identify questions/concepts each student is struggling with, praise good study habits, and assess student performance not only at a personal level but also in comparison with peers.

To tell teachers more than whether question responses were correct, we developed some additional insights about responses.

3.1.1. Answer speed:

An item is flagged as answered relatively fast or slow based on the learner's expected time on the item given their working speed and whether the learner is answering faster or slower than 90% of the other students answering the item. Based the learner's history in a given subject, a Bayesian personalized estimate is kept of his or her working speed. The working speed is updated only based on items the student answered correctly, to keep the estimate from plummeting when a student is just skipping through questions. The estimate is based on a linear mixed model of the logarithm of the response time, with the learner's working speed estimate calculated relative to the average time spent on the item by other learners. E.g., if a learner's response time is faster than 90% of other learners' response times but consistent with that learner's working speed, the item is not flagged as too fast.

3.1.2. Item difficulty:

Based on the learner's ability estimate and question difficulty, questions are categorized as hard (<50% probability of getting the question correct), easy (>80% probability of getting the question correct) or medium for a given learner. Ability estimates are based on an adjusted version of Bayesian Item Response Theory models (Bock & Mislevy, 1982; Van der Linden & Glas, 2000) developed for adaptive testing, which allows the ability estimate to be updated after each question. Because reports are generated on an edition level, the final ability estimate and question difficulty estimates represent how well a learner did compared to other learners at the end of that edition. At the start of each edition, the prior probability distribution is set to the average of the priors from the three previous editions, with a wide standard deviation to allow for a different ability level for the topic at hand.

3.1.3. Guessing:

We developed a general model for estimating thresholds for response times that are short enough to suggest that students probably guessed the answer (See Wise & Kong, 2000; Baker et al. 2006 for discussion on rapid response times). This model applies across all question types and is based on the distribution of response data and corresponding pass-rates on a per-question basis. Using this model, we were able to categorize responses as "guessed" much more accurately than simply setting an arbitrary response time for all questions. Comparing response times to pass-rates, most questions have a region of low response times with low pass-rates and a region of higher response times with higher pass-rates. Then the pass-rate gradually declines for even higher response times. The log normal distribution shares a similar shape, and therefore makes a good function to model response time vs. outcome. As an example, Figure 1 shows a model for one math question after having optimized four coefficients. These models have low mean squared error (~0.05) compared to actual response time vs. outcome data. We found that our model needed at least 50 correct and 50 incorrect responses to be reliable. Of the 58,806 questions for which our analytic platform had responses and response times, our modeling algorithm assigned a

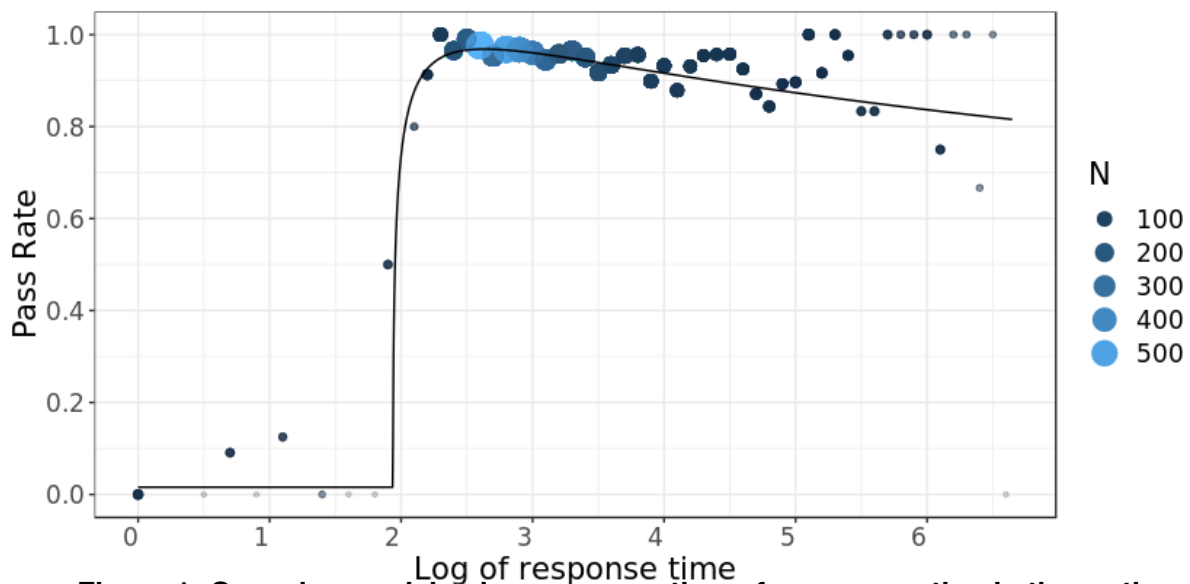


Figure 1: Guessing model using response times for one question in the math curriculum

default guessing threshold of up to one second for 69% of questions, specific thresholds greater than one second for 29%, and no guessing threshold for 1.6% (these were cases in which the percentage correct at one second was almost as high as or higher than the percentage correct at the middle 20% of response times for correct answers).

The combination of personalized answer speed, item difficulty, and item correctness produces insights regarding sets of items. Based on the individual ability estimate and the estimated item difficulty of the items in the next test in the curriculum for each learner, we also record an estimate of that learner's predicted performance on the upcoming test. We then use this estimate to provide further insight to the tutors (e.g., to congratulate or encourage the learner to do their best).

As learners work through the curriculum, we also provide monthly reports to the tutors, summarizing the learner's activity for the month as well as trends across months. This helps the tutor evaluate student learning and growth, praise improving study behaviors, and celebrate achievements.

4. EVALUATION OF INTERVENTION

Our reports were provided to all users of our partner's platform, so a direct control group was not available for the evaluation of the program. We evaluated efficacy of the product in two ways using linear mixed models:

1. *A Difference in difference analysis of historical data:* We compared the differences in test scores in the current year with those in the previous year, before and after launch of the teacher reports. This method accounts for the seasonal differences in course material, but the individual students are of course different. To mitigate this we included the random effects for difficulty of particular curricular material and individual learner ability. Month and year were modeled as fixed effects, and we also included interaction

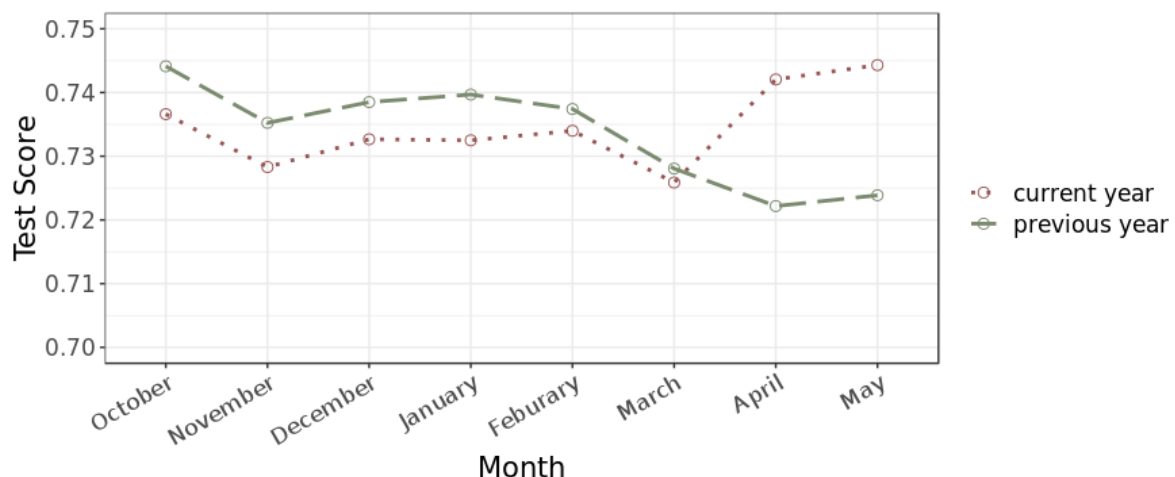


Figure 2: Mean test scores for Korean, for the current year and the previous year. Our intervention program was implemented in February.

effects between month and year. We used over 1.2 million individual scores of ~40,000 learners for the subject Korean. We included data for eight months from each year; the

data selected for this year covered three months before the implementation of our tutoring support service (February 2018) and five months after. Figure 2, shows the mean test scores for Korean during the previous year (2017) and the current year. The differences between current year and the previous year were essentially constant until January, with overall test score in the previous year being slightly higher than the current year. Starting from February, the current year scores start to catch up, and by April they outperform the previous year's scores. The increase ranged between 0.4 and 3.6 points across all the subjects, Math, Korean, Social Studies and Science on the scale of 0–100 points. Statistically significant, positive interaction effects start around one month after the implementation of the program, indicating that the test scores relative to last year have shown improvement after the start of the service.

2. A analysis based on frequency of report utilization by teachers: The historical analysis

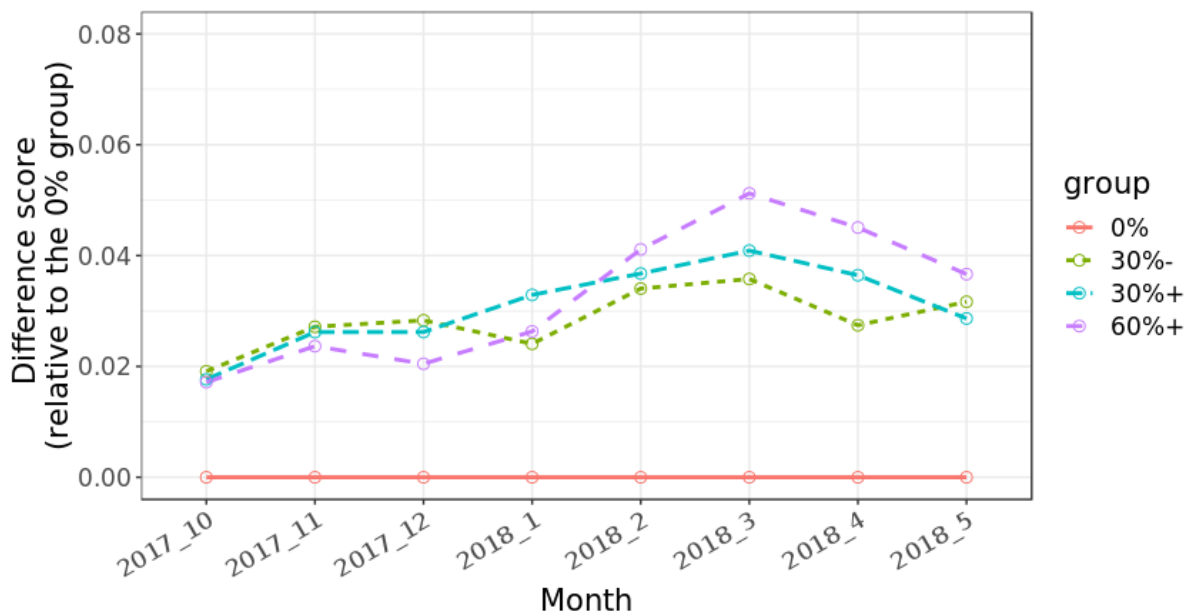


Figure 3: Differences in the mean test scores of students whose reports were viewed, grouped by the fraction of their reports viewed.

does not measure a direct effect of intervention by the teachers who are empowered by our reports. To evaluate a more direct effect we compared the test scores of students as a function of the rate at which their reports were viewed by the teacher. Our hypothesis was that teachers empowered by the personalized insights would provide timely intervention and over time positively influence student behavior. In addition to the year/month fixed effects and learner/material random effects, we grouped teachers based on what proportion of the students' reports they viewed and used the resulting group membership as a fixed effect to model the test scores. We analyzed all the subjects, Math, Korean, Social Studies and Science for which we provided reports. Figure 3 shows the difference in scores of the students whose teachers viewed their reports at different frequencies compared to those whose reports were not used at all, for Math. The performance of students whose teachers are in the never-viewed group is clearly worse than that of students whose teachers are in the three sometimes-viewed groups,

but those differences increase after January 2018, indicating an improvement associated with teachers viewing reports. As shown in Table 1, the improvements in scores relative to the never-viewed group range from 1.02 to 3.07 points depending on the rate of teacher viewing and time of the year. The statistical significance of these differences is indicated in parentheses and explained in the footnote.

Table 1: Differences in test scores between the indicated group and the group without any report views. Statistical significance indicated in parenthesis¹

% of reports viewed	2018/1	2018/2	2018/3	2018/4	2018/5
< 30%	0.72(*)				
30% - 60%	0.52(.)	0.65(*)	0.73(*)	0.81(**)	0.93(**)
> 60%	0.98(***)	1.12(***)	1.20(***)	1.69(***)	1.40(***)

5. CONCLUSION

In this paper we have described a real-world instance of learning analytics indirectly supporting ~200,000 learners through personalized weekly and monthly reports sent to those learners' teachers. These reports characterize a variety of learning-relevant study behaviors to help teachers identify and correct bad habits, praise and reinforce good habits, and optimally direct each learner's study efforts. Two types of analyses comparing scores before and after implementation of personalized insights to tutors suggest a positive effect on test scores, especially for students whose reports are frequently viewed by their teachers.

REFERENCES

- Baker, R., Koedinger, K. R., Corbett, A. T., Wagner, A. Z., Evenson, S. et al.. Adapting to When Students Game an Intelligent Tutoring System. International Conference on Intelligent Tutoring Systems, 2006, Jhongli, Taiwan. 2006. <hal-00190177>
- Becker, S.A., Cummins, M., Davis, A., Freeman, A., Glesinger Hall, C. & Ananthanarayanan, V. (2017). NMC Horizon Report: 2017 Higher Education Edition. Austin, Texas: The New Media Consortium.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, 6(4), 431-444.
- Chen, E., Heritage, M. & Lee, J. (2005). Identifying and Monitoring Students' Learning Needs With Technology. *Journal of Education for Students Placed at Risk*, 10 (3), 309-332.
- Wise, S. L. & Kong, X. *Applied Measurement in Education*, 2005, Montreal, April, 2005 Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

¹ significance: . p < .1; * p < .05; ** p < .01; *** p < .001

Implementation of Learning Analytics to Optimize Learning and Learning Environments: Tertiary Instructor Perspectives

Julie (Junli) Wei

Faculty of Arts

The University of British Columbia

julie.w@ubc.ca

Fred Cutler

Faculty of Arts

The University of British Columbia

fred.cutler@ubc.ca

Leah P. Macfadyen

Faculty of Education

The University of British Columbia

leah.macfadyen@ubc.ca

Sanam Shirazi

Faculty of Arts

The University of British Columbia

sanam.shirazi@ubc.ca

ABSTRACT: The evolution of learning analytics (LA) systems and tools offers unprecedented opportunities to make use of insights from learning data to promote effective teaching and learning practices. However, a significant gap still exists between what is possible, and what is being applied in practice. Little is known about instructors' interests and concerns in relation to implementing LA for supporting classroom practices. This study aims to address the gap by identifying tertiary teachers' interests and concerns regarding the implementation of LA in teaching and learning. Interviews and surveys were used to collect responses from faculty members of a large research-intensive university. Findings reveal tertiary instructors' degree of familiarity with LA, and their attitudes to, interest in and concerns about using LA tools in various contexts. Discussions about how to take actions to enhance teaching and learning practices with the implementation of LA are provided.

Keywords: implementation, learning analytics, tertiary teachers, interest, concerns

1 INTRODUCTION

Today's higher education institutions, tertiary instructors and students are faced with a complex set of challenges. While many factors may contribute to the efficacy of learning, understanding which combination of these factors will most effectively improve student learning remains a pressing challenge. Learning analytics (LA) has evolved as a field of research and practice that employs data driven methods to understand and improve student learning processes and outcomes (Macfadyen & Dawson, 2010). In order to help maximize student learning success, instructors must more actively engage in the design, development and implementation of LA (Ferguson, Macfadyen, Clow, Tynan, Alexander & Dawson, 2014). It is therefore crucial to identify instructors' interests and concerns about using LA as well as provide appropriate forms of support and training. We developed a survey and analyzed it to explore the interests of faculty members in implementing LA, and present our first-level findings here. Based on these findings, we propose strategies to support instructors that will allow them to benefit more from available LA.

2 METHODS

All faculty members from the largest faculty of a public research university were invited via email to answer our online survey. One hundred and eighteen instructors, from a total number of 659, completed the survey yielding a response rate of 16%.

Our survey had a special ‘educational’ function. Recognizing the widespread lack of knowledge about LA among educators and practitioners, the survey first introduced eight LA tools/methods (i.e., use cases) and the potential benefit of each to teaching or learning, before asking respondents to rate their level of interest in using these tools. It was hoped that this approach to survey design may help collect more meaningful and informed responses. The survey also included questions about instructors’ teaching experience and contexts, the course levels and class sizes that they typically teach, their familiarity with LA, their opinions as to whether the university and its faculties should devote significant resources to LA and support for instructors to use LA, their level of interest in using eight LA tools, and their interest in participating in a funded LA project. Space for open comments was also included, to elicit any other questions or opinions that the respondents might have about using LA to address student learning success meaningfully.

3 RESULTS

Survey response data was collected online and visually analyzed in Tableau 10.3. The primary findings are shown in Table 1 below. Content analysis was performed on the qualitative data obtained from open-ended questions. Salient thematic areas were identified and coded. Note: not all respondents answered all questions. Results presented below indicate the number of respondents to each question.

3.1 Teaching Contexts of the Respondents

Many respondents teach several levels of undergraduate and graduate courses with varying class sizes, but a majority teaches 3rd and 4th year students. In addition, the proportion of instructors that teaches small and medium classes is much larger than the proportion that teaches large classes. However, due to insufficient information on teaching experience, discipline, career stage, appointment type, and the nature of the course they teach, we cannot determine whether our pool of respondents is representative of the faculty as a whole.

3.2 Awareness of Learning Analytics

Respondents were invited to indicate their level of awareness and understanding of LA on a Likert scale from “Only vaguely know or not know” to “know it very well”. The majority of respondents (87%, n=118) reported that they only vaguely knew or did not know what LA refers to, and only 13% felt that they knew it very well.

3.3 Attitudes to Devoting Resources to LA and Support for Use of LA

More than 70% of the respondents thinks the university should devote resources to LA and support for instructors to use LA. This result is as expected, as we assume that most are aware of the rising

importance of LA in general, even if they are unfamiliar with methods, tools or implementation approaches.

3.4 Interest in Using Eight Tools for LA

The survey provided respondents with eight examples of LA tools and asked them to rate their level of interest in each tool in each context. Preliminary results are shown in Table 1 below. The first column lists eight tools that could be used for LA. The second displays four LA contexts that each tool could be implemented within. The third shows associated instructors' interests in using the tools. It should be noted, the original 0-10 Likert Scale is aggregated into three categories below (i.e., not interested, neutral or interested) for easier viewing.

Table 1: Instructors' interest in using eight LA tools in four contexts

LA Tools	Implementation Context	Instructor Interest		
		Not..	Neutral	Interested
1. Visualizing student enrolment pathways	<i>during a course</i>	36%	29%	34%
	<i>inform course change</i>	30%	33%	36%
	<i>research</i>	43%	15%	41%
	<i>program planning</i>	17%	33%	50%
2. Tracking progress and giving feedback	<i>during a course</i>	49%	24%	27%
	<i>inform course change</i>	40%	37%	23%
	<i>research</i>	54%	21%	25%
	<i>program planning</i>	46%	40%	14%
3. Monitoring student and class activity in the course site, in real time	<i>during a course</i>	31%	30%	39%
	<i>inform course change</i>	43%	28%	29%
	<i>research</i>	50%	20%	30%
	<i>program planning</i>	49%	28%	23%
4. Monitoring student activities in your course's online discussion forums	<i>during a course</i>	38%	32%	30%
	<i>inform course change</i>	41%	39%	20%
	<i>research</i>	50%	25%	25%
	<i>program planning</i>	56%	35%	9%
5. Measuring the impact of student engagement with course material on their course grades or other indicators of learning	<i>during a course</i>	40%	25%	35%
	<i>inform course change</i>	41%	25%	34%
	<i>research</i>	56%	13%	31%
	<i>program planning</i>	48%	30%	22%
6. Making better use of student performance data to inform curriculum redesign	<i>during a course</i>	48%	26%	26%
	<i>inform course change</i>	39%	27%	34%
	<i>research</i>	45%	19%	36%
	<i>program planning</i>	22%	28%	50%
7. Knowing students before the first class	<i>during a course</i>	18%	25%	57%
	<i>inform course change</i>	29%	31%	39%
	<i>research</i>	43%	23%	34%
	<i>program planning</i>	23%	33%	43%
8. Helping students monitor their own level of preparation for class	<i>during a course</i>	25%	30%	45%
	<i>inform course change</i>	33%	31%	36%
	<i>research</i>	53%	17%	30%
	<i>program planning</i>	58%	21%	21%

Overall, survey results indicate that more instructors want to use LA at the course and program level for planning and adapting to student real needs. Specifically, they have a high preference for using LA in:

- *knowing who their students are, before a class begins (57%)*
- *visualizing student enrollment pathways to plan curriculum, course offerings and sharing of teaching knowledge (50%)*
- *making better use of student performance data to inform curriculum redesign (50%)*

Conversely, fewer instructors indicated interest in using LA for teaching and learning-related research. To our surprise, given the level of interest among LA researchers in developing student-facing analytics tools, a majority indicated no interest in applications designed to assist students with self-regulated learning. Specifically, greater than 50% of respondents indicated NO interest in using LA for:

- *helping students monitor their own preparation for class (58%), as well as monitoring students' online discussion activities (56%), both in the context of program planning*
- *measuring the impact of student engagement with course material (56%), tracking progress and giving feedback (54%), helping students monitor their own preparation for class (53%), monitoring student online activity and class activity in the course site (50%), all in the context of doing research on teaching and learning*

3.5 Interest in Participating in a Funded Project

The majority of respondents (62%, n=97) indicated that they are not interested in participating in a funded LA project to answer some of their questions, drawing on expertise from the faculty and university LA professionals. Possible reasons for this are discussed at the end of this paper.

3.6 Findings from Open-Response Questions

In addition to the preliminary findings above, open-response data from the survey was coded and analyzed using NVivo 12 software, with the goal of complementing the quantitative findings of this study. This analysis revealed that instructors wanted to use LA to answer questions such as: What factors affect student's learning outcomes? How to better assess and evaluate students? Some instructors expressed their concerns about using LA in their comments. It seems some of their major concerns are:

- *Time pressure: time seems to be the biggest concern, as the instructors felt that learning a new software, and collecting and analyzing data take time, which would add a burden to their time- constrained schedule.*
- *Cost vs Benefits: many instructors were worried that LA might become yet another instance where they needed to learn a lot but achieve very little.*
- *Limitations of LA: some instructors felt that LA could not do much to students who fall outside the LA net. LA is also over-and-above the core tasks of humanities education such as critical thinking, reflection.*

4 DISCUSSION

This survey takes initial steps towards identifying the interests of faculty members in using LA tools to maximize student learning success. Although the majority of the instructors stated that they do not know what LA refers to, they agreed that the university and its faculties should be devoting significant resources to LA and to support for instructors to use LA. This is probably because their faculty has been playing a leading role in employing LA, with a wide variety of LA projects underway. These instructors thus have some experience using LA to understand “what is happening” in classrooms, though they might lack detailed knowledge of the broader field of LA.

Instructors demonstrated a variation in their level of interest. Overall, many instructors appeared to appreciate that LA could provide them with actionable information for their teaching and their department. While such information might be used in a variety of ways, these instructors appreciate the ability to plan ahead, such as planning curriculum and course offerings based on greater knowledge of who their students are before the first class, and visualizing student enrollment pathways.

Instructors are also interested in using LA to discover solid evidence that can inform course and curriculum redesign. It is challenging for instructors to remember details of each student’s activity and performance, especially in large classes. However, LA can offer visual data presentations to simplify learner monitoring as they encounter course elements over time. These can help instructors identify problems early in a course that need to be addressed, as well as justify their concern by confirming or disputing whether changes should be made.

In general, it seems instructors care most about the practical use of LA, rather than in using LA to support research on teaching & learning. They also do not value LA information that they don’t believe to be meaningful or actionable. For example, instructors showed little interest in monitoring student activity in online discussion forums, or in measuring the impact of student engagement with course material.

In summary, instructors want to be mindful of student background information and help students get on track with the course to increase the probability of success. They thus prefer to use LA to improve classroom practice such as planning ahead, early identification and intervention, and informing changes. It was unexpected that the majority of instructors indicated no interest in participating in a funded LA project, even though they agreed that LA could be a fantastic opportunity to improve student learning and the university should devote significant resources to LA. Two factors may be causing this interesting phenomenon:

- First, as demonstrated in their open response, many instructors were concerned about time constraints and workload, cost and benefits of LA, as well as the limitations of LA.
- Second, trust in LA fidelity and usefulness needs to be established. It is common for users to be wary of new technologies, and LA is no exception. Some instructors felt that the LA tools can be inaccurate, impersonal, or intrusive, discouraging them from investing time in them.

5 CONCLUSIONS

Organizational, technological, and pedagogical environments can create both barriers and opportunities for successful implementation of LA. An important conclusion that can be drawn from this small study is that effective LA implementation calls for regular communication with instructors about LA, and provision of appropriate levels of training to support their LA efforts.

First, establishing transparency and building trust in the use of LA is essential. Instructors must feel confident that data collection and analysis is transparent and accurate, and that these processes are continually evaluated and refined. If instructors mistrust and disbelieve the data or its utility, they will be unlikely to invest time and energy in LA.

Second, user interfaces must be easy to understand and training/support opportunities must recognize and work with their time constraints. This could promote the broad adoption and use of LA at the course level.

Third, alignment of LA options to instructor's values and pedagogical goals, as well as the institution's strategic direction will encourage uptake. The process of understanding and optimizing learning requires a thorough understanding of how learning can be facilitated and supported, and how various factors can affect learning. Creating and illustrating robust connections between LA, instructors' pedagogical intent, and learning science should be strengthened. LA implementation needs to be scaffolded, aligned, and adjusted to fit the institution and its instructors, rather than asking instructors to adjust their direction and pedagogical practice to support the use of LA.

Fourth, it must be clear that LA tools and methods are being implemented responsibly and ethically. Instructors should not only be trained to understand LA and its possible benefits and limitations. They also need to develop skills with the technologies, with data interpretation and in understanding related ethical issues, including but not limited to how they should responsibly assess and appropriately use the data, what they should and should not do with sensitive personal information.

While this survey allowed us to capture data about many aspects of our instructors' LA interest, further studies are needed investigate key issues in more detail. We also plan to involve more instructors in the future. Follow-up focus group interviews will be conducted to further identify barriers that prevent instructors from using LA, and to discover which conditions would better support instructors in benefiting from LA.

REFERENCES

- Macfadyen, L. P. & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, 54, 588-599.
- Ferguson, R., Macfadyen, L. P., Clow, D., Tynan, B., Alexander, S., & Dawson, S. (2014). Setting learning analytics in context: Overcoming the barriers to large-scale adoption. *Journal of Learning Analytics*, 1, 120-144.

Improving Dashboard Usability: A Case Study

Author(s):

Dr. Bradley Coverdale
University of Maryland University College
Bradley.coverdale@umuc.edu

Dr. Matthew Hendrickson
HelioCampus
Matthew.hendrickson@umuc.edu

ABSTRACT: This information in this study will be conveyed via **Presentation** format. Today, dashboards are used to explain pertinent information for decision making at higher education institutions. However, a tool is only effective if it is used. This study explores the process one online higher education institution completed to improve existing electronic dashboards to meet users' needs. Conducting a needs analysis with stakeholders revealed three areas of improvement. Changes were made and new versions of the dashboards were presented to the user groups for feedback. After the final version has been approved by administration, the new dashboards will be demonstrated to users. Activity logs will be monitored and analyzed for potential trends.

Keywords: dashboard usage, benchmarks, user experience, optimization, process improvement

1. INTRODUCTION

Today, the stakes for success in higher education are higher than before, especially for adult students who already have a career. These students are interested in obtaining a new or updated set of skills to improve their job performance or perhaps transition to a new career. Institutions that serve these students are evaluated based on how well they deliver the desired skills, especially at an online institution (Viberg, Hatakka, Balter, & Mavroudi, 2018). Student data is captured across all levels of their educational experience, from inquiring about available programs to enrolling in the first course and eventually completing their credential and progressing into their careers as alumni. However, so much data can be overwhelming when desiring to make decisions about improving the educational experience (Bill and Melinda Gates Foundation, 2015; Vatrappu, Teplovs, Fujita, & Bull, 2011). It can be difficult to determine what data is noise and what data is relevant (Picciano, 2012; Qu & Chen, 2015; Siemens & Long, 2011; Smith, 2013). One tool that helps with this challenge is electronic dashboards which display one set of data as well as potential trends. Yet, a tool is only effective if it is being used, otherwise it needs to be improved and catered to the users' needs. This paper is a case study that explores the challenges with modifying dashboards to be more relevant to users as well as highlights user feedback following the changes.

2. LITERATURE REVIEW

Before evaluating the case study it is important to understand how others have used or evaluated the dashboard user experience.

2.1 Why are dashboards the right medium to display data?

Dashboards provide an interface for visual analytics which allows for understanding analytical methods through visual interfaces (Thomas & Cook, 2005). These visual techniques can identify patterns in data that may not be apparent to those without statistical backgrounds (Ndukwe, Daniel, & Butson, 2018). The organization of data can help the decision making process shift from reactive to proactive (Fernandez, McClain, Brown-Williams, & Ellison, 2015; Smith, 2015). Eventually, the analytic models displayed in the dashboards could inform students about their rate of completion based on current behaviors as well as behaviors that can be implemented to improve a score (Smith, 2015). This could help adult students who may need to prioritize their time and decide which action will lead to the highest rate of success.

However, one of the criticisms of dashboards is that most designs are box, bar, or line plots instead of something that is more innovative or provides more interaction (Vieira, Parsons, & Byrd, 2018). Additionally, some research has observed that it can be a challenge to create one dashboard that can provide sufficient information for decision making that applies across all courses and instructors (Aljohani, Daud, Abbasi, Alowibdi, Basher, & Aslam, 2018). One dashboard may not display the needed information in a way that benefits all users.

2.2 Determining if a dashboard is effective

One of the challenges in evaluating a dashboard is evaluating the learning analytics results that the dashboard displays. In fact, one research study reviewed papers related to higher education and learning analytics and observed that only 9% of the 252 studies reviewed presented any evidence related to improving student learning outcomes (Viberg, Hatakka, Balter, & Mavroudi, 2018). This suggests that while there are many studies discussing the practice or potential models for learning outcomes, very few are able to describe how to improve them, which could be problematic when trying to design a dashboard to map learning outcomes. Another challenge in determining effectiveness is the limitation of the dashboard results. Often, dashboards tend to report outcome feedback (how is the student doing) instead of process feedback (how can the student improve?) (Sedrakyan, Malmberg, Verbert, Jarvela, & Kirschner, 2018).

3. SAMPLE

The participants in the case study include all of the dashboard users. This includes 32 program chairs, 6 deans, and 17 program specialists from the undergraduate school in addition to 39 program chairs, 6 deans, and 12 program specialists from the graduate school. Each role has limited viewing capacities of

the dashboards based on their job responsibilities. For example, deans can view reports on all of programs in a particular school whereas program chairs can only view dashboards that pertain to an individual program. Currently, the dashboard system is broken into categories to allow for quick access: Enrollment Trends, Faculty, and Success Metrics. An analysis of the dashboard activity logs revealed that only 33% of users logged in to the dashboard between January 2018 and March 2018. However, there is not a clean way to understand which dashboards are more effective based on user actions. Anecdotal responses suggest that many users do not utilize the dashboards because they are not catered to relevant needs or are not user-friendly. Further initial viewings were suspect as the numbers (i.e. enrollment) did not align with periodic reports without proper customization/limitations.

4. RESEARCH DESIGN

The research design was created and implemented by a third party data analytics team, which is wholly integrated into the university. Over the span of many months, the team conducted iterative feedback sessions, gathering feedback from the Academic Affairs teams. This feedback was solicited through open discussions, email, and product iterations. All major Academic Affairs teams were included, ranging from the Chief Academic Officer and Deans to the Program Chairs, Program Managers, and Assessment Managers.

While each session was designed to attain specific goals, the overarching goal was to both distill the critical reporting needs of the Academic Affairs leadership team and determine the best method to provide timely and actionable data. Throughout these sessions, all voices were heard. Individual feedback was collected through a shared spreadsheet. After the feedback was collected, the analytics team organized the feedback by theme and potential audiences (Personas). The resulting Personas were used to create a tracking sheet that allowed the Academic Affairs team to review past feedback and check on the project status. The order in which the project progressed was based upon institutional priorities as defined by the Academic Affairs leadership team.

Iterative and continual interaction has been critical in maintaining momentum on both sides. This continual interaction ensured the resulting reports were built in alignment with Academic Affairs Personas and initial feedback. It also enabled the Academic Affairs teams to utilize portions of the final reports prior to their full completion. By providing intermediary reports, individual teams were able to drive impact sooner than a traditional rollout. The iterative nature also allowed for educational opportunities regarding the reporting tool, data structures, and ideas for usage.

5. RESULTS

A resonating message heard from all levels of the Academic Affairs administration throughout this process was a clear need for actionable reporting. The legacy reporting required waiting for a full term to complete, resulting in reactionary responses. New reports were required to allow the Academic Affairs teams to become proactive in managing their organization. Although the first priority was to assess ongoing course success and trending over time, other themes often overlapped. This overlap

increased the speed of report development and adoption. The other themes included program health and assessment of learning outcomes.

Suites of reports were developed and are under development. Each of these suites aimed to provide a depth of reporting never seen by the Academic Affairs teams. Initial success has been seen through the deployment of the first of many report suites - course completion. The resulting report suite was crafted to tell a story which resonated with each of the Personas. A user of the report starts with a landing page designed to surface critical information, spurring further analysis. Next, the user sees a high level summary incorporating drill-down filter ability to further investigate the critical issues surfaced on the landing page. Additional layers are added as appropriate to allow the user to investigate, providing customization by the end users.

Using course completion as an example, the landing page surfaces courses under a predetermined threshold. This threshold was set by a combination of course subject area and level, smoothing variation. Next, the user can see all course completion rates and trending information for the past three years. Finally, the user can investigate the grade distributions of every course. The idea, again, is to allow the user to investigate from a high level (e.g., all courses), pick those of interest, and diagnose the drivers that may be causing an increase or decrease in course completion rates.

Further, each step was designed with a Persona in mind. For instance, the Deans can readily view school level summaries, while the Associate and Vice Deans can drill into program and specialization area summaries. Program Chairs, Program Managers, and Assessment Managers can drill to the individual assignment level data to ascertain the drivers of improving or declining course completion rates.

6. SUMMARY

The objective of this case study is to improve the dashboard experience with each user. There is a plethora of information captured in higher education, and it needs to be synthesized in a clear way that can support important decisions that are made to impact the institution. Currently, the focus is on dashboards related to course success that will allow deans and program chairs to analyze trends of results to determine potential interventions as well as course structure changes. Final feedback on the process will be collected from all users after each major report theme launch. These results will be integrated into revisions of the currently deployed reports, creation of new drafts, and additional development in the future.

7. AUDIENCE TAKEAWAY

At the conclusion of this presentation, the audience should understand that challenges that one institution faced when trying to improve the existing visual dashboard as well as the solutions that were designed to overcome those challenges. Audience members can take this experience and implement these solutions it to their own organizations as applicable.

REFERENCES

- Aljohani, N.R., Daud, Al., Abbasi, R.A., Alowibdi, J.S, Basher, M., & Aslam, M.A. (2018). An integrated framework for course adapted student learning analytics dashboard. *Computers in Human Behavior*, xxx, 1-12.
- Bill & Melinda Gates Foundation (2015). Teachers know best: Making data work for teachers and students. Retrieved from <https://s3.amazonaws.com/edtech-production/reports/Gates-TeachersKnowBestMakingDataWork.pdf>.
- Fernandez, M., McClain, D., Brown-Williams, B. & Ellison, P. (2015). PBIS in Georgia Department of Juvenile Justice: Data dashboard and radar reports utilized for team data-based decision making with facility team leader perspectives. *Residential Treatment for Children*, 32, 334-343.
- Ndukwe, I.G., Daniel, B.K., & Butson, R. (2018). Data science approach for simulating educational data: Towards the development of teaching outcome model (TOM). *Big Data and Cognitive Computing*, 2,24,1-18.
- Picciano, A. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, 16(3), 9–20.
- Qu, H., & Chen, Q. (2015). Visual analytics for MOOC data. *IEEE computer graphics and applications*, 35(6), 69–75.
- Sedrakyan, G., Malmberg, J., Verbert, K., Jarvela, S., & Kirschner, P.A. (2018). Linking learning behavior analytics and learning science concepts: Designing an learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, xxx, 1-15.
- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 30.
- Smith, M. (2015). Output from statistical predictive models as input to eLearning dashboards. *Future Internet*, 7, 170-183.
- Smith, V. S. (2013). Data dashboard as evaluation and research communication tool. In T. Azzam & S. Evergreen (Eds.), *Data visualization, part 2. New Directions for Evaluation*, 140, 21–45.
- Thomas, J. J., & Cook, K. A. (2005). Illuminating the path: The research and development agenda for visual analytics. IEEE Computer Society.

Vatrapu, R., Teplov, C., Fujita, N., & Bull, S. (2011). Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. Proceedings of the 1st international conference on learning analytics and knowledge (pp. 93–98). ACM.

Viberg, O., Hatakka, M., Balter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98-110.

Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122, 119-135.

Identification of sample comparability issues during the iterative design of game-based cognitive assessments

Rebecca Kantar, David K. Laing, Matthew A. Emery, Sonia D. Doshi, Yao Xiong, and Erica L. Snow

Imbellus
info@imbellus.com

ABSTRACT: Imbellus develops cognitive assessments designed to evaluate the skills and abilities of its users within the context of game-based simulations. This paper describes work from our ongoing collaboration with a best-in-class management consulting firm, McKinsey & Company, to build a simulation-based assessment that evaluates applicants' cognitive skills and abilities. A major challenge inherent in producing a game-based cognitive assessment in an industry setting is to balance the need for iterative design improvements and the need for validation with large samples. The work presented here describes a four-step process for assessing the impact of a given design change on the comparability of pre-change and post-change samples. We also discuss implications and future work.

Keywords Cognitive Assessments, Iterative Design, Score Validation

1 INTRODUCTION

Imbellus develops cognitive assessments designed to evaluate the skills and abilities of its participants within the context of game-based simulations. Imbellus assessments evaluate how people think instead of what they know. We design each assessment to a client's unique work environment. The work presented in this paper will focus on our ongoing collaboration with a best-in-class management consulting firm, McKinsey & Company. These assessments evaluate McKinsey & Company's incoming candidates on their cognitive skills and abilities. Each task within our assessments requires participants to engage in a series of problem-solving challenges that are representative of McKinsey & Company's work environment. As such, Imbellus assessments provide opportunities for candidates to exhibit capabilities required for success on the job. These tasks are set in multiple contexts within an abstracted natural world environment (involving varying terrain, plants, and wildlife) to evaluate meaningful learning of problem-solving capabilities through a far transfer application (Perkins & Salomon, 1992).

2 OVERVIEW OF THE PATHOGEN TASK

The focal point of our analysis is one task within Imbellus assessments referenced as the Pathogen Task. The Pathogen Task is designed to measure problem-solving constructs such as decision-making, critical thinking, and situational awareness. We conducted an extensive literature review to define and operationalize these constructs among the broader problem-solving ontology that formed the foundational layer of the Imbellus assessments. This theoretical problem-solving framework and an evaluation of the firm's nature of work informed our task design. We conducted a cognitive task analysis, in partnership with the firm, to understand the problem-solving domain, as it relates to their work. This analysis ensured structural parity of our problem-solving framework to our domain of interest (Shraagen et. al, 2000).

In the first version of this task, participants engage in the following steps: 1) diagnose an infected species in a desert environment, 2) select a viable treatment method for this species, 3) identify critical populations for whom to prioritize treatment, and 4) calculate dosages of treatments to be deployed. The final phase of this task requires participants to input formulas and solve for dosages using quantitative data provided in a table. There are six formulas and solutions that participants must input through text-entry boxes. Specifically, participants calculate three groups of chemical and solvent formulas and solutions that make up the whole dosage. We evaluate participants on their accuracy and their editing process for every formula entered. Within the Pathogen Task, we created a total of 23 item scores. These scores are designed to capture the participants' cognitive processes stealthily. In this paper, the Pathogen Task's scores are referred to with anonymized IDs, as their exact content is out of scope.

3 CURRENT STUDY

There is a significant challenge inherent in producing a simulation-based cognitive assessment in an industry setting: during the assessment's pilot phase, the iterative cycles of design often interfere with validation efforts. For example, test-takers may give feedback that a given component of the assessment is suboptimal, perhaps due to the instructions, the user interface, or even the core mechanics of the task. This realization may lead to the implementation of design improvements, but may consequently introduce difficulties in comparing samples of candidates whom we assessed before and after the change. The literature on methods of diagnosing the effects of changes in simulation-based cognitive assessments is limited. In this study, we describe a four-step process for evaluating the impact of a given design change on the comparability of scores from the pre-change and post-change samples. We demonstrate how this process helped us diagnose anomalous behavior in one of our predictive models and to reach a deeper understanding of the Pathogen Task's 23 item scores.

4 METHOD

As we piloted the Pathogen Task, we noticed that one of our predictive models was assigning every new participant a score above the previous median. These results were a cause for concern, as we had no strong reason to believe that the newer participants had significantly different characteristics than those whose results had been used to train the initial ranking model. Instead, we suspected that a critical test condition had changed at some time in recent months, which was impacting the score ranges in our assessment. Since the pilot phase of our assessment occasionally involved small updates to the assessment design, we suspected that the anomalous predictions were occurring due to one of these changes. To diagnose this problem, we used a four-step process:

1. Identify an assessment design change that may have caused the anomalous predictions.
2. For each item score, test the null hypothesis that the pre-change and post-change samples have the same mean values.
3. For the item scores that are significantly different, investigate the lower-level behavioral patterns that are relevant to the computation of those item scores.

4. Assess the differences in behavioral patterns in light of the design change that was identified in step 1.

The sample we used includes 1232 participants, with 66.7% of them participating in the pre-change build and 33.3% in the post-change build. Although the sample size is imbalanced for pre- and post-change builds, the English proficiency level and life-long game experience are similar for the two groups.

4.1 Step 1. Identify a candidate design change that may have caused the anomalous predictions.

We made the following changes to the Pathogen Task:

1. Refined and edited the instructional text for typos and confusion in the dosage calculation section.
2. Changed the Guidebook (i.e., help center stores environmental information) to always open to the Index page so that participants could navigate to a page about pathogen species types and understand how to access it.
3. Removed a UI element to filter data on one phase of the task because it was widely unused.
4. Fixed a bug that was causing the task to restart unexpectedly.

We thought that the refinement and editing of the instructions in the dosage calculation section of the Pathogen Task were directly associated with the anomalous predictions. Because these scores were repeated measures of performance on the same tasks with different numeric inputs, they correlated with each other. This correlation led us to believe that the instructional text change may have reduced the rate of repeated penalization for a single misunderstanding. This repeated penalization could explain why a seemingly small design change could have had an outsized impact on the behavior of our model.

4.2 Step 2. For each item score, test the null hypothesis that the pre-change and post-change samples have the same mean values.

Next, we evaluated which items were truly different before and after the instructional text change. A standard *t*-test would not be appropriate because some of our scores do not follow a normal distribution. Instead, a Wilcoxon–Mann–Whitney test was used to determine independence (Holthron, 2008). We corrected the resulting *p*-values for multiple comparisons using the Bonferroni method (“Bonferroni Correction”).

4.3 Step 3. For the item scores that were significantly different, investigate the lower-level behavioral patterns that were relevant to the computation of those item scores.

Dosage calculations accounted for fifteen of the seventeen significantly different item scores in the Pathogen task. Moreover, all the fifteen average post-change scores were higher than the average

pre-change scores. Our goal became to determine whether the pre-change participant population were penalized at different rates for making repeated errors of the same type.

4.4 Step 4. Assess the differences in behavioral patterns in light of the design change that was identified in Step 1.

We found that the post-change participants made fewer errors than the pre-change participants. Of the errors the post-change population made, a smaller percentage of errors were distinct across multiple inputs. An example of a distinct error type is to omit a multiplication factor related to the stage of the pathogen infection. In the pre-change sample, 64% of participants made an error of the same type more than once, whereas only 27% of participants made an error more than once in the post-change sample. This finding supported our hypothesis that the change in instructional text in this section was unfairly favoring the participants in the later sample. We present the detailed results in the next section.

5 RESULTS

After the design changes, fifteen Pathogen Task scores were significantly different (see Figure 1). Of these scores, all but the first two are computed directly from the dosage calculation section, where the instructional text change was made.

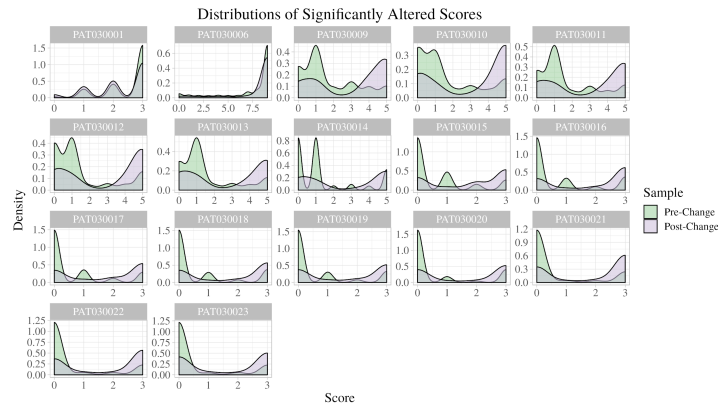


Figure 1: Distribution Comparison of the Significantly Altered Scores

The primary challenge in comparing the pre-change and post-change behaviors that led to the dosage calculation scores was that we needed to convert free-form equation inputs into generalized descriptions of error patterns that we could compare across different numeric prompts. For example, each response required converting a ratio to a fraction, but not all responses started from the same ratio. We needed to be able to compare equation inputs from prompts with different ratios so that we could recognize when both inputs represented the same computational error even though the prompt used different numbers. To solve this problem, we wrote an R script that evaluates equation inputs as mathematical expressions and matches the resulting value to a predefined list of common computational errors. This script allowed us to account for and describe approximately 75% of the text inputs we collected from 1232 test-takers.

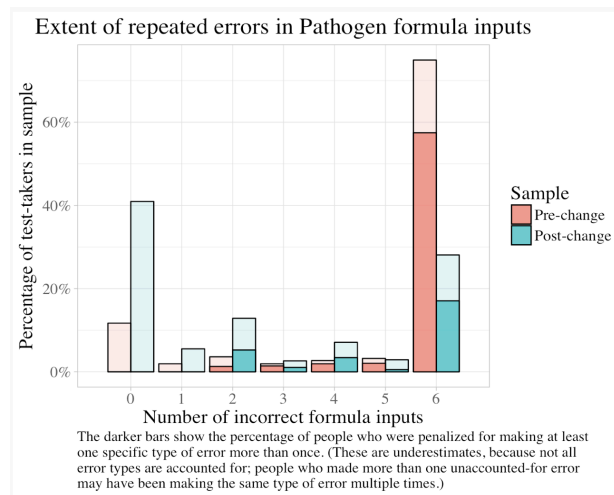


Figure 2: Pathogen Formula Error Frequencies

The most striking difference between the pre-change and post-change samples was that in the pre-change sample, over 75% of participants made errors on all six formula inputs, whereas in the post-change sample, only 28% of participants made errors on all six inputs (see Figure 2). This difference alone would likely have been enough to account for the anomalous predictions made by our model, but there was a further difference that was critical from the perspective of learning science theory. Namely, in the pre-change sample, almost 60% of participants made errors on all six inputs and also made the same type of error more than once. By contrast, less than 20% of participants in the post-change sample made errors on all six inputs and also made the same type of error more than once. These results mean that the participants who were assessed with the pre-change sample were penalized much more heavily for what may have been a single misconception, as a result of compounding errors.

Results from these analyses revealed significant differences across seventeen item scores. In the pre-change assessment, we found that applicants were significantly more likely to make errors in the Pathogen Task's formula input boxes and that those errors were more likely to be compounded. This fact implies that the UI and instructional changes that occurred added clarity to the Pathogen Task, specifically during the dosage calculation section. We used this information to adjust our predictive models and future task design. To fix the predictive models we aggregated the pathogen formula scores to reduce multicollinearity arising from very similar features. We also added a pre-change / post-change flag in the dataset to help the model distinguish between the two builds. As we develop new tasks for Imbellus assessments in the future, the results of this occurrence have influenced design decisions by limiting the need for text-based formula inputs to restrict the over-penalization of compounding errors.

6 DISCUSSION

Imbellus assessments evaluate how you think, not just what you know, by measuring the cognitive processes of test-takers as they engage in complex problem-solving tasks. Getting tasks right requires multiple iterations of design and score development. The work presented here describes a four-step process for assessing the impact of a given design change on the comparability of pre-change and post-change samples. Specifically, this process helped us diagnose anomalous behavior

in one of our predictive models, and to reach a deeper understanding of the Pathogen Task's 23 item scores. Model improvements and future analyses inform upcoming iterations of our assessments. Identifying this difference in our dataset emphasized the impact that a seemingly minor change can have on an assessment, as a whole. In simulation-based assessments like these, we have found that the interdependencies of assessment design, from text to UI elements, can cause reverberating consequences in the data, as we found here. This process could be applied to other assessments of a similar nature and complexity, more broadly, if the suspected affecting change(s) are isolatable during the diagnosis. If the number of changes made between pre- and post-test were more extensive, it is likely that identifying the change and correcting for it in our model may have required a different diagnostic process.

Our assessments will be used for large-scale field testing of the McKinsey & Company candidates with an expected sample size of over 3000 participants. We will use this sample to test the replicability of our analyses and item score iterations, and also explore the effects of individual differences, geographical effects, and the impact of prior knowledge. In the Spring of 2019, we will begin to operationalize the first generation of Imbellus assessments within the firm's recruiting pipeline. This will provide recruiters with an additional data point that can help them assess applicants for hire.

Beyond replicating and expanding the current Imbellus assessments, we are beginning to develop assessments that go beyond evaluating applicants' problem-solving skills and abilities, to measuring other dimensions of their cognition. In the future, we plan to leverage both simulation-based and traditional text-based assessment formats to take a mosaic approach to understand what a person is like across different dimensions and in different scenarios.

REFERENCES

- Implementing a Class of Permutation Tests: The coin Package | Hothorn | Journal of Statistical Software. (n.d.). <https://doi.org/10.18637/jss.v028.i08>
- Perkins, D. N., & Salomon, G. (1992). Transfer of learning. *International encyclopedia of education*, 2, 6452-6457.
- Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (Eds.). (2000). Cognitive task analysis. Psychology Press.
- Weissstein, Eric W. "Bonferroni Correction." From MathWorld--A Wolfram Web Resource. <http://mathworld.wolfram.com/BonferroniCorrection.html>

LAViEW: Learning Analytics Dashboard Towards Evidence-based Education

Rwitajit Majumdar¹, Arzu Akçapınar¹, Gökhan Akçapınar^{1,2}, Brendan Flanagan¹,
Hiroaki Ogata¹

Kyoto University¹, Hacettepe University²
{majumdar.rwitajit.4a, ogata.hiroaki.3e} @kyoto-u.ac.jp

ABSTRACT: Learning analytics dashboards (LAD) have supported prior finds that visualizing learning behavior helps students to reflect on their learning. We developed LAViEW, a LAD that can be easily integrated with different learning environments through LTI. In this paper, we focus on the context of eBook-based learning and present an overview of the indicators of engagement that LAViEW visualizes. Its integrated email widget enables the teacher to directly send personalized feedbacks to selected cohorts of students, clustered by their engagement scores. These interventions and dashboard interactions are further tracked to extract evidence of learning.

Keywords: BookRoll, LAViEW, Student Engagement, Visual Analytics, Intervention widgets

1 INTRODUCTION

One of the key issues in this data driven era in education is to find evidence of learning from analyzing the log data itself. It would have impact in designing ways to increase the students' engagement, especially for at-risk students who have low motivation to the course. In today's technology enhanced learning scenario we can collect learning logs of students and analyze them. A learning analytics dashboard (LAD) assists easier and useful interpretation by different stakeholders based on the visualized information. Learners can view the different indicators presented in dashboards, triggering them to reflect and examine their learning behavior and learning outcomes (Durall, E. and Gros, B. 2014). The teacher can use LAD to get a pulse of the class and analyze if there is any problem. Typically, we envision that the learning analytics system developer would visualize various indicators based on the data that a particular system gathers and the features that are extracted from them. Then a teacher can identify a problem based on the defined indicators. For example, in our context BookRoll is an e-book reader and an issue of low engagement may be indicated in terms percentage completion of content. The teacher sets the level of indicators to identify any problem. For instance, a completion lower than 60% may be considered low engagement for that content.

Currently none of the LADs capture this preference of the teachers to relate problems and indicators and assist them to plan interventions directly from the dashboard. This paper presents LAViEW (Learning Analytics Visualizations & Evidence Widgets), a LAD that supports the users to analyze learning logs and gather evidence of learning. Figure 1 gives sample visualizations in the current version of the LAViEW dashboard. Readers can access the system at live.let.media.kyoto-u.ac.jp/analysis to explore the features with anonymized dataset.



Figure 1: Sample information and visualizations in LAVIEW Dashboard.

1.1 Our LA Framework

We developed our dashboard based on our earlier proposed framework (Flanagan B., and Ogata H., 2017). This framework helps us to collect anonymous learning logs of students. For example, teachers can use a LMS to coordinate a course and upload reading content in BookRoll linked to the LMS. While students use BookRoll for browsing course material, their reading behaviors can be anonymously logged. The eBook system in our context assists instructors to support students' in-class learning activities. It has features to highlight important and difficult to understand text. Students can add memos or bookmark important pages. Learning Logs of eBook reading is recorded in Learning Record Store (LRS) as an eXperience API (xAPI) statements. Next, the analytics engine helps to analyze the log data and extract features and recording in MySQL database. This processed data is visualized in the dashboard. All these processes work in real-time. The framework applies two-way anonymization to the student data. In the logs, students are represented by UUID to ensure their privacy. However, when user logs in to the system via LTI, based on their roles, s/he can see the converted student ids. The framework is also very flexible to connect to any other behavior sensors which has LTI.

This first version of LAVIEW was deployed in October 2017 across 3 universities which used BookRoll, the digital textbook reader, as the learning behavior sensor. In Kyoto university as of 1 February 2019, the LA system had collected 795401 logs about student's reading behavior. The other novelty in the current implementation is the inclusion of a learning evidence extraction system, the evidence portal that captures the interactions of the dashboard users while they monitor learning, analyzing problems, implementing solutions based on the learning widgets in the dashboard and reflecting on the results. The current updated version will be deployed even at school level across several districts in the country from the next school term.

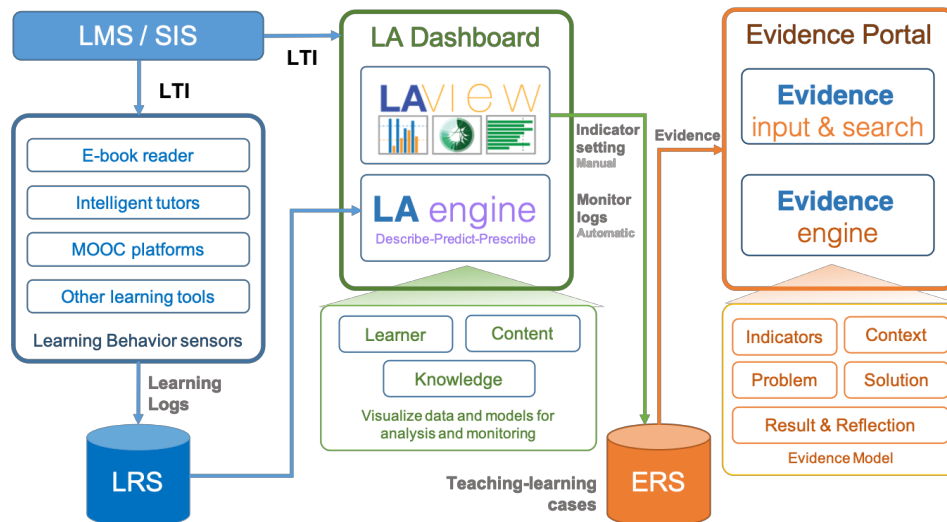


Figure 2: LEAF framework and the LA Dashboard.

2 LAVIEW: LEARNING ANALYTICS DASHBOARD

Our dashboard LAVIEW (Learning Analytics Visualizations & Evidence Widgets) can be added as an external tool in LMS and accessed by both teachers and students. LAVIEW automatically handles the role from the LTI and displays different panels of graphs based on customized views. When the user login to the system, they need to select the content and the period of time they want to analyze from Context Selector panel (see Fig 1 as reference). According to the user's selection the data in every panel is updated. We created an Overview panel which gives aggregated statistics about selected course. In this section both teachers and students can see average statistics of the class on the bottom and selected student's record on the top. Overview information is split into four groups of information, each group having specific color which is also used in the title of the graphs which belong to that group. The current dashboard provides information regarding *Learners & Content*, *Engagement*, *Learning Traces* and *Learning Outcome*. User gets the number of students and pages of the ebook in the selected course. For engagement we visualize indicators such as time spend on eBook, completion percentage of the content, average engagement rating of the class or selected student, and total number of interactions that students made. Learning Traces are the interactions that the students do with the BookRoll content to create annotations like yellow and red marker highlights, memos written or bookmark put. The learning trace section gives the count and the content related to each trace. For Learning Outcome, we link the performance scores gathered in the LMS and the knowledge points based on our content-knowledge (Flanagan, B., Majumdar, R., Akçapınar, G., Wang, J. and Ogata, H. 2018).

The dashboard contained charts with information on display or as pop ups on click interactions. To assist users, we added an overlay panel to every graph which gives explanation about each graph to the users. Additionally, such an implementation also helps to track usage of graphs and collect data for researching regarding visual approaches in LAD in terms of effectiveness, efficiency or other criteria that pertain to learning (Klerkx, J., Verbert, K. and Duval, E. 2014). The opensource implementation of LAVIEW APIs also makes it easy to add novel visualizations and widgets to the dashboard. Thus, it has the potential to visualize collected data from multiple data sources that are connected through LTI.

In this paper we consider the teacher as our primary user and conceptualize the following user goals for the LAVIEW. (1) monitoring a class of students, (2) provide feedback and intervention through dashboard, (3) increase engagement of students.

3 SUPPORTING ACTIONABLE ANALYTICS WITH LAVIEW: ILLUSTRATION OF AN INTERVENTION FOR LOW ENGAGEMENT

3.1 Purpose

Increasing students' engagement is one of the important features to increase students' success. However; especially at-risk students who are disengaging from coursework, it is difficult to identify students' engagement in large class sizes (Field, J., Lewkow, N., Burns, S. and Gebhardt, K. 2018). For this reason, using computers, to 'observe' students 'in situ', that is, while students are occupied in learning activities is an appropriate way to measure the engagement. Systems like dashboards with potential to visualize large amounts of data about students' behavior, is being harnessed to improve learning interactions and to personalize the learning experience (Liu, M. 2015). We enable teachers to identify at-risk students based on their engagement score and integrated intervention widget of emailing that helps to send clear guidance on how to improve personalized for different cohorts of students.

3.2 Design

According the student's usage of BookRoll, we created 9 indicators to define their engagement. An aggregated value is computed as an Engagement Score. That Engagement Score is visualized in three Engagement Graphs as shown in Fig. 3. The first one (Fig 3a.) visualizes the breakdown of the Engagement Score showing parameters value of each of the nine indicators. The number in the center (43) is the overall engagement score. To see the value of each indicator you can hover on each segment on the donut graph. Leaderboard (Fig. 3b.) is a table that users can see engagement score of the all students in the class and their ranking among other students. Weekly Engagement graph (Fig. 3c) visualizes the engagement score computed across the activity in that week. It can be used to compare individual's weekly engagement with average class engagement by comparing the point values in that week. Further looking at the lines gives a temporal trend. Green line shows the average score of the class and red line shows the student's score.

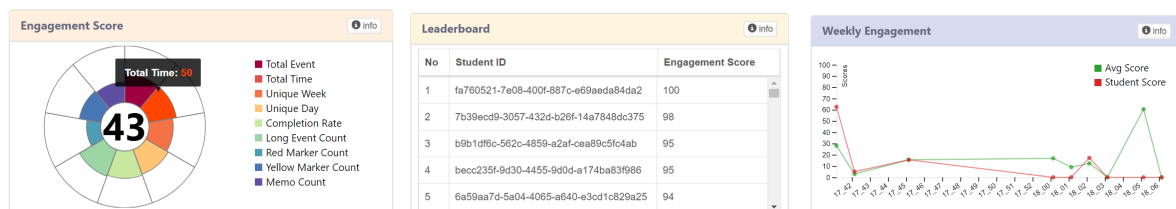


Figure 3: a. Engagement Score b. Leaderboard c. Weekly Engagement

3.3 Email Intervention Widget

Based on the engagement score LAVIEW also affords the teacher to plan intervention such as sending emails (see Fig4. For the currently implemented version of sending email). The system automatically clusters 3 different cohorts of students, *Good*, *OK* and *At-risk*. Teacher can select the

students in that cohort and send them a personalized email. Then after a chosen period of time, the teacher can receive a report regarding the indicators to assess the result of the intervention. The interface is presented in Figure 4 and workflow of the instructor is presented in Figure 5..

Figure 4: Email widget in LAVIEW

3.4 Extracting evidence from teaching-learning practice

We propose an evidence portal (Majumdar, R., Akçapınar, A., Akçapınar, G., Flanagan, B. and Ogata, H. 2018.) which would have the provision to record all the information that is part of the above workflow. It records the criteria of the classification of students from the learner model, the teacher can input additional details of the context, the description of the indicators of the problem, the solution plan of intervention regarding this case and its result. The ERS stores this as a single record along with the automatically linked context from the LMS and the search parameters of the LAVIEW. We call each record as a teaching-learning case (TLC). Context anonymized dataset in the LRS can be used to retrieve the whole case details during evidence search. We give a sample xAPI data that would be stored in the ERS corresponding to the email sending activity in Figure 6.

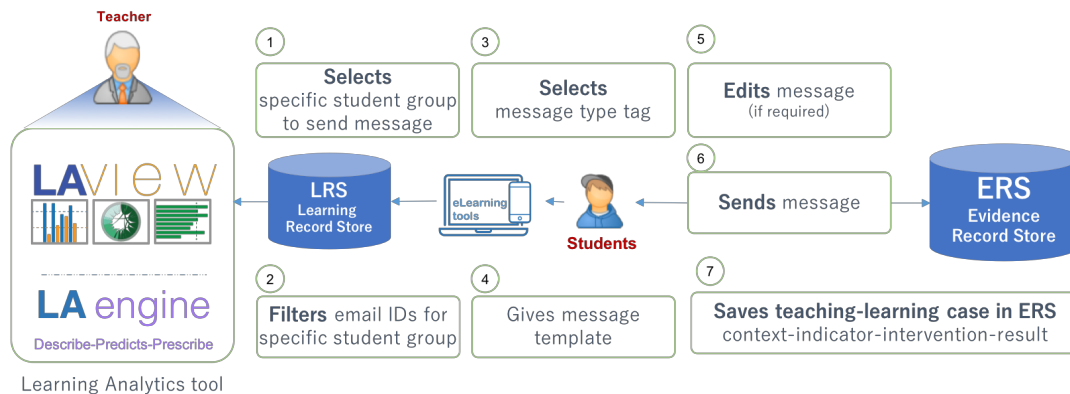


Figure 5: Workflow for intervention

xAPI Statement

```

{"version": "1.0.0",
 "actor": {
  "objectType": "Agent",
  "account": "User account",
  "homePage":
    "LAVIEW Dashboard",
  },
 "verb": {
  "id": "Planning",
  },
 "object": {
  "objectType": "Activity",
  "id": "#URI#",
  "definition": "Contents Name",
  "description": "Contents ID",
  "extensions": {
    "course_detail": "Course Detail",
    "content_name": "Content Name",
    "indicator":
      "Visualization Name",
    "problem_class":
      "problem engagement",
    "problem_description":
      "Low engagement",
    "solution_class":
      "Low Percentage Completion",
    "solution_description":
      "E-mail Intervention",
    "reflect_description":
      "Solution description",
    "reflection_description":
      "Reflection description",
  },
  "result": {
    "extensions": {
      "result": {
        "result_description":
          "Result description",
        "report_link": "Result link",
        "rating": "Rating",
        "timestamp": "Timestamp",
      },
    },
  },
}

```

Figure 6: Sample structure of the xAPI log of Teaching-Learning Case in the ERS.

Our approach to commence an evidence-based practice in education supported by technology starts with systematically gathering indicators of learning in a specific scenario and then analyzing visualized indicators in the analytics dashboard to identify problems (Ogata, H. et.al. 2018). Teacher can design intervention to mitigate it and then monitor its effectiveness. We believe technology can help to capture this process and reflect on the effectiveness of the practice as evidence. Conceptualizing such an evidence analytics system in education would push the boundaries of existing learning analytics infrastructures towards a technology-enhanced and evidence-based education and learning.

ACKNOWLEDGEMENT

This research was partly supported by JSPS Grant-in-Aid for Scientific Research (S) Grant Number 16H06304, Research Activity Start-up Grant Number 18H05746 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

REFERENCES

- Durall, E. and Gros, B. (2014). Learning Analytics as a Metacognitive Tool. *Procs. of the 6th Int. Conf. on Computer Supported Education - Volume 1*, Portugal, 2014, 380–384.
- Flanagan, B. and Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. *Procs. of the 25th Int. Conf. on Computers in Education*, pp.333-338,
- Flanagan, B., Majumdar, R., Akçapınar, G., Wang, J. and Ogata, H. (2019). Knowledge Map Creation for Modeling Learning Behaviors in Digital Learning Environments. *Companion Procs. of 9th LAK*.
- Klerkx, J., Verbert, K. and Duval, E. (2014). Enhancing Learning with Visualization Techniques. *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop, Eds. New York, NY: Springer New York, 2014, pp. 791–807.
- Field, J., Lewkow, N., Burns, S. and Gebhardt, K. (2018). A Generalized Classifier to Identify Online Learning Tool Disengagement at Scale. *Procs. of the 8th LAK*, New York, NY, USA, 2018, pp. 61–70.
- Liu, M. (2015). Using Particle Swarm Optimization Approach for Student Engagement Measurement. *International Journal of Learning, Teaching and Educational Research*, vol. 11, no. 1, Apr. 2015.
- Majumdar, R., Akçapınar, A., Akçapınar, G., Flanagan, B. and Ogata, H. (2018). Learning Analytics Dashboard Widgets to Author Teaching-Learning Cases for Evidence-based Education, *Companion Procs. of 9th LAK*.
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M.N. and Flanagan, B. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning, *Proceedings of the 26th Int. Conf. on Computers in Education*, pp. 486-489, 2018.11.26.

Towards Enhancing Conceptual Knowledge in Algebra through Diagrammatic Self-explanation in an Intelligent Tutoring System

Tomohiro Nagashima

Human-Computer Interaction Institute, Carnegie Mellon University
tnagashi@cs.cmu.edu

ABSTRACT: In mathematics education, it is still unclear how instruction can support learning conceptual knowledge, procedural knowledge, and connections between them. Particularly, research suggests that it is harder to acquire conceptual knowledge than procedural knowledge in algebra. Tape diagrams, a representation used to visualize a relationship between quantities in an equation, have been studied to explore their potential benefits in supporting conceptual knowledge; however, their effectiveness is still not entirely clear especially for low-ability students. To effectively foster students' conceptual knowledge in algebra, we propose a novel instructional approach integrating self-explanation into the use of tape diagrams in an intelligent tutoring system. The proposed study will design and test this approach, called diagrammatic self-explanation, where students manipulate tape diagrams as a way of self-explanation. The study will make contributions to the fields of learning analytics and the learning sciences by 1) establishing an effective instructional strategy using the combination of tape diagrams and self-explanation and 2) designing an adaptive tutor for equation solving with tape diagrams.

Keywords: Conceptual Knowledge, Tape Diagrams, Self-explanation, Equation Solving, Algebra, Intelligent Tutoring Systems.

1 INTRODUCTION

1.1 Conceptual Knowledge in Mathematics

One of the biggest challenges in mathematics education is how to support learning conceptual knowledge (CK), procedural knowledge (PK), and connections between them (Schneider, Rittle-Johnson, & Star, 2011). A widely-accepted view is that the development of CK and PK is interactive, where the development of one type of knowledge leads to the other and vice versa (Rittle-Johnson & Schneider, 2014). Yet, it has been found more difficult to gain CK compared to PK in some mathematics fields, including algebra (Matthews & Rittle-Johnson, 2009). Despite the importance of fostering CK, however, current teaching practices are too often focused on teaching procedures without offering explanations on conceptual understanding of such procedures (National Council of Teachers of Mathematics, 2014).

Conceptual knowledge is a complicated notion for which researchers have adopted a variety of definitions. Crooks and Alibali (2014) suggest using two types of definitions: “general principle knowledge” and “knowledge of principles underlying procedures” (p. 366). The

former refers to the fundamental and general knowledge about the domain, such as rules and definitions whereas the latter involves “knowing why certain procedures work for certain problems and knowing the purpose of each step in a procedure” (Crooks & Alibali, 2014, p.367). We will adopt these definitions of CK in the proposed study.

1.2 Tape Diagrams

One promising way of fostering CK in mathematics instruction is the use of diagrams. Diagrams and other types of external representations have been studied extensively in mathematics education with their effects generally proven to be positive (Mayer, 2005). In algebra, one type of diagram that is thought to be helpful for students is tape diagrams (Murata, 2008). Tape diagrams visually depict relationships among quantities in an equation problem (Figure 1). They are consistently used in mathematics instruction in Japan and Singapore, two of the countries where students perform far better than the world’s average on international mathematics tests (Murata, 2008). In the United States, studies have empirically demonstrated the benefits of the presence of tape diagrams on learning among middle school students (Booth & Koedinger, 2012; Chu, Rittle-Johnson, & Fyfe, 2017; Koedinger & Terao, 2002). However, it is also suggested that prior knowledge may mediate the benefits: several studies indicate that low-ability students were incapable of translating their conceptual understanding of the quantitative relationship across different representations (e.g. tape diagrams and algebraic equations, tape diagrams and word problems) (Booth & Koedinger, 2012; Chu, Rittle-Johnson, & Fyfe, 2017). This implies the need for more careful design and additional scaffolds to support not only high-ability students but also low-ability students.

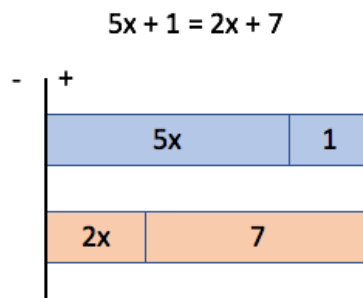


Figure 1: An example of a tape diagram representation. Tape diagrams visualize the relationship between the quantities in the equation.

1.3 Self-explanation

Another potential approach to boosting CK is self-explanation. Self-explanation is an instructional strategy in which students generate explanations in order to understand newly-introduced information by connecting it with their prior knowledge (Rittle-Johnson, Loehr, & Durkin, 2017; Wylie & Chi, 2014). It has consistently been shown effective in a variety of domains both in a paper-and-pencil format and in a computer environment (Wylie & Chi, 2014). In mathematics education, research has illustrated that self-explanation is effective in promoting CK and PK (Rittle-Johnson et al., 2017), but it has also been found that the effectiveness of self-explanation in promoting CK is limited in a classroom context (Rittle-Johnson et al., 2017). In an effort to facilitate robust learning in an authentic context,

however, researchers have studied and demonstrated that intelligent tutors can effectively enhance students' CK, particularly through scaffolding self-explanations via providing a list of possible responses or self-explaining in a fill-in-a-blank form (e.g. Rau, Aleven, & Rummel, 2015). This suggests that intelligent tutoring software, when designed appropriately, can meaningfully foster CK through self-explanation prompts. As self-explanation can potentially support students' understanding of tape diagrams and to help them connect different representations, it is worthwhile to explore the potential of integrating self-explanation into tape diagram use.

The proposed study will make contributions to the fields of the learning sciences and learning analytics by 1) establishing an effective instructional strategy involving self-explanation in equation solving with tape diagrams and by 2) designing and evaluating an adaptive tape diagram tutor which adapts scaffold/prompt types and levels to students' prior knowledge.

2 PROPOSED STUDY AND RESEARCH QUESTIONS

Towards developing an effective instructional strategy for enhancing CK, we propose a novel approach of integrating self-explanation into the use of tape diagrams for equation solving. In doing so, we will use tape diagrams not as an additional representation to algebraic equations or word problems, but rather as an active constructive activity which we call diagrammatic self-explanation, where students are asked to manipulate tape diagrams when solving an equation. We will develop diagrammatic self-explanation activities in Lynnette, a web-based intelligent tutor for equation solving that supports practices across problems of varying difficulty (Long & Aleven, 2017). As is common in Intelligent Tutoring Systems (ITS), Lynnette offers step-by-step personalized guidance based on students' inputs, and personalized problem selection.

In diagrammatic self-explanation, it is hypothesized that manipulating tape diagrams can help develop CK, such as the concept of mathematical equivalence (Matthews, Rittle-Johnson, McEldeen, & Taylor, 2012) and variables, through visualizing quantitative relationships. Figure 2 illustrates an example of diagrammatic self-explanation. In this example, students are asked to manipulate the tape diagram following the transformation steps shown on the left and identify an error.

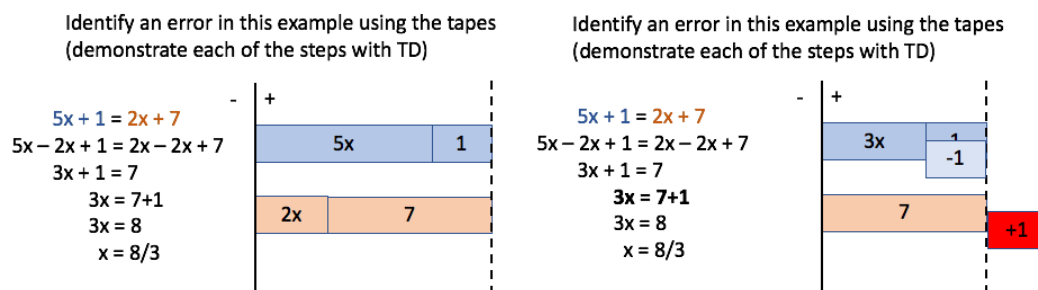


Figure 2: Diagrammatic self-explanation with an incorrect example. The tape diagrams on the left show $5x + 1 = 2x + 7$ (original equation) and the diagrams on the right show $3x = 7 + 1$ (intermediate step) .

To the best of our knowledge, the instructional potential of “manipulable” tape diagrams in equation solving has not been investigated to date, presumably due to the difficulty of manipulating tapes on paper. Also, in equation solving (e.g. Looi & Lim, 2009), tape diagrams have traditionally been used in the problem representation phase (e.g. constructing tape diagrams given a word problem) but not in the problem solution phase (Mayer, 1985). As constructive activities such as sketching and drawing have been shown effective for learning (e.g. Wu & Rau, 2018), we believe that diagrammatic self-explanation, where students manipulate tape diagrams in the problem solution phase, can be beneficial to students. Our proposed study will address the following research questions:

- *Will diagrammatic self-explanation enhance students’ CK?*
- *Will diagrammatic self-explanation enhance low-ability students’ CK?*

3 RESEARCH PLAN AND CURRENT STATUS

3.1 User-centered Investigation on the Use of Tape Diagrams

Despite their popularity and success in Asian countries, tape diagrams are not necessarily a familiar representation in other countries, including the United States (Murata, 2008). In order to explore the design, potential instructional impacts, and any difficulties associated with tape diagrams, we will first conduct qualitative user research with teachers and students in middle schools in the US. Specifically, we will conduct interviews and task analysis, followed by iterative prototyping co-designing with teachers and testing of the prototypes in Lynnette with students. The findings would inform the instructional design of diagrammatic self-explanation and necessary training for teachers and students, as well as the design of the ITS interface.

3.2 Planned Experimental Study

Once we have perfected our design through prototyping, we will conduct an experiment examining the effectiveness of diagrammatic self-explanation in equation solving.

3.2.1 Participants and Materials

Seventh- and eighth-grade students at schools in the US will participate in the study, which will take place as part of their regular mathematics instruction. Equation solving activities will be prepared in Lynnette. We will also develop pre- and post-assessments on CK in equation solving based on past studies with Lynnette as well as from mathematics education literature.

3.2.2 Study Design

The proposed study will conduct an *in vivo* experiment (i.e. a rigorously controlled experiment in a natural classroom setting) examining whether the diagrammatic self-explanation can improve students’ performance on CK. Students will be randomly assigned to either of three conditions. The first condition involves diagrammatic self-explanation as a way of solving algebraic equations. Students in the second condition will solve algebraic equations but tape diagrams will be shown as an additional reference to the algebraic equations, rather than as manipulable tape diagrams. In the third condition, students will solve algebraic equations without tape diagrams.

3.2.3 Procedure

Students will first be asked to work on the pre-test in the ITS. After being assigned to either of our three conditions, they will be asked to solve the equation problems with diagrammatic self-explanation prompts or no prompts. All groups will receive the problems at the same difficulty level and the total amount of time spent will be matched across the conditions. Following that, students will be asked to complete the online post-test.

3.3 Expected Results

We hypothesize that students in the tape diagram conditions (first and second conditions) will perform better than the no tape diagram condition.

Regarding our second research question on individual differences, when we compare the results from the first and second conditions, we expect to see an interaction effect between math ability, assessed by the pre-test, and the type of tape diagram use. Specifically, we expect that diagrammatic self-explanation will help both low-ability and high-ability students while using tape diagrams as a reference (no manipulation) will only be effective for high-ability students.

3.4 Current Status

We have reviewed the literature on related topics and have started the qualitative investigation of the design of the interactive tape diagrams with teachers and students.

4 FUTURE PLAN: ADAPTIVE TAPE DIAGRAM TUTOR

Our proposed study will test whether diagrammatic self-explanation can promote the learning of CK in equation solving. Although what follows might change depending on the results of our first study, we plan to develop an adaptive tape diagram tutor, which would vary the prompt/scaffold type and type of tape diagram representation based on students' level of CK, because it is likely that students' individual differences in prior knowledge influence whether and how much students benefit from the use of tape diagrams.

To explore this approach, we will investigate whether we can use students' interaction data from Lynnette to identify their levels of CK when they work on activities. Once we identify their levels of CK, we would be able to provide different diagrammatic scaffolds or activities to avoid over-scaffolding or under-scaffolding and to meaningfully support students' conceptual understanding in Lynnette (e.g. constructing tape diagrams from a list of possible tape options, selecting a correct tape diagram representation from the list of possible answers). We believe that our first study will provide the foundation for the idea of the adaptive tutor.

REFERENCES

Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem

- solving: Development of diagram use. *British Journal of Educational Psychology*, 82(3), 492–511.
- Chu, J., Rittle-Johnson, B., & Fyfe E. R. (2017). Diagrams benefit symbolic problem-solving. *British Journal of Educational Psychology*, 87(2), 273-287.
- Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, 34(4), 344–377.
- Koedinger, K. R., & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebraic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 24, No. 24).
- Long, Y., & Aleven, V. (2017). Educational game and intelligent tutoring system: A classroom study and comparative design analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3).
- Looi, C. K., & Lim, K. S. (2009). From bar diagrams to letter-symbolic algebra: a technology-enable bridging. *Journal of Computer Assisted Learning*, 25(4), 358-374.
- Matthews, P.G. & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology*, 104, 1-21.
- Matthews, P.G., Rittle-Johnson, B., McEldoon, K., & Taylor, R.T. (2012). Measure for measure: What combining diverse measures reveals about children’s understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education*, 43, 316-350
- Mayer, R. E. (1985). Mathematical ability. In R. J. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 127-150). San Francisco: Freeman.
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Press.
- Murata, A. (2008). Mathematics teaching and learning as a mediating process: the case of tape diagrams. *Mathematical Thinking and Learning*, 10(4), 374–406.
- National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston VA: Author.
- Rau, M. A., Aleven, V., & Rummel, N. (2015). Successful learning with multiple graphical representations and self-explanation prompts. *Journal of Educational Psychology*, 107(1), 30.
- Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, 49(4), 599–611.
- Rittle-Johnson, B., & Schneider, M. (2014). Developing conceptual and procedural knowledge of mathematics. *Oxford handbook of numerical cognition*, 1102-1118.
- Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology*, 47(6), 1525–1538.
- Wu, S. P. W., & Rau, M. A. (2018). Effectiveness and efficiency of adding drawing prompts to an interactive educational technology when learning with visual representations. *Learning and Instruction*, 55, 93-104.
- Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd Ed., pp. 413-4320. Cambridge University Press.

Investigating the Effectiveness of Online Learning Environments for Complex Learning

Charlotte Larmuseau

KU Leuven, KU Leuven KULAK campus Kortrijk ; Itec-Imec

charlotte.larmuseau@kuleuven.be

ABSTRACT: It is increasingly relevant that online learning environments reflect the complexity of the real life and deal with authentic, complex tasks. Carrying out complex learning tasks requires students to actively engage in different problem-solving skills. Therefore, it is important that learning environments are designed in a way that complex learning is supported. The effectiveness of learning environments to promote complex learning is dependent of external and internal conditions. External conditions are related to the instructional design of the learning environment. Internal conditions encompass learners' cognitive, metacognitive, affective and motivational characteristics. External and internal conditions should be aligned with each other. To have more insight into how the interrelationship of internal and external conditions influences interactive behavior, multichannel data is incorporated in the different studies consisting of log data, physiological and self-reported data. Findings should provide insight into how the effectiveness of online courses for complex tasks can be supported.

1.1 Introduction

As society and work environments become more interconnected and complex, it is increasingly relevant that online learning environments reflect the complexity of the real life and focus on 21st century skills i.e., *external conditions*. Carrying out complex learning-tasks requires students to actively engage in a dynamic process of analyzing, looking for possible solutions, decision making and implementation. Therefore, educators and instructional designers should realize that students (e.g., novices) need ample instructional support to make their problem-solving process more efficient and effective (Slof et al., 2010). Therefore, in order for online courses for complex learning material to be effective, the instructional design must respond to *learners' internal conditions*, namely, their cognitive, metacognitive, affective, and motivational characteristics. The interrelationship between the internal and external conditions influence students' interactive behavior and subsequently their learning outcomes (Rienties & Toetenel, 2010). Accordingly, to investigate the effectiveness of online learning, we should apply measurement methods to link the interrelationship of learners' *internal and external conditions* with their *interactive behavior* and *learning outcomes* as shown in Figure 1. These findings can give insight into how learners cope with complex online learning material and how they can be supported. Taking into account the goal and the method to achieve that goal, this research project is part of the research field of learning analytics. Specifically, because in the different studies we measure, collect, analyze and report data about learners and their context, for purposes of understanding and optimizing learning and the environments in which it occurs (Long & Siemens, 2011). This research project draws particular attention to the need to align learning analytics with the existing body of research knowledge about learning and teaching in order to understand interactive behavior and how it can be related to students' learning processes. Findings should provide insight into how the effectiveness of online

courses for complex tasks can be supported. To achieve this goal the following research questions are the focus of attention during my PhD project:

- *How does the interrelationship between internal and external conditions influence the quantity and quality of use of an online course and students learning outcomes?*
- *How can the effectiveness of an online course, containing complex learning material, be improved by adapting the instructional design (e.g., selecting adequate support)?*

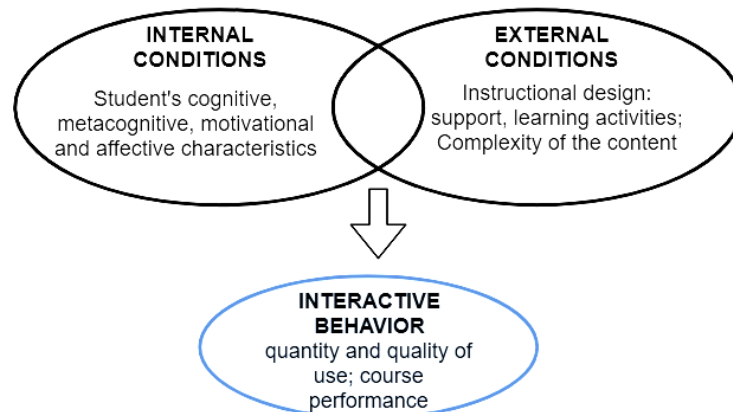


Figure 1: Schematic overview of the theoretical framework

1.2 Theoretical Framework

Complex learning can be defined as the integration of knowledge, skills and attitudes (van Merriënboer, 1997). Different researchers have investigated how learning environments for complex tasks should be designed (Merrill, 2002). van Merriënboer (1997) developed the four component instructional design model (4D/ID-model). The basic claim of 4C/ID is that all environments for complex learning can be described in terms of four interrelated components (1) learning tasks, (2) supportive information, (3) procedural information and (4) part-task practice. Learning tasks are meaningful whole tasks experiences based on real-life tasks from professional or daily life and typically require the integrated use of knowledge, skills and attitudes i.e., complex learning. Learning tasks are the backbone of the educational program to which the other three components are connected. The 4C/ID-model is aligned with the existing body of research knowledge about learning and teaching. Nevertheless, a learning environment that incorporates an instructional design based on a strong theoretical framework does not ensure that the online course will be effective. The *external conditions*, namely the complexity of the task and the instructional design should be aligned with *students' internal conditions*. Firstly, *cognitive characteristics* should be taken into account as complex learning often imposes high cognitive load for novice learners which may seriously hamper learning (Sweller, 2010). This phenomenon can be explained by Cognitive Load Theory (CLT) which uses current knowledge about the human cognitive architecture to develop the instructional design for complex learning. Basically, the human cognitive architecture consists of an effectively unlimited long-term memory, which interacts with a working memory that has limited processing capacity (Sweller, 2010). On the other hand, the content must also be challenging enough to keep the learner motivated. Recent studies claim that a minimal guided task prior to explicit instruction, might also be beneficial for novice learners (Kalyuga & Singh, 2015). It is therefore important to find a balance in which the learner is cognitively challenged without requiring too much mental load (i.e., cognitive overload). As online courses give learners autonomy and control over the time and the

pace at which learners work, students' online self-directed learning (e.g., metacognitive strategies) is especially important. Given this freedom, we assume that learners can manipulate their own cognitive load (e.g., by consulting more or less support; cognitive strategies). Accordingly, cognitive load and self-directed learning appear to be strongly connected in an online learning context (Boekaerts, 2017; de Bruin & van Merriënboer, 2017). Moreover, in the context of online courses, *motivational and affective beliefs* have been identified as critical factors influencing successful learning since the online context poses high demands on learners' motivation and persistence. The importance of motivational and affective beliefs increases even further when dealing with complex tasks as learners will have to be sufficiently motivated and perceive these tasks as useful to deal with the complexity (Larmuseau et al., 2018). Accordingly, to understand *learners' interactive behavior* and how this is related to learning processes, we have to have insight into the interrelationship between students' internal and external conditions (i.e., instructional design and the complexity of the content).

1.3 Discussion of the project and added value

Former research in the field of learning analytics and educational data mining has demonstrated much potential for understanding and optimizing the learning process (Baker & Yacef, 2009). Nevertheless, there are still some areas that are not sufficiently clear. The first shortcoming relates to *the methodology*. Former studies indicate the difficulty to capture learning processes. This research project aims at contributing to the research field in findings methods to measure cognitive load during online complex learning. Accordingly, following studies will incorporate physiological data in order to have a more continuous measurement of learning processes. This will be explored by means of log-data and the manipulation of task complexity. We also want to specifically look for methods to analyze effective use of different components in online courses using log data. Secondly, there is a need for more *theory-informed research*. Computational aspects of learning analytics should be well integrated within existing educational research (Gašević, et al., 2015). The challenge for learning analytics is to establish plausible relationships between models derived from trace data and "learning" that have utility for educators and are interpretable by them. Therefore, is this project we aim at linking interactive behavior (i.e., focus on strategy use) and learning outcomes with well-founded educational theories and theoretical models with a specific focus on the 4C/ID model when designing the instructional design of the online courses and a focus on cognitive load and strategy use (cognitive and metacognitive strategies; Griffin et al., 2013).

1.4 Research methodology and ethical considerations

To investigate these interrelationships the research project will use different sources of data. More specifically, students' interactions with the virtual learning environment will be captured and stored. In my previous studies course activity and time spent were incorporated in the study, but this provides little insight into the learning process itself. In my following studies I would like to investigate effective use of the different components by investigating the coverage of the content needed to solve the tasks, and the effect of use of support on performance. Subjective measurements (i.e., students' motivation, perceived cognitive load etc.) will be retrieved from self-reported data and knowledge tests. For the fifth study we might also consider incorporating think-aloud data in order to have more insight into metacognitive processes. For some studies additional data will be collected from wrist-worn devices measuring physiological aspects of the learners that reflect the sympathetic nervous system (SNS), namely galvanic skin responses (GSR), heart rate variability (HRV; Nourbakhsh et al., 2012). Structural equation modeling (SEM) techniques are used to investigate relationships between students' individual differences, students' use and learning

outcomes (Gašević et al., 2015; Milligan, 2018). The design of the following studies will be within-subject design where we will compare several conditions (while including covariates). Accordingly, we will use repeated measures and/or linear mixed modeling (LMM). Depending on (the amount) of data, predictive modeling is also considered. As we will work with physiological data and log-data in the second research cycle as illustrated in section 1.5 we will submit an application for the Social and Societal Ethics Committee (SMEC).

1.5 Current status of the work and results achieved so far

This project can be divided into two different research cycles: In the *first research cycle*, three studies were conducted. The *first study* investigated the influence of students' acceptance of a 4C/ID based online course on students' use and learning outcomes. Content of the online course was teaching French as a foreign language (i.e., complexity based on the need for integration of skills, attitudes and knowledge; van Merriënboer, 1997), and students could use the course for three weeks. Findings of SEM suggest that students' perceived usefulness of an online course can be used as an indicator of their quantity of use of the online course in an ecological valid context. By contrast, perceived ease of use has no influence on the quantity of use. Furthermore, results show that the quantity of use of the online course has a positive influence on the students' learning gain. The study has been published as: Larmuseau, C., Evens, M., Elen, J., Van Den Noortgate, W., Desmet, P., & Depaepe, F. (2018). The Relationship Between Acceptance, Actual Use of a Virtual Learning Environment and Performance. *Journal of Computers in Education*, 5. 95-111. <https://doi.org/10.1007/s40692-018-0098-9>.

A *second study* investigated the influence of students' cognitive (i.e., prior knowledge) and motivational (i.e., task value and self-efficacy) characteristics on students' quantity of use of the four different components of a 4C/ID-based online course and measured how students' quantity of use of the four components of the 4C/ID model, influence students' learning gain, controlling for students' cognitive and motivational characteristics. Content of the course was learning French as a foreign language and students could use the online course for two weeks. SEM indicates that students' characteristics influenced differences in use of the four components. Students' with higher task value consulted more supportive information. Additionally, students with lower prior knowledge consulted more part-task practice. Furthermore, the use of the learning tasks, procedural information and mainly students' prior knowledge significantly contributed to students' learning gain. This study is published as: Larmuseau, C., Elen, J., & Depaepe, F. (2018). The Influence of Students' Cognitive and Motivational Characteristics on Students' Use of a 4C/ID-based Online Learning Environment and Their Learning Gain. *Proceedings of the 8th International Conference on Learning Analytics & Knowledge - LAK '18*, 171–180. <https://doi.org/10.1145/3170358.3170363>.

A *third study* investigated the influence of the perceived instructional quality on students' acceptance. Moreover, this study investigated the impact of technology acceptance and the perceived instructional quality on both the quantity of use and task performance. SEM indicates that the perceived instructional quality has a significant positive influence on students' perceived usefulness and perceived ease of use. Furthermore, students' perceived instructional quality has a positive influence on task performance, but not on the quantity of use. This study was published as: Larmuseau, C., Desmet, P. & Depaepe, F. (2018). Perceptions of instructional quality: impact on

acceptance and use of an online learning environment. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2018.1509874>.

Findings of study 1 to 3 made it clear that students' learning gain during complex learning were mainly influenced by students' prior knowledge, which might indicate that students with a low level of prior knowledge experienced high cognitive load, which had a negative influence on their learning gain. Accordingly, in the *second research cycle* we aim at conducting two studies where we investigate how we can link cognitive load by manipulating the complexity of several learning tasks and effective use of support in an online course.

In a *fourth study* we aimed at investigating how problem complexity influences self-reported cognitive load in an online course. Participants were 62 future primary school teachers. The complexity of the task was manipulated by increasing the element interactivity for the high complex task (Sweller, 2010). In the low complex task one element was questioned each time, and consequently students had to apply a rule or procedure. By contrast, the high complex task required learners to engage in a series of cognitive activities such as analyzing, decision making, implementing and evaluating, while holding several procedures and rules in mind. In order to solve these tasks effectively, the same amount of support was provided during both tasks (assuming a similar level of extraneous load). The aim of the study was threefold. First, we investigated differences in the experienced cognitive load while solving a high and low complex problem. Secondly, we examined whether students' self-efficacy and strategy use (i.e., retrieved from log-data) influences the different types of perceived cognitive load, controlled for prior knowledge. In a third phase, we investigated the influence of students' perceived cognitive load on task performance, controlled for students' self-efficacy and prior knowledge. There were also 15 students who wore wearables that measured their skin conductance and skin temperature. For those students we studied differences in these physiological data between the high and low complex problem. Moreover, we investigated whether there was a relationship between the physiological and self-reported data. This study "*Multichannel data for understanding cognitive affordances during complex problem solving*" will be presented during the LAK19-conference.

In a *fifth study* we want to investigate if differences in the instructional design in an online course has an influence on cognitive load and effective use of support during high and low complex problem solving. In this study students will have to participate at two study conditions where they have to solve eight statistical problems that differ in complexity. In condition one, they do not receive targeted support (but it is available). In the second condition they get just-in-time procedural information on top of the more general supportive information (i.e., guided condition). We will investigate whether this affects their strategy use (e.g., less or more effective), cognitive load, perceived stress and task performance during low and high complex problem solving. In this study we will again incorporate physiological data. Compared to study 4, we will conduct more self-reported measurements and offer more problems that differ in complexity.

REFERENCES

- Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*. <https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314>
- Boekaerts, M. (2017). Cognitive load and self-regulation: Attempts to build a bridge. *Learning and Instruction*, 51, 90–97. <https://doi.org/10.1016/j.learninstruc.2017.07.001>
- de Bruin, A. B. H., & van Merriënboer, J. J. G. (2017). Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, 51, 1-9. <https://doi.org/10.1016/j.learninstruc.2017.06.001>
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget : Learning analytics are about learning. *TechTrends*, 59, 64-17. <https://doi.org/10.1007/s11528-014-0822-x>
- Griffin, T. D., Wiley, J., & Salas, C. R. (2013). *Supporting Effective Self-Regulated Learning: The Critical Role of Monitoring*. In International Handbook of Metacognition and Learning Technologies. <https://doi.org/10.1007/978-1-4419-5546-3>
- Kalyuga, S., & Singh, A. M. (2016). Rethinking the Boundaries of Cognitive Load Theory in Complex Learning. *Educational Psychology Review*, 28. <https://doi.org/10.1007/s10648-015-9352-0>
- Kitto, K., Shum, S. B., & Gibson, A. (2018). Embracing Imperfection in Learning Analytics. . *Proceedings of the 8th International Conference on Learning Analytics & Knowledge - LAK '18*, 4(10), 451–460. <https://doi.org/10.1145/3170358.3170413>
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 31–40.
- Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, 50(3), 43–59. <https://doi.org/10.1007/BF02505024>
- Milligan, S. K. (2018). Methodological Foundations for the Measurement of Learning in Learning Analytics. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 466-470. <https://doi.org/10.1145/3170358.3170391>
- Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. a. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. *Proceedings of the 24th Conference on Australian Computer-Human Interaction OzCHI '12*. <https://doi.org/10.1145/2414536.2414602>
- Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior*, 60, 333–341. <https://doi.org/10.1016/j.chb.2016.02.074>
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22. <https://doi.org/10.1007/s10648-010-9128-5>
- Slof, B., Erkens, G., Kirschner, P. A., Janssen, J., & Phielix, C. (2010). Fostering complex learning-task performance through scripting student use of computer supported representational tools. *Computers and Education*, 55, 1707-1720. <https://doi.org/10.1016/j.compedu.2010.07.016>
- van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: a four-component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology Publications

The Influence of Geo-Cultural Background on MOOC Learning Trajectories; Mapping Divergence and Similarities

Saman Rizvi

Institute of Educational Technology (IET), The Open University, UK
saman.rizvi@open.ac.uk

ABSTRACT: The possibility of learning in massive, open, and online learning environment leaves an impression of reachability and diversity. However, a large body of MOOC literature have indicated minimal participation from certain cultural clusters, especially from less-privileged strata of the globe. Previous research has identified cultural and regional elements as strong predictors for MOOC engagement, but pedagogical aspects of learning environment have also been found to affect learners' retention. My PhD aims to empirically test and verify on how learners from different geo-cultural background progress in MOOCs, and if their progress is influenced by pedagogical design decisions, such as learning design. Using methods associated with Learning Analytics (LA) and Educational Process Mining (EPM), my PhD will focus on temporal dynamics of learners' end-to-end progression in distinct MOOC learning designs, with a consideration of geo-cultural belonging. The research findings aim to provide useful and actionable insights on MOOC learning designs, leading to potentially more inclusive and diverse MOOCs.

Keywords: Learning Analytics (LA), Educational Process Mining (EPM), Massive Open Online Courses, Learning Design, Cultural Clusters.

1 INTRODUCTION (THE PROBLEM)

As per a formal definition, Massive Open Online Courses (MOOCs) are 'courses designed for large numbers of participants, that can be accessed by anyone anywhere as long as they have an internet connection, are open to everyone without entry qualifications, and offer a full/complete course experience online for free.' (Jansen & Schuwer, 2015). At first, MOOCs were expected to address the global disparity in education. However, some argue that at present, international participation in leading MOOCs continue to present a form of 'intellectual neo-colonialism'. The reason is that majority of active learners and mainstream MOOC providers belong to just a few developed countries (Kizilcec et al., 2017). Like Bozkurt and Aydın (2018) recently noted that 'most of the participation originates from developed, Western, Anglo-Saxon cultures.'

Despite being one of the most innovative and progressive learning phenomena, MOOCs are facing a range of challenges, including a failure to increase overall learners' retention. This problem is particularly severe for international learners. Extensive research into cultural diversity found that cultural background can be one of the primary indicators for educational attainment (Bozkurt & Aydın, 2018). Other studies on online learning (Cai et al., 2017; Kizilcec et al., 2017) found that cultural elements/regional factors influence learning and can be used to potentially predict learners' integration. These findings, however, apply to the context of national or local population, and there is generally a lack of knowledge about a wider regional and geo-cultural effect on MOOC learning. Despite cultural and regional elements being identified as strong predictors for MOOC completion,

pedagogical aspects of learning environment have also been found to affect learners' retention. The issues of success in online learning are closely linked to the Learning Designs. In general, Learning Design (LD) can be described as the process of designing pedagogically informed learning activities to support learners while remaining aligned with the curriculum (Rizvi et al., 2018). In a recent work (Nguyen et al., 2017) found a strong link between LD and successful learning outcomes, suggesting that 'LD could explain up to 60% of the variance of the VLE-engagement time'. However, most of the theoretically-driven work on LD so far, have only been done in formal blended or online education environment and not in MOOCs, thus suggesting a paucity in research. Nonetheless, based on the strong results from previous research in formal online settings, it is expected that Learning Design is likely to play an important role in successful learning in massive, open, online environment as well.

Possibly one of the most controversial debates in both formal and informal education is how to define success. On the one hand, researchers propose that assessment results or earning a certificate, engagement and learning gains are a good reflection of learning. On the other hand, there is an argument that those measures are inadequate because they are only based on outcome variables only (Joksimović et al., 2017) and learning process is a better reflection of actual learning. Likewise Bogarín et al., (2018) suggests that learning is not always evidenced by academic grades. Instead, learning itself is processual. The processual nature of learning can be observed and measured via interaction and engagement with a variety of learning and assessment activities (e.g., video, discussion, quiz, article), and is guided by learners' intentions. Hereby, the term temporal dynamics used in my research has twofold meanings; the engagement-duration, and sequential progression through various activities (Rizvi et al., 2018). Thus, taking the diversity, size and informal or semi-informal way of learning in MOOCs as well as relative flexibility in assessments, it seems more appropriate to look at learning processes in MOOC learning environment.

While previous studies have indicated that regional or cultural constructs significantly impact MOOC learning behavior, to the best of my knowledge no research has looked at the geo-culture and main and dominating temporal learning paths (time-based analyses of processual learning paths) that are followed by large numbers of learners in a course. As such, this PhD project aims to fill this gap in knowledge by probing the temporal dynamics in processual nature of learning in a variety of MOOC learning designs, in context of various geo-cultural clusters. This research highlights the need for an increased intercultural awareness among MOOC instructors, designers and providers. Another major contribution this PhD offers to the existing body of literature is a comprehensive, deeper understanding of the nature of learning processes in different MOOC learning designs. Furthermore, this research will outline theoretical implications and enrich our understanding of learning in MOOC. The findings from this PhD will help to outline practical steps for the practitioners to develop MOOCs that will not only be accessible globally, but which will be globally inclusive as well.

2 THEORETICAL FRAMEWORK(S)

2.1 Geo-Cultural Framework

Few conceptual frameworks have received considerable attention from the researchers investigating the relationship between learning and various dimensions of culture. An important framework is GLOBE (Global Leadership and Organizational Behavior Effectiveness) by House et al. (2004), which used nine cultural dimensions from 62 countries to empirically devise those countries into ten

distinct geo-cultural groups. The Extension of GLOBE societal clusters (Extended-GLOBE) by Mensah & Chen (2013), introduced five more variables which were also perceived to be necessary to define culture in external terms: (1) racial/ethnic distribution; (2) religious distribution; (3) world region or geographic proximity; (4) major language; and (5) (British) colonial heritage. Most importantly they included previously ignored (previous category: Uncategorized) countries. MOOC learners have planetary footprints, where learners reside in all continents and originate from numerous countries. Taking into account the extent of coverage of previous cultural frameworks, using Extended-GLOBE societal clusters will allow a better understanding of geo-culture and its effect on MOOC learning.

2.2 OU Learning Design Initiative (OULDI) Framework

This research has theoretical groundings in the conceptual framework for Learning Design recommended in The Open University Learning Design Initiative (OULDI) project. Same framework provides the foundation FutureLearn MOOCs, which is the primary source of data in this research. The formal taxonomy (Table 1 in Appendix A) was developed by (Conole, 2012) which described LD as reusable, adaptable description or template which aims to ‘make the structures of intended teaching and learning – the pedagogy – more visible and explicit thereby promoting understanding and reflection’. The framework has been empirically tested in large-scale studies (Nguyen et al., 2017; Rienties et al., 2017), although not necessarily in FutureLearn courses. For the interpretations of temporal engagement behaviors, this PhD employs OULDI, along with methods that can map sequence and duration of learning activities (section 4.2).

3 MAIN RESEARCH QUESTIONS

In this PhD, a set of well-connected empirical studies have been designed to answer following RQs:

RQ1: To what extent can main and dominating temporal learning paths be identified in a MOOC Learning Design (i.e., in a MOOC do significantly large subgroups of learners follow a particular learning path before dropping out)?

RQ2.A: To what extent does association with a geo-cultural cluster impacts temporal learning paths in a MOOC Learning Design? (i.e., What temporal learning path learners from a geo-cultural cluster follow, as they progress in a course)?

RQ2.B: To what extent are behavioral patterns (from RQ1 and RQ2.A) of geo-cultural clusters similar or dissimilar in different MOOC Learning Designs?

RQ3: With the help of learners’ reflections, and temporal process models from RQ1 and RQ2, how can we suggest meaningful, actionable insights from investigating the broader geo-cultural and pedagogical factors that may make MOOC learning more sustainable, diverse and inclusive?

4 PROPOSED METHODOLOGY

4.1 Data Sources and Ethical Considerations

Building on my initial work set in analyzing five OU modules using decision-tree modelling (Rizvi et al., 2018), and follow-up work analyzing learning design decisions in one Futurelearn MOOC (Rizvi et

al., 2018), my work in progress is using data from four FutureLearn MOOCs from year 2017. The MOOCs followed different learning designs and had a (relatively) large international (Non-Anglo) learners' population. The courses will be offered at least few more times so would provide an opportunity to get perception or experience data from learners during the last (qualitative) phase of this PhD research.

Table 1. Description of four Open University FutureLearn MOOCs (Work in progress)

MOOC	Discipline	Learners
MOOC 1	Nature & Environment and Science, Eng. & Math	2086
MOOC 2	Tech & Coding and Business & Management	981
MOOC 3	Business & Management	1927
MOOC 4	Languages & Cultures and Study Skills	11763

The research protocol for my research project has recently been assessed, and fully approved by the Open University Human Research Ethics Committee (OU-HREC).

4.2 Data Analysis Methods

The enormity of volume of data constantly being generated by the systems requires advanced analysis methods which are scalable, comprehensible, and yet easy to implement by non-technical stakeholders. Therefore, to develop learners' temporal navigational patterns, I am using Educational Process Mining (EPM). EPM is the application of Process Mining techniques in educational domain (Bogarín et al., 2018). In Process Mining, the term *Variant* refers to an end-to-end sequence of activities followed by a significant number of cases. In our case, each of these subgroups (Variants) demonstrates a learning process, and all variants follow a learning trajectory (Fig.1 Appendix A). This PhD is focused on understanding such end-to-end learning processes in various learning designs, and to see if this progression is affected by geo-culture. For process map construction, *Discovery* software will be used, which implements *Fuzzy Miner* algorithm (Günther & Van Der Aalst, 2007) to create elaborative, uncomplicated process map or identify infrequent variants (learner subgroup).

5 CURRENT STATUS OF WORK AND FUTURE WORK

My project will be conducted through a series of studies addressing respective RQs (Figure.1).

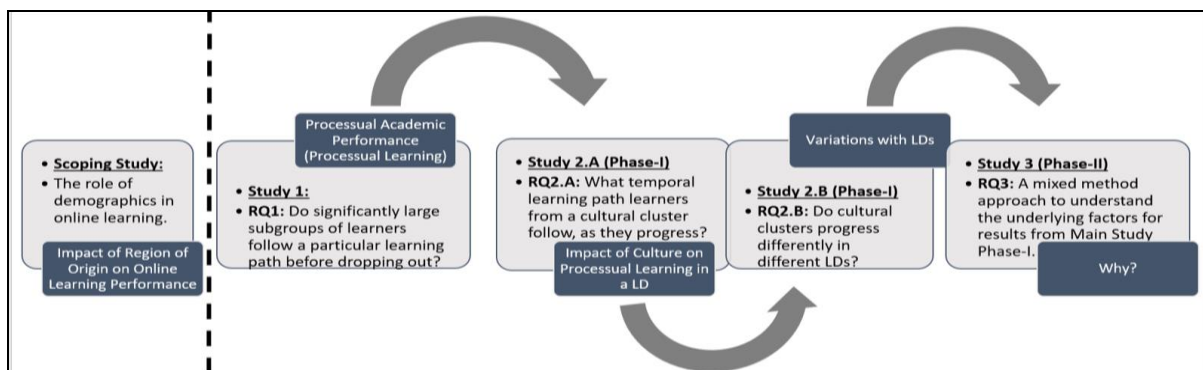


Figure 1. Illustration of the PhD research plan

I have recently discussed this project at Young Researchers Track, in 19th AIED18 conference (Rizvi et al., 2018) and received really useful feedback from the academics and senior researchers.

5.1 Scoping Study

This scoping study helped me to identify what is known about various demographic attributes in relation to learners' dynamic progress in online courses by examining whether, and how, six demographics (i.e., region of origin, multiple deprivation levels, education, age, gender, and disability) affected outcomes of assignment over time. The study employed machine learning based predictive models on data from 3,908 UK-based learners enrolled in four Open University courses. The results suggested that (a) region of origin remained highly predictive of the temporal performances throughout the course, and (b) a change in Learning Design can potentially influence temporal academic performance. Overall results comprised a journal paper, have been submitted to a prestigious journal (Rizvi et al., 2018). The findings provided bases for my proposed PhD studies.

5.2 Study 1 (Work-in-progress)

This study addresses RQ1. To confirm the consistency, the analyses were repeated on different LDs from different disciplines. Some preliminary results support the propositions that (a) MOOC learning is processual and learning trajectories can be mapped and compared with the LDs (the pathways a learner is expected to follow), (b) engagement-duration is a key temporal aspect which should not be left unexplored or unmapped, (c) learners exhibit varied clicking behaviors which are suggestive of natural groupings, as well as distinct psychological dispositions or intentions. Some preliminary results were presented at JURE18 (EARLI) conference, and at ACM Data'18 conference (Rizvi et al., 2018). Finally, more finer results have recently been submitted to LAK19 as a short research paper.

5.3 Study 2 (Future Work)

Phase-I: Study 2.A, Study 2.B: Study2 Phase-I (quantitative), will answer RQ2.A, RQ2.B by exploring how academic performance (which is processual and is linked closely with LD) varies with geo-cultural background. The additional variable of IP based-location will be extracted and converted to respective geo-cultural cluster. Analysis methods will be like as were used in Study 1.

5.4 Study 3 (Future Work)

Phase-II: Study 3: This study will close the gap between extracted knowledge and its potential use. I will explore the underlying factors behind findings from Study 1 and 2 by using the qualitative methods (post-course survey). I will examine if results from Study 2 can be explained in the light of learners' self-reported learning experiences with various type of learning activities (such as videos, forum, assessments). Also, if temporal learning paths were a result of intentional navigation? As well as dependent upon, and guided by, the learning designs. I will try to understand (i) to what extent the deviations in pathways were intentional? and in relation to that (ii) which pedagogical/design strategies would be useful for instructors/developers? This, in turn, will explain why culture does (or doesn't) have an impact on processual nature of learning? Since this part of research will be executed sometime later in my 2nd year, I remain open to other approaches and methods.

6 PRACTICAL IMPLICATIONS AND UNIQUE CONTRIBUTION

While recent evidence indicates impact of course designs on learners' engagement, to the best of my knowledge no study has linked learning design with the geo-cultural diversity in MOOCs at such

scale. By linking learning design principles with activity-engagement from all around the globe, this PhD project aims to understand better how one can design and implement effective MOOCs for ALL learners. This could potentially lead to the workload estimation especially when majority of MOOC learners are adults, with several personal responsibilities (family, other extracurricular activities) and working full-time/part-time. It is even more critical to control overall difficulty level and prevent overestimation (or underestimation) of workload expected from diverse learners. The proposal can be expended to a personalized learning system that considers cultural preferences (like any other good recommender system). The ethical considerations of such implication, however, remain a topic of interest to future researchers and to me. The deeper understanding of a learning environment, which has all the inherent potentials to support global diversity in Education, will result in useful, actionable insights the stakeholders, researchers and practitioners can use to design all-inclusive MOOCs. I believe that this could improve overall engagement in MOOCs as a consequence.

7 ACKNOWLEDGMENT

This work is supported and funded by the Leverhulme Trust, Open World Learning.

8 REFERENCES

- Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1).
- Bozkurt, A., & Aydın, İ. E. (2018). Cultural Diversity and Its Implications in Online Networked Learning Spaces. In *Supporting Multiculturalism in Open and Distance Learning Spaces* (pp. 56–81). IGI Global.
- Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education*, 105, 1–13.
- Conole, G. (2012). *Designing for learning in an open world* (Vol. 4). Springer Science & Business Media.
- Günther, C. W., & Van Der Aalst, W. M. (2007). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *International Conference on Business Process Management* (pp. 328–343). Springer.
- House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.
- Jansen, D., & Schuur, R. (2015). *Institutional MOOC strategies in Europe status report based on a mapping survey conducted in October - December 2014*. EADTU.
- Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., ... Brooks, C. (2017). How Do We Model Learning at Scale? A Systematic Review of Research on MOOCs. *Review of Educational Research*, 0034654317740335.
- Kizilcec, R. F., Davis, G. M., & Cohen, G. L. (2017). Towards Equal Opportunities in MOOCs: Affirmation Reduces Gender & Social-Class Achievement Gaps in China. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 121–130). ACM.
- Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355(6322), 251–252.
- Mensah, Y., & Chen, H.-Y. (2013). Global clustering of countries by culture—an extension of the GLOBE study.
- Nguyen, Q. (2017). Unravelling the dynamics of learning design within and between disciplines in higher education using learning analytics.
- Nguyen, Q., Rienties, B., & Toetenel, L. (2017). Mixing and matching learning design and learning analytics. In *International Conference on Learning and Collaboration Technologies* (pp. 302–316). Springer.
- Rienties, B., Nguyen, Q., Holmes, W., & Reedy, K. (2017). A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK. *Interaction Design and Architecture (S)*, 33, 134–154.
- Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior*, 60, 333–341.
- Rizvi, S., Rienties, B., & Khoja, S. (2018). The Role of Demographics in Online Learning; A Decision Tree Based Approach. *Manuscript Submitted for Publication*.

- Rizvi, S., Rienties, B., & Rogaten, J. (2018). Investigation of Temporal Dynamics in MOOC Learning Trajectories: A Geocultural Perspective. In *International Conference on Artificial Intelligence in Education* (pp. 526–530). Springer.
- Rizvi, S., Rienties, B., & Rogaten, J. (2018). Temporal Dynamics of MOOC Learning Trajectories. In *Proceedings of the International Conference on Data Science, E-learning and Information Systems. DATA'18*. ACM. <https://doi.org/10.1145/3279996.3280035>

The Effects of Discussion Strategies and Learner Interactions on Performance in Online Mathematics Courses: An Application of Learning Analytics

Ji Eun Lee

Department of Instructional Technology & Learning Sciences, Utah State University

jieun.lee@aggiemail.usu.edu

ABSTRACT: In higher education, a widely used online instructional method to enhance learners' engagement, presence, and achievement is asynchronous online discussions. Yet studies demonstrating their effectiveness, especially in high-failure rate courses like mathematics, remain elusive. The objectives of the study are to investigate 1) what online discussion strategies are associated with positive student performance, 2) to what extent do different structures designed into online discussions impact the kinds of learner interactions, and 3) what types of learner interactions are associated with positive student performance. In particular, by applying a set of text mining and data mining techniques (e.g., Classification and Regression Tree), this study analyzes clickstream and textual data automatically collected by a Learning Management System (LMS) for five consecutive years at a university located in the western U.S. The results of study will inform instructors and instructional designers how to design the better online mathematics courses.

Keywords: Asynchronous online discussion, Online mathematics courses, Classification and Regression Tree (CART), Automated content analysis

1 BACKGROUND

Mathematical skill is one of the core competencies for the 21st century (Dede, 2010). It is not only a foundation for all Science, Technology, Engineering, and Math (STEM) disciplines but also helps learners solve complex problems and make important connections to other fields (Chen, 2013). A recent study found that mathematical ability also influences career success and accomplishments (Lubinski, Benbow, & Kell, 2014).

1.1 Problem Statement: Challenges in College Mathematics

"Mathematics courses are the most significant barrier to degree completion" (Saxe & Braddy, 2015, p.28).

Despite the importance of math skills, high failure rates in college math courses have become a growing concern in the United States. One report found that approximately 50% of students do not pass college algebra courses with a grade of C or above (Saxe & Braddy, 2015). The negative experiences in math courses also affect degree completion. The result of a nation-wide study indicated that negative experiences in math courses, such as poor performance or withdrawal, were associated with not just leaving STEM majors, but also led to a higher probability of dropping out of college (Chen, 2013). More seriously, while the number of students taking online courses is rapidly increasing, online math courses showed even worse results, with a 20% higher failure/withdrawal rates (62%) compared to face-to-face courses (43%) (Jaggars, Edgecombe, & Stacey, 2013).

1.2 Possible Solution

In online learning environments, one of the widely used instructional methods to enhance learners' engagement, presence and achievement is asynchronous online discussions, a type of Computer-Supported Collaborative Learning (CSCL) (Hew, Cheung, & Ng, 2010; Ke & Xie, 2009). Many previous studies have shown that using asynchronous online discussions had significant effects on increasing students' achievement (Pettijohn II & Pettijohn, 2007), critical thinking skills (Maurino, 2007), and engagement (Salter & Conneely, 2015). In mathematics education, it is also important to involve activities that develop mathematical thinking and communication skills to increase students' mathematical understanding and success. A number of studies have also demonstrated that the use of online discussions has helped in decreasing math anxiety (Liu, 2008), the creation of correct and new ideas (Chen et al., 2012), and achievement outcomes (Tunstall & Bossé, 2015).

However, the use of online discussions does not always lead to productive interactions or knowledge construction. Many studies reported that student often exhibited low participation rates, low levels of critical thinking or knowledge construction (Hew et al., 2010; Maurino, 2007). Indeed, several empirical studies have revealed that learners exhibited a higher level of engagement or performed better in effectively designed and structured online discussions (Darabi, Liang, Suryavanshi, & Yurekli, 2013; Salter & Conneely, 2015). Thus, it is important to offer well-designed and domain-specific support to engage learners in meaningful activities and discourse.

Nonetheless, instructors seldom implement strategic online discussions that are purposefully designed or structured (Darabi et al., 2013). In addition, in terms of research, several gaps were identified. First, although there have been numerous studies in CSCL field, most of the studies tended to focus on students' behaviors or interactions, rather than instructor involvement (Maurino, 2007). Little research has investigated effective design strategies, such as the design of activities or discussion tasks, that reinforce meaningful student interactions (Ke & Xie, 2009). Second, although the implementation of online discussions has been less successful in mathematics learning contexts compared to other academic disciplines (Nason & Woodruff, 2004), the effective use of online asynchronous discussions has seldom been studied in mathematics learning contexts (Maurino, 2007).

1.3 Objectives and Research Questions

To address these challenges in research and practice, the aim of this study is twofold. The first is to explore the effective discussion strategies that enhance meaningful learner interactions and achievement outcomes in online introductory mathematics courses. The second is to investigate learner behaviors and interaction patterns that lead to better learning outcomes. In particular, by using a learning analytics approach, this study analyses large-scale data automatically collected by a Learning Management System (LMS) for five consecutive years at a university located in the western U.S.

To examine the relationship between instructors' use of discussion strategies, learners' interactions and learning outcomes, a research model was created based on Biggs's 3P model (Biggs, 1991) (Figure 1). The 3P model assumes that the four factors, student characteristics, teaching context, students' approaches to learning, and learning outcomes are interrelated and affect each other.

Among the four factors, this proposed study focuses on the relationship between teaching context (instructors' use of discussion strategies), students' approaches to learning (learner interactions) and learning outcomes (performance).

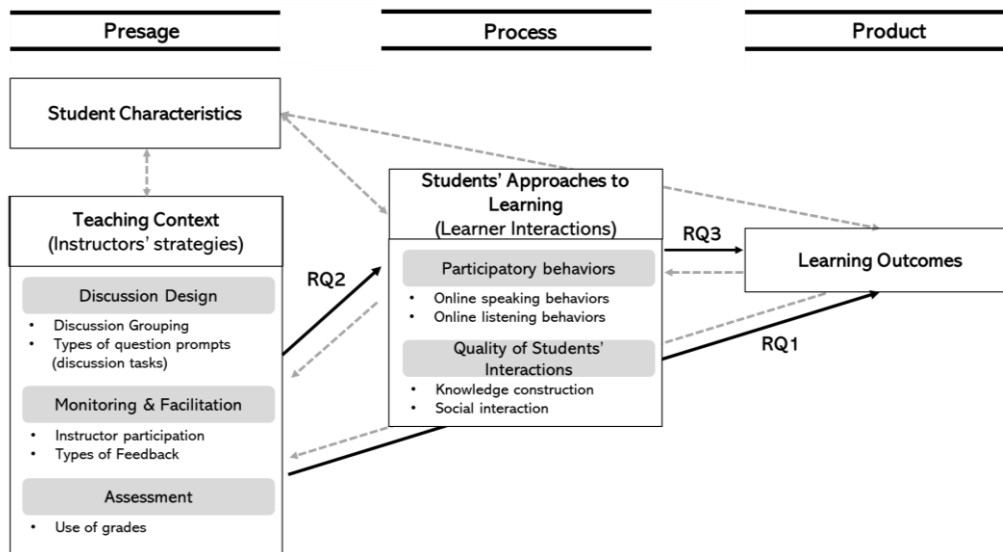


Figure 1: The research model adopted from Biggs' Presage-Process-Product (3P) model

The specific research questions are as follow: For online introductory mathematics courses:

1. What online discussion strategies are associated with positive student performance?
2. To what extent do different structures designed into online discussions impact the kinds of learner interactions?
3. What types of learner interactions are associated with positive student performance?

2 METHODOLOGY

The proposed study uses a data-driven approach by applying *learning analytics* techniques. In recent years, an increased interest in learning analytics has emerged due to the rapid growth of online education. One of my previous studies (Lee & Recker, 2018) reviewed 47 studies that used learning analytics methods. The results of the systematic review showed that most studies focused on learner behaviors, while remarkably few studies looked at instructor or course related data, which is similar to a trend in CSCL research (Maurion, 2007). In addition, the vast majority of the work has used quantitative data capturing learner interactions, such as simple counts of user activities, whereas few studies have sought to examine textual or content data.

2.1 Research Context and Sample

This study will use data automatically collected by a Learning Management System (LMS), Canvas, used at a public university located in the Western U.S. The Canvas system records a log of all of students' and instructors' interactions, with dates and timestamps, as well as student/instructor textual data, such as discussion prompts, messages, and replies. These Canvas data are made available to an academic-support (AS) unit at the university, which then anonymizes the data to

protect user privacy. The AS unit then makes the data available as multiple files for further analysis. The sample for the study includes instructors and students in fully-online introductory (0 and 1000 levels) mathematics/statistics courses offered between 2011 fall and 2015 summer semesters. The sample consists of four levels of hierarchy, course, students, activities, and events/actions. Figure 2 summarizes the number of courses, students, discussion topics, and discussion messages posted by the instructors and the students.

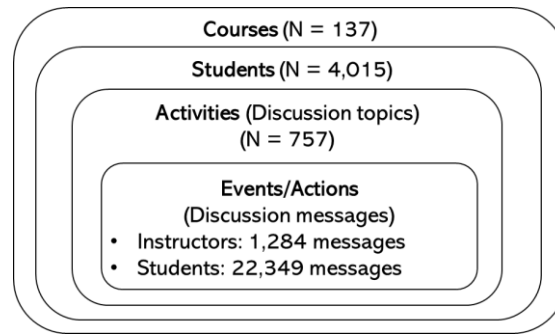


Figure 2: Summary of sample sizes with the different levels of hierarchy

2.2 Research Design and Procedures (Current status of work)

This study uses a quantitative and non-experimental research design. The study is guided by the Knowledge Discovery in Databases (KDD) process, which is a widely used process frameworks in data mining, learning analytics, and educational data mining research (Baker & Yacef, 2009). Figure 3 summarizes the research procedures. Up until now, data pre-processing and hand-coding a subset of the textual data are completed in order to train the machine learning algorithms.

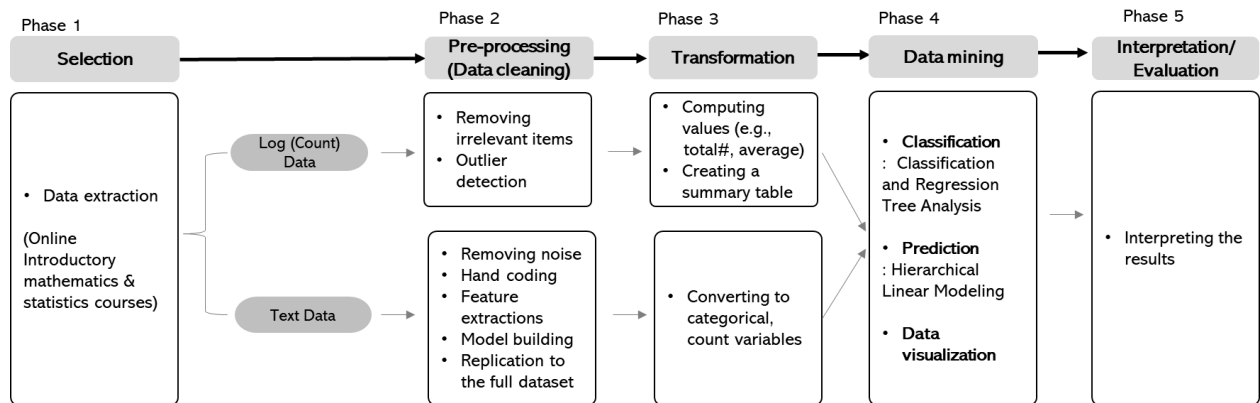


Figure 3: Research Procedures guided by the KDD process

2.3 Measurement and Data Analysis

For Research Question 1 (course-level analysis) and Research Question 2 (course-level analysis), instructors' use of online discussion design strategies is measured in terms of three constructs: discussion design, discussion monitoring and facilitation, and discussion assessment. The constructs, sub-constructs, categories, and how each variable is measured are summarized in the Appendix.

2.3.1 Innovation: Automatic Analysis of Online Discussion Data

To measure “types of discussion tasks,” “types of feedback,” and “qualitative aspects of learner interactions” (See Table in the Appendix), this study applies automated analyses of online discussions using a text mining tool *LightSIDE* (Mayfield, Adamson, & Rosé, 2013). There are several advantages of using automatic content analysis (Mu, Stegmann, Mayfield, Rosé, & Fischer, 2012). First, it helps reduce the time required for analyzing the huge body of online discussions by hand as well as training human coders, thus accelerating the progress of research. Also, it enables researchers to analyze discussions messages along multiple dimensions at the same time. Further, it can inform the design of adaptive collaborative learning support, such as individualized feedback or scaffolds, to enhance the quality of learners’ knowledge constructions during online discussions.

Finally, Table 1 summarizes the input variables, outcome variables, analysis methods and tools used in the study.

Table 1: Summary of variables, analysis methods, and tools used in the study

	Input variables	Outcome variables	Analysis methods	Tools
Data pre-processing			-Data cleaning -Content analysis (Text mining)	SQL server management studio, LightSIDE
RQ1. What online discussion strategies are associated with positive student performance?	Instructors’ use of discussion strategies	Average of students’ final grades in each course (out of 4.00)	Decision Tree: Classification and Regression Tree (CART)	R studio (rpart package)
RQ2. To what extent do different structures designed into online discussions impact the kinds of learner interactions?	Instructors’ use of discussion strategies	Different Level of learners’ interactions	-Kruskal-Wallis H Test -Descriptive statistics	R studio
RQ3. What types of learner interactions are associated with positive student performance?	Level of learners’ interactions	Students’ final grades (out of 4.00)	Hierarchical Linear Modeling (HLM)	R studio

REFERENCES

- Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–16.
- Biggs, J. (1991). Approaches to learning in secondary and tertiary students in Hong Kong: Some comparative studies. *Educational Research Journal*, 6, 27–39.
- Chen, G., Chiu, M. M., & Wang, Z. (2012). Social metacognition and the creation of correct, new ideas: A statistical discourse analysis of online mathematics discussions. *Computers in Human Behavior*, 28(3), 868–880.
- Chen, X. (2013). *STEM Attrition: College students’ paths into and out of STEM fields* (NCES 2014-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

- Darabi, A., Liang, X., Suryavanshi, R., & Yurekli, H. (2013). Effectiveness of online discussion strategies: A meta-analysis. *American Journal of Distance Education*, 27(4), 228–241.
- Dede, C. (2010). Comparing frameworks for 21st century skills. In J. Bellanca & R. Brandt (Eds.), *21st Century Skills: Rethinking how students learn* (pp. 51–76). Bloomington, IN: Solution Tree Press.
- Hew, K. F., Cheung, W. S., & Ng, C. S. L. (2010). Student contribution in asynchronous online discussion: a review of the research and empirical exploration. *Instructional Science*, 38(6), 571–606.
- Jaggars, S. S., Edgecombe, N., & Stacey, G. W. (2013). *What we know about online course outcomes. Research overview*. Community College Research Center, Columbia University. Retrieved from <http://eric.ed.gov/?id=ED542143>
- Ke, F., & Xie, K. (2009). Toward deep learning for adult students in online courses. *Internet and Higher Education*, 12(3–4), 136–145.
- Lee, J. E., & Recker, M. (2018). What do studies of learning analytics reveal about learning and instruction? A systematic literature review. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, Design, and Technology* (pp. 1–37). Cham, Switzerland: Springer International Publishing.
- Liu, F. (2008). Impact of online discussion on elementary teacher candidates' anxiety towards teaching mathematics. *Education*, 128(4), 614–630.
- Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life paths and accomplishments of mathematically precocious males and females four decades later. *Psychological Science*, 25(12), 2217–2232.
- Maurino, P. S. M. (2007). Looking for critical thinking in online threaded discussions. *Journal of Educational Technology Systems*, 35(3), 241–260.
- Mayfield, E., Adamson, D., & Rosé, C. (2013). *LightSIDE: Researcher's user manual*. Retrieved from http://www.cs.cmu.edu/~cprose/LightSIDE_Researchers_Manual_Draft3.pdf
- Mu, J., Stegmann, K., Mayfield, E., Rosé, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, 7(2), 285–305.
- Nason, R., & Woodruff, E. (2004). Online collaborative learning in mathematics: Some necessary innovations. In T. S. Roberts (Ed.), *Online collaborative learning: Theory and practice* (pp. 103–131). Hershey, PA: Information Science Publishing.
- Pettijohn II, T. F., & Pettijohn, T. F. (2007). Required discussion web pages in psychology courses and student outcomes. *Journal of Instructional Psychology*, 34(4), 256–263.
- Salter, N. P., & Conneely, M. R. (2015). Structured and unstructured discussion forums as tools for student engagement. *Computers in Human Behavior*, 46, 18–25.
- Saxe, K., & Braddy, L. (2015). *A common vision for undergraduate mathematical sciences program in 2025*. Retrieved from <https://www.maa.org/sites/default/files/pdf/CommonVisionFinal.pdf>
- Tunstall, S. L., & Bossé, M. J. (2015). Promoting numeracy in an online college algebra course through projects and discussions. *Numeracy*, 8(2), 1–23.
- Van Der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4), 475–511.
- Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y. T. (2013). Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors. *Instructional Science*, 41(2), 323–343.

APPENDIX

Constructs		Categories	Measures	Types of variables	Data sources
Instructors' use of discussion strategies (Course-level analysis)					
Discussion design	Grouping	Whole-class	Whether the students in each course are assigned to a group or not	Categorical	Log data
		Small group			
		Mixed			
	Types of discussion tasks (Ke & Xie, 2009)	Open-ended	Total # of open-ended discussion tasks	Continuous	Textual data
		Closed-ended	Total # of closed-ended discussion tasks	Continuous	
		Others	Total # of discussion tasks fall into neither open-ended nor closed-ended	Continuous	
Monitoring and Facilitation	Monitoring	Instructor participation	Total # of discussion views by an instructor	Continuous	Log data
			Total # of discussion posts by an instructor	Continuous	
	Types of Feedback (Kleij, Feskens, & Eggen, 2015)	Elaborated feedback	Total # of elaborated feedback provided by an instructor (e.g., providing an explanation)	Continuous	Textual data
		Correctness of the answer	Total # of feedback regarding the correctness of the answer by an instructor	Continuous	
		Providing the correct answer	Total # of feedback providing the correct answer by an instructor (e.g., Revising the student's incorrect responses by providing the correct answer)	Continuous	
	Assessment	Use of grades	Graded	Whether the discussion messages are graded or not	Categorical
Not-graded					
Mixed					
Learner interactions (Student-level analysis)					
Participatory behaviors (Wise et al., 2013; 2014)	Online speaking	Quantity	Total number of new messages made by a student	Continuous	Log data
			Average message length (in words)	Continuous	
		Breadth	Percent of threads with a minimum of one message posted	Continuous	
	Online listening (attending to other's posts)	Quantity	Total number of replies made by a student	Continuous	
			Total number of views of (any) discussion threads by a student	Continuous	
		Breadth	Percent of threads read at least once	Continuous	
Qualitative aspects of Interactions	Social Interactions		Total # of messages regarding social interactions (e.g., emotional expressions)	Continuous	Textual data

Constructs		Categories		Measures	Types of variables	Data sources
(Ke & Xie, 2009)	Knowledge constructions	Sharing Information (K1)	Total number of messages regarding sharing information (e.g., simply adding facts)		Continuous	
		Egocentric elaboration (K2)	Total number of messages elaborating one's own arguments		Continuous	
		Allocentric elaboration (K3)	Total number of messages comparing or synthesizing peers' multiple perspectives		Continuous	
		Application (K4)	Total number of messages regarding the application of new knowledge		Continuous	
Outcome variables						
Performance	RQ1: Average of students' final grades in each course (out of 4.00)				Continuous	Log data
Learners' interactions	RQ2: Measures of descriptive statistics of learner interactions				Continuous	
Performance	RQ3: Students' final grades (out of 4.00)				Continuous	

The Use of Learning Analytics in a Blended Learning Context

Author: Elise Ameloot

Ghent University

Elise.Ameloot@UGent.be

[DOCTORAL CONSORTIUM] ABSTRACT: Blended Learning (BL) has many opportunities for flexible learning, but it also poses some challenges. One of these challenges is to keep students motivated. As described by self-determination theory, a prerequisite for motivation are students' basic needs for autonomy, relatedness and competence. An opportunity that is often overlooked by educational scientists is the use of Learning Analytics (LA) to promote students' motivation. Therefore, the general goal of this research project is to examine if and how LA can support students' motivation in an authentic BL context in teacher education. This research goal is investigated through a mixed-method design-based research approach. Preliminary results confirm that students' initial motivation is low. Further results will be discussed, as well as implications for practice and research.

Keywords: Learning Analytics; Self-Determination Theory; Blended Learning; Teacher Education

1 CONTEXT

In teacher education the increasing diversity of students' characteristics is a worldwide phenomenon (Preston et al., 2010). Additionally, technology has become an essential part of society and offers many opportunities for education (Brand-Gruwel, 2012; Rubens, 2013). Blended Learning (BL) is characterized by a deliberate combination of online and face-to-face interventions to investigate and support learning in an instructional context (Boelens, Van Laer, De Wever, & Elen, 2015). It is important to implement BL in teacher education since it enables more flexible education responding to the earlier mentioned diversity in teacher education (Irvine, Code, & Richards, 2013; Laurillard, 2014). Besides, student teachers are the new generation of teachers who can disperse the opportunities of BL and technology enhanced learning in general (Cabero Almenara, del Carmen Llorente Cejudo, & Puentes Puente, 2010; Delfino & Persico, 2007). Consequently, BL is one of the twelve projects in the strategic plan of Ghent university (Belgium).

As stated in the BL review study of Boelens, De Wever and Voet (2017), fostering an affective learning climate is one of the key challenges in designing BL environments. A pitfall is the decrease of students' motivation during the learning process (Laurillard, 2014; Osguthorpe & Graham, 2003), especially when students' basic needs are not fulfilled (Rubens, 2013). A pilot study of Ameloot & Schellens (2018) reaffirms this and indicates that students' basic psychological needs for autonomy, relatedness and competence were not fulfilled in a BL environment. Follow-up research should investigate how to stimulate the motivational component in BL environments using LA (Ameloot & Schellens, 2018). So far, there has been a general lack of research about the added value of using Learning Analytics (LA) (Tempelaar, Rienties, & Giesbers, 2015) to promote students' basic needs.

2 THEORETICAL FRAMEWORK

2.1 Students' Basic Needs as Components of Motivation

Self-determination theory is a broad and strongly validated framework to gain insight into students' motivation (Ryan & Deci, 2000). Within this framework, motivation is classified as autonomous and controlled motivation. It is important to foster students' autonomous motivation, because this is associated with a high degree of self-determination (Vansteenkiste & Soenens, 2015) and various positive learning outcomes (Reeve, Deci, & Ryan, 2004). It is maintained that students' autonomous motivation can be increased when a learning environment facilitates the satisfaction of the basic needs for autonomy, relatedness and competence. As stated in figure 1 the basic needs can be promoted respectively by offering autonomy support, involvement and structure. These actions are consolidated under the heading of basic need supportive teaching (Vansteenkiste & Soenens, 2015).

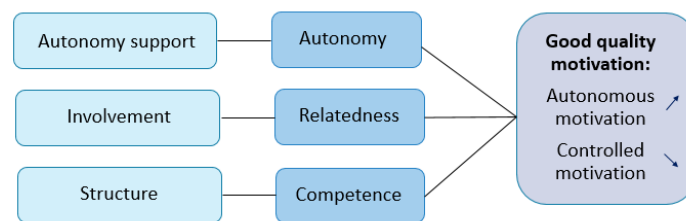


Figure 1: Basic need supportive teaching (based on Vansteenkiste & Soenens, 2015)

Research shows that there is little relatedness and autonomy in BL environments (Rubens, 2013). However, the students must be motivated to complete e.g. an entire learning path on their own. Thus, it is important to foster students' motivation in BL environments (Rubens, 2013).

2.2 Learning Analytics

Long and Siemens (2011) refer to the first LAK conference to define this concept: "LA is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." (p. 34) It covers a wide range of analytics and occurs on different levels of education. The focus of this study is on the micro-level, which mainly addresses the needs of instructors and students, and aims at a single course (Drachsler & Kalz, 2016; Shum, 2012). Clow (2012) describes LA as a cyclic process. This process starts with the student. The interaction between the students and the digital learning environment provides data, as for example the duration to complete a learning task, background information of students or information related to students learning activities on a Learning Management System (LMS). The data can be analysed and visualized. Through this LA, instructors might understand students' needs better, so they could provide optimal feedback and make informed instructional decisions. Based on this information, the instructor can organize an intervention (Clow, 2012). Hence, LA can impact on both students' learning and instructors' design and management of the learning environment (Mangaroska & Giannakos, 2018; Matuk, Linn, & Eylon, 2015). It is presumed that LA applied within learning design provide opportunities for more personalized learning experiences and can increase students' satisfaction (Mangaroska & Giannakos, 2018; Schmitz, Limbeek, Greller, Sloep, & Drachsler, 2017).

Furthermore, establishing and investigating the connection between what instructors do with the data and how this data is relinked to students, is essential for a good implementation (Clow, 2012). In a BL context, LA might help to narrow the gap between online and face-to-face interventions by offering the instructor insight into students' online activities (Ginns & Ellis, 2007; Tempelaar, Rienties, Mittelmeier, & Nguyen, 2018). However, the link between LA and motivation is underdeveloped (Tempelaar, Rienties, & Giesbers, 2015). Current research suggests that the use of LA needs to be investigated (Pardo, Powuet, Martinez-Maldonado, & Dawson, 2017) in order to enhance students' motivation (Rubens, 2013). In addition, little is known about which data is perceived as useful by instructors and the perceptions of students and instructors of LA (Wise & Shaffer, 2015). Research on LA design decisions, such as which data can be collected and how, is needed (Jivet, Specht, Scheffel, & Drachsler, 2018; Mangaroska & Giannakos, 2018; Verbert et al., 2014).

3 PURPOSE OF THE STUDY

LA data allows instructors to make informed instructional decisions to provide a more personalized learning environment (Clow, 2012; Laurillard, 2014; Matuk et al., 2015). Our hypothesis states that through the use of LA data instructors can enhance students' basic need satisfaction through adequate adaptations of the learning environment and offering appropriate formative feedback. Therefore, the purpose of this dissertation is to examine how LA can be used to enhance students' basic needs in an authentic BL context. Furthermore, another aim involves gaining insight into students' and instructors' general perceptions of LA. In the LAK research field, evidence-based studies in an authentic learning environment are rare. This doctoral dissertation will also help to address this research gap.

4 METHODOLOGY

A mixed method approach is used in this research project gathering both quantitative and qualitative data. A pre-post-test design is set up to collect quantitative data, through the use of both existing questionnaires such as the basic need satisfaction scale (Chen et al., 2015) and newly developed questionnaires about students' perceptions of using LA. Finally, focus groups or interviews are organized to gather more qualitative data (Howitt, 2011). The following sections focus on the research questions and the design-based research approach that is carried out in this research project.

4.1 Research Questions

The central focus of this research project is how LA can support students' basic need satisfaction in a BL context in teacher education. More precisely, the following research questions are formulated:

- RQ 1: How do students teachers' experience learning in a BL context?
- RQ 2.1: What are student teachers' perceptions of using LA from their experiences as student in the BL context?
- RQ 2.2: What are student teachers' perceptions of using LA from their future teacher perspective?
- RQ 3: What are instructors' perceptions of using LA?
- RQ 5: What is the impact of using LA on students' basic psychological need for autonomy, relatedness and competence?
- RQ 6: What are important design requirements and design propositions for using LA in a BL context?

4.2 Design-Based Research Approach

To answer the research questions, a design based research approach is used. This is a systematic yet flexible methodology using design, implementation and iterative analysis (e.g. intervention studies), with the goal to develop design principles and theories to improve educational practices (McKenney & Reeves, 2012). The interventions are designed and implemented in an authentic setting: in this case the course Powerful Learning Environments in teacher education of Ghent University ($N = \pm 300$).

5 CURRENT STATUS OF THE WORK AND RESULTS

In 2016 and the beginning of 2017 a review of the literature was conducted. Furthermore, in a pilot study (December 2016, $N = 164$ students) a questionnaire was used to gather preliminary results about student teachers' initial basic needs satisfaction in a BL environment; items were rated on a 5-point Likert scale, ranging from 1 *completely disagree* to 5 *completely agree* (Chen et al., 2015). Based on a one-sample t-test, the basic needs for autonomy ($M: 2,534$; $SD: 0,864$) and relatedness ($M: 2,707$; $SD: 0,780$) score significantly lower ($p < 0,001$) than the neutral score of 3 (on a 5-point Likert scale). Based on these findings, it can be concluded that students' initial motivation is low. This reaffirms the challenge to investigate how this motivational component may be enhanced (Rubens, 2013), by using LA. In the earlier mentioned pilot study, student teachers' perceptions of using LA were gathered as well. Descriptive results indicate that these perceptions are rather diverse. The findings show that student teachers are not yet convinced about the added value of LA. Yet, it has to be noted that these perceptions are based on a hypothetical use of LA and not on student teachers' experiences, because there was no LA intervention conducted (Ameloot & Schellens, 2018).

Following up, in 2017 the first quasi-experimental intervention study was being conducted ($N = \pm 261$ students). This intervention focusses on the proximal effect of how LA can support the basic need for relatedness. The quasi-experimental study was conducted during 3 months in module five of the course Powerful Learning Environments. In this course the module starts with an online learning path in which central concepts are presented. The second part of the module consists of a face-to-face workshop which focuses on a specific technological tool. Every workshop was organized twice both in the control and experimental condition. Participants were distributed over the control ($n = 139$) and experimental ($n = 118$) condition. In the control condition, a regular implementation of the online learning path took place and the instructor organized a general workshop. In contrast, in the experimental condition different LA were gathered and offered to the instructor, enabling the instructor to make adequate adaptations in the face-to-face workshop and offer appropriate formative feedback (see Figure 1). The design of the intervention study is illustrated in figure 3.

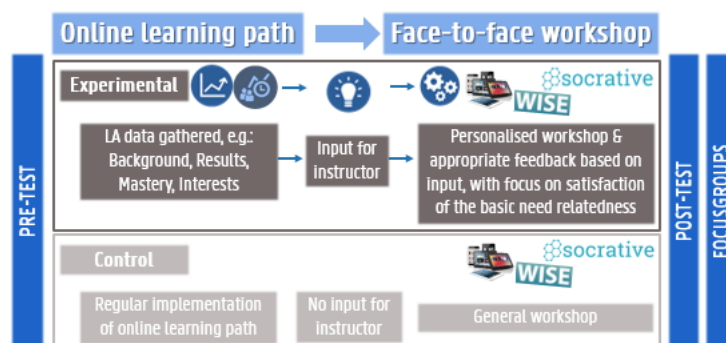


Figure 2: Design of the intervention study

The LA were gathered on group level per group of students who followed the same face-to-face workshop at the same moment. This data was displayed on the one hand on a LA dashboard and on the other hand in additional graphs because the dashboard did not generate all the data automatically. The design is guided by the framework of Self-Determination Theory (Ryan & Deci, 2000) and the LA literature (Clow, 2012; Long & Siemens, 2011). Three types of data were gathered:

1. The first type of data were *general statistics* gathered by the LMS about for example students' assignments and task on time.
2. The second type of data was gathered through extra questions about *students' degree of understanding, interests, and desires* for the face-to-face workshop. Next to multiple choice questions, also some open questions were asked to gather more enriched data. This first and second type of data is important to monitor students' learning, enabling the instructor to improve the instruction and provide appropriate formative feedback during the face-to-face workshop. By asking and using this information, students could feel more related.
3. The third type of data was information about *students' previous education and background characteristics*, gathered through extra questions and displayed in a graph. This type of data is interesting because of the highly diverse student population in teacher education (Delfino & Persico, 2007). This information should enable the instructor to personalise the face-to-face workshop, which is expected to enhance students' basic need for relatedness.

Based on a one-way repeated measures ANOVA, the results revealed no significant effects of the intervention on both relatedness within students and relatedness between students and instructor. Nevertheless, the results indicate a significant main effect for time on relatedness between students and instructor (Wilks' Lambda = .941, $F(1, 255) = 15.856$, $p < .001$). All students felt more related to the instructor than they expected beforehand. It can therefore be assumed that students positively experienced the organization of the module in general.

Overall, students' perceptions towards the use of LA are positive. Regarding the *general statistics* (e.g. students' task on time and assignments), students of the experimental condition agree significantly less with the disadvantages than students of the control condition after the intervention was conducted (Wilks' Lambda = .976, $F(1, 255) = 6.281$, $p = .013$). Students of the experimental condition are more positive against the idea that *general statistics* stimulate connectedness and personalization (Wilks' Lambda = .982, $F(1, 255) = 4.560$, $p = .034$). Table 1 presents the descriptives. It seems possible that these results are due to the positive experiences of students of the experimental condition with the LA intervention. No significant differences between conditions were found for the scale added value of LA ($p > 0.05$). Other quantitative results can be presented at the doctoral consortium.

Table 1: Descriptives for LA situation focusing on *general statistics*

Condition	Scale	Pretest M(SD)	Posttest M(SD)
Control	Added value	4.00(.61)	4.05(.50)
	Disadvantage	2.67(.86)	2.78(.73)
	Connectedness and personalization	3.54(.63)	3.56(.54)
Experimental	Added value	3.98(.71)	4.03(.66)
	Disadvantage	2.86(.78)	2.72(.84)
	Connectedness and personalization	3.42(.75)	3.63(.61)

Preliminary results based on the focus groups indicate that students clearly perceived the implementation of LA as useful. Some students argue that they preferred a specific type of data. For example, statistics focusing on *students' degree of understanding, interests and desires* was indicated as highly valuable. Besides, the majority of the students of the experimental condition found it stimulating when their needs were explicitly considered. They also experienced that the instructor had adapted the content of the workshop to their personal needs, making the adaption formative in nature (Ferguson et al., 2017). It is important that the instructor does something with the information gathered by the LA. Closing the loop through effective interventions that reach the learners is a crucial aspect in the design of LA interventions (Clow, 2012). Other implications for practice and challenges for further research can be presented and discussed during the doctoral consortium.

Based on a design-based research approach (McKenney & Reeves, 2012), the second quasi-experimental study will be carried out in October 2018. This intervention focusses on the supporting role of LA for the instructor to foster an autonomy and competence supporting learning environment. Interviews with the teachers and students are planned, in addition to the pre-post quantitative student questionnaires. Findings from this study will provide directions for further investigation.

REFERENCES

- Ameloot, E., & Schellens, T. (2018). Student Teachers' Perceptions of Using Learning Analytics in a Blended Learning Context. *Proceedings of INTED2018 Conference*, (March), 4508–4516.
- Boelens, R., De Wever, B., & Voet, M. (2017). Four key challenges to the design of blended learning: A systematic literature review. *Educational Review Research*, 22, 1–18. <https://doi.org/10.1016/j.edurev.2017.06.001>
- Boelens, R., Van Laer, S., De Wever, B., & Elen, J. (2015). *Blended learning in adult education: Towards a definition of blended learning*. *Adult Education and Training*. Retrieved from <http://www.iwt-alo.be/wp-content/uploads/2015/08/01-Project-report-Blended-learning-in-adult-education-towards-a-definition-of-blended-learning.pdf>
- Brand-Gruwel, S. (2012). *Leren in een digitale wereld: Uitdagingen voor het onderwijs*. Heerlen. Retrieved from <http://dspace.ou.nl/bitstream/1820/4456/1/ORATIE-def 04-10-2012.pdf>
- Cabero Almenara, J., del Carmen Llorente Cejudo, M., & Puentes Puente, A. (2010). Online students' satisfaction with blended learning. *Scientific Journal of Media Literacy*, 18(35), 149–156. <https://doi.org/10.3916/C35-2010-03-08>
- Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E., Van der Kaap, J., ... Verstuyf, J. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, 39(2), 216–236. <https://doi.org/10.1007/s11031-014-9450-1>
- Clow, D. (2012). The Learning Analytics Cycle: Closing the loop effectively. In S. Buckingham Shum, D. Gasevic, & and R. Ferguson (Eds.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK12)* (pp. 134–138). Vancouver, BC, Canada: New York: ACM.
- Delfino, M., & Persico, D. (2007). Online or face-to-face? Experimenting with different techniques in teacher training. *Journal of Computer Assisted Learning*, 23(5), 351–365. <https://doi.org/10.1111/j.1365-2729.2007.00220.x>
- Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): a reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning*, 32(3), 281–290. <https://doi.org/10.1111/jcal.12135>
- Ferguson, R., Barzilai, S., Ben-Zvi, D., Chinn, C. A., Herodotou, C., Hod, Y., ... Whitelock, D. (2017). *Innovating Pedagogy 2017: Open University Innovation Report 6*. UK. Retrieved from <https://iet.open.ac.uk/file/innovating-pedagogy-2017.pdf>
- Ginns, P., & Ellis, R. (2007). Quality in blended learning: Exploring the relationships between on-line and face-to-face teaching and learning. *The Internet and Higher Education*, 10(1), 53–64. <https://doi.org/10.1016/j.iheduc.2006.10.003>
- Howitt, D. (2011). Focus groups. In G. Van Hove & Claes (Eds.), *Qualitative research and educational sciences: A reader about useful strategies and tools* (pp. 109–131). Hampshire, Great Britain: Pearson Education Limitation.
- Irvine, V., Code, J., & Richards, L. (2013). Realigning higher education for the 21st-century learner through multi-access learning. *MERLOT Journal of Online Learning and Teaching*, 9(2), 172–186. Retrieved from <http://search.proquest.com/openview/938f9422f6700897a753a827d37f539a/1?pq-origsite=gscholar>
- Jivet, I., Specht, M., Scheffel, M., & Drachsler, H. (2018). License to evaluate: preparing learning analytics dashboards for educational practice. In *Proceedings of International Conference on Learning Analytics and Knowledge*. <https://doi.org/10.1145/3170358.3170421>
- Laurillard, D. (2014). *Thinking about blended learning: A paper for the thinkers in residence programme*. London. Retrieved from http://www.kvab.be/denkensprogramma/files/DP_BlendedLearning_Thinking-about.pdf
- Long, P., & Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46, 30–32. <https://doi.org/10.1145/2330601.2330605>
- Mangaroska, K., & Giannakos, M. N. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, PP(1), 1. <https://doi.org/10.1109/TLT.2018.2868673>
- Matuk, C., Linn, M., & Eylon, B.-S. (2015). Technology to support teachers using evidence from student work to customize technology-enhanced inquiry units. *Instructional Science*, 43(2), 229–257. <https://doi.org/10.1007/s11251-014-9338-1>
- McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. New York, NY: Routledge.

- Osguthorpe, R. T., & Graham, C. R. (2003). Blended Learning Environments. *Quarterly Review of Distance Education*, 4, 227–233. <https://doi.org/Article>
- Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers & Education*, 54(1), 222–229. <https://doi.org/10.1016/j.compedu.2009.08.005>
- Pardo, A., Powuet, O., Martinez-Maldonado, R., & Dawson, S. (2017). Provision of Data-Driven Student Feedback in LA & EDM. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (First edit, pp. 163–174). Society for Learning Analytics Research. <https://doi.org/10.18608/hla17>
- Preston, G., Phillips, R., Gosper, M., McNeill, M., Woo, K., & Green, D. (2010). Web-based lecture technologies : Highlighting the changing nature of teaching and learning. *Australasian Journal of Educational Technology*, 26(6), 717–728. <https://doi.org/10.14742/ajet.v26i6.1038>
- Reeve, J., Deci, E. L., & Ryan, R. M. (2004). Self-determination theory: A dialectical framework for understanding socio-cultural influences on student motivation. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited* (pp. 31–60). Greenwich, CT: Information Age Press.
- Rubens, W. (2013). *E-learning: Trends en ontwikkelingen*. Tilburg: InnoDoks Uitgeverij.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist Association*, 55(1), 68–78. <https://doi.org/10.1037/110003-066X.55.1.68>
- Schmitz, M., Limbeek, E. Van, Greller, W., Sloep, P., & Drachsler, H. (2017). Opportunities and challenges in using Learning Analytics in Learning Design. *EC-TEL 2017: Data Driven Approaches in Digital Education*, 10474, 209–223. <https://doi.org/10.1007/978-3-319-66610-5>
- Shum, S. B. (2012). *Learning analytics (UNESCO policy brief)*. Retrieved from <http://verstekapro.ru/>
- Tempelaar, D., Rienties, B., Mittelmeier, J., & Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior*, 78, 408–420. <https://doi.org/10.1016/J.CHB.2017.08.010>
- Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167. <https://doi.org/10.1016/j.chb.2014.05.038>
- Vansteenkiste, M., & Ryan, R. M. (2013). On Psychological Growth and Vulnerability: Basic Psychological Need Satisfaction and Need Frustration as a Unifying Principle. *Journal of Psychotherapy Integration*, 23(3), 263–280. <https://doi.org/10.1037/a0032359>
- Vansteenkiste, M., & Soenens, B. (2015). *Vitamines voor groei. Ontwikkeling voeden vanuit de Zelf-Determinatie Theorie*. Leuven: Acco.
- Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Assche, F. Van, Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6), 1499–1514.
- Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2(2), 5–13. <https://doi.org/10.18608/jla.2015.22.2>

Analytics for the Measurement of Process Dimensions of Self-Regulated Learning and Feedback Impact

John Saint

School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
john.saint@ed.ac.uk

ABSTRACT: Blended learning environments have the potential to provide educators with valuable insights into learner behaviours and strategies. Capturing and analysing learner data, using traditional frequency-based statistical methods, is a challenge if the objective is to understand self-regulated learning (SRL) as a dynamic process. Current research on SRL has recognised the potential of data science methods for analysis of temporal processes. To explore this potential, this research aims to 1) improve the measurement of SRL by deriving micro-level processes from trace data; 2) analyse these micro-level processes for temporal associations; 3) explore how such temporal associates between micro-level processes are correlated with learning strategies; and 4) assess the impact of formative data-driven feedback on these SRL processes. We have undertaken two preliminary studies and found that certain temporal activity traits relate to performance in the summative assessments attached to the course, mediated by strategy type. In addition, more strategically minded activity, embodying learner self-regulation, generally proves to be more successful than less disciplined reactive behaviours.

Keywords: Learning Analytics, Self-Regulated Learning, Micro-Level Processes, Process Mining, Student Feedback

1 MOTIVATION

Self-regulation is a key skill for strategically mature students as it informs how effectively they process feedback (both internal and external) and act upon it (Winne & Hadwin, 1998). In addition, administering developmental feedback to students has a significant effect on their learning journey and academic performance (Hattie & Timperley, 2007). From a metacognitive viewpoint, exponents of self-regulated learning (SRL) are able to succeed by assessing, planning, assimilating, organising, and self-evaluating in an ongoing cycle (Zimmerman, 1990). Therefore, significant benefits can be realised in identifying, articulating, and optimising patterns of SRL. Much of the research around measuring SRL, however, is based not on authentic process data, but on variants of self-report data capture e.g. Bannert, Reimann, & Sonnenberg (2014), and Greene & Azevedo (2009).

Data generated in blended learning environments, specifically those from learning management systems (LMS), provide opportunities for researchers to unlock insights into learner behaviours and strategies. The use of LMS trace data is a promising alternative to self-report data, as it eliminates potential issues of data objectivity and reliability. However, raw trace data cannot, in and of itself, represent self-regulation processes as theorised in well-known models. Therefore, the derivation of

SRL macroprocesses and associated microprocess, as demonstrated by Siadaty et al. (2016c), provides a strong methodological platform on which the current study can build.

2 RESEARCH AREA

2.1 Modelling SRL

Winne (1996) identifies three key aspects of SRL: 1) Cognitive Tactic; 2) Cognitive Strategy; 3) Metacognition. This articulates a learner's management of their own cognitive tactics, and the development of an overarching knowledge management strategy, encompassing self-awareness. Zimmerman's model of SRL also provides a strong and conceptually interpretable model: Self-regulation is presented as a cycle of forethought (planning), performance (of learning event), and self-reflection (Zimmerman, 2000).

In the context of Learning Analytics (LA), Winne advises underpinning SRL research with a proven SRL model. and provides a framework for mapping trace data events to 'inferences' and categorising them to phases of his SRL model (Winne, 2017). There are studies that harvest pure trace data to unlock insights into cognitive tactics and learning strategies e.g. Lust, Vandewaetere, Ceulemans, Elen, & Clarebout (2011), Lust, Elen, & Clarebout (2013), and Kovanović, Gašević, Joksimović, Hatala, & Adesope (2015). They stop short, however, of truly articulating SRL. This study aims to use LMS-generated event logs as its main trace data source, thus diluting the empirical shortcomings of self-report data, yet retaining the vital characteristics of SRL.

2.2 SRL Microlevel Process Analysis

Micro-level process analysis is one of the responses to the challenges of capturing and identifying SRL. This analytical mode presents a way of contextualising sequences of engagement events into recognised categorisations of SRL. These categorisations are themselves categorised by macro-level processes, which form the main constructs of the chosen SRL model. Significant work in this area was pioneered by Greene and Azevedo (2009) and further explored by Cleary & Zimmerman (2012). Siadaty, Gašević, & Hatala (2016c) build on this substantially by developing a hybrid self-report/trace-based protocol of SRL microanalysis. They posit an SRL model which positions 1) Planning, 2) Engagement, and 3) Evaluation & Reflection as its macro-level processes/SRL phases. Micro-level processes, such as Task Analysis or Working on Task, are categorised to their corresponding macro-level processes. Siadaty et al.'s method was empirically validated in two studies on the self-regulatory patterns of knowledge workers in technology-enhanced environments: Siadaty, Gašević, & Hatala (2016b), and Siadaty, Gašević, & Hatala (2016a). These studies provide a critical empirical SRL bedrock. They do not, however, explore inter/intra-strategy temporal differences, and the research subjects are knowledge workers, not further/higher education students. Additionally, although the impact of various scaffolding interventions is assessed, it does not represent a true study of feedback as a mediator of SRL. This study seeks to address this empirical gap.

2.3 Event-based Process Analysis

Process Mining (PM) is an analytical discipline that straddles data mining, machine learning, and business process modelling. Being data-driven but process-centric, we can view it as a missing link

between data science and process science (van der Aalst, 2016). In many PM studies, the processes are of a tactic-level granularity. This study aims to harness PM for micro and macro-level analysis. PM provides a vital temporal dimension that is not afforded by traditional statistical methods e.g., Lust et al. (2011). Some studies recognise time as a dimension, but this is restricted to measurement of time on task, and not a reflection of true inter-process temporal dynamics e.g., Kovanović et al. (2015). The current study seeks to unlock insights into the temporal sequence of study activities as exhibited by exponents of SRL.

2.4 Learner Strategy Detection

Bannert et al. (2014) use process mining techniques to analyse think-aloud data logged from a student-group's navigation through an LMS. Their aim is to provide a comparison of process models of high and low performing students. Lust et al. (2011) use clustering to identify user-profiles through learner behaviours, identifying profiles through frequency of activity. Kovanovic et al. (Kovanović et al., 2015), Jovanovic et al. (Jovanović, Gašević, Dawson, Pardo, & Mirriahi, 2017), and Fincham et al. (Fincham, Gasevic, Jovanovic, & Pardo, 2018) all demonstrate sophisticated deployments of learner clustering around user-profiles and strategic learning sequences. The resultant group comparisons are insightful but leave a clear empirical gap for intra and inter-strategy articulation in the context of SRL micro-analysis. The current study aims to provide this specific comparative analysis.

2.5 Student Feedback

Providing quality feedback in HE is inherently challenging. These challenges have intensified with the increasing massification of education. Two significant studies, both using the same high-volume LMS trace data, provide valuable insights into the impact of customised automated feedback. Pardo et al. (Pardo, Jovanovic, Dawson, Gašević, & Mirriahi, 2017) demonstrate a positive association between feedback messaging and both student satisfaction and assessment performance. Fincham et al. (Fincham et al., 2018) detected tactical transition and strategic improvement (linked to assessment performance) as the result of feedback interventions. The current study aims to build on this research to assess the impact of feedback on SRL patterns in learners.

2.6 Research Questions

RQ1. To what extent can micro-level SRL processes be derived from trace data collected by conventional virtual learning environments?

RQ2. To what extent can temporal associations between micro-level SRL processes be derived through the analysis of trace data?

RQ3. What are the differences in SRL micro-level processes exhibited by students following different learning strategies?

RQ4. What is the impact of conditionally administered, analytics-based formative feedback on the micro-level SRL processes of students who follow different learning strategies?

3 METHODOLOGY

The trace data for this study come from two sources: The first were collected from an LMS attached to a computing course at an Australian university. The datasets provide LMS trace data from four cohorts of a course, spanning 2014 to 2017 (Pardo & Mirriah, 2017). The course was based on a flipped classroom pedagogy and the data relate to students' engagement with the online activities as preparation for the face-to-face learning sessions. Each time a student engaged with an element of the LMS, a learning event record was generated. These events, which are collectively called trace data, provide the source for our analyses.

The starting point of PM is a dataset in the form of an event log. The required elements to run a PM algorithm are: Case, a process instance; Activity, a well-defined step in a broader process; Timestamp, providing the temporality that is key to this study. Each LMS event record contains a student ID number (which serves as our PM case), a completed study action (which serves as our PM activity), and a timestamp.

To identify micro-level SRL processes (RQ1), we extract trace data and utilise the mapping method outlined in the Siadaty (and associated) studies i.e. trace event → micro-level process → macro-level process/SRL construct. To address RQ2, we will build on the PM techniques explored in our preliminary study (see section 4) and extract temporal relationships from the SRL microprocesses. To address RQ3, we will cluster students in strategy groups, using the methods employed by Bannert et al. (2014) and Fincham et al. (2018). We will perform pair-wise comparative analyses, using appropriate PM algorithms, to articulate the differences in learner strategies from a temporal micro-level perspective. Finally, will consolidate these methods to measure the impact of feedback interventions on learner behaviours, mediated by strategy type (RQ4).

4 PRELIMINARY RESULTS

We have undertaken two preliminary studies, using the 2014 cohort LMS data; one study was presented at the EC-TEL 2018 conference as a full research paper (Saint, Gasevic, & Pardo, 2018). The second study, building on the EC-TEL paper, but employing micro-level parsing in a recognised model of SRL, has been submitted to LAK 19. In both cases, we employed the R package pMineR (Gatta et al., 2017) to train process models using first order Markov chains. The focus of both studies is the analysis of tactical cognitive processes in a temporal/stochastic context, and how it informs learning strategy and performance. We found that certain temporal activity traits relate to performance in the summative assessments attached to the course, mediated by strategy type. In addition, more strategically minded activity, embodying learner self-regulation, generally proves to be more successful than less disciplined reactive behaviours.

5 FUTURE AGENDA & PUBLICATION PLAN

It is anticipated that this doctoral submission will be a linked collection of published journal articles. These publications will be interspersed by conference paper submissions to future Learning Analytics and Knowledge (LAK) conferences. Building on the preliminary studies, we hope to expand and formalise a process mining/microprocess methodology and replicate it across the remaining LMS

cohorts (2015-2017). Finally, it is hoped that data harvested from subsequent studies will provide a means of testing the generalisability, and scalability of a consolidated LA methodology.

REFERENCES

- Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, 9(2), 161–185. <https://doi.org/10.1007/s11409-013-9107-6>
- Cleary, T. J., & Zimmerman, B. J. (2012). A Cyclical Self-Regulatory Account of Student Engagement: Theoretical Foundations and Applications. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 237–257). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-2018-7_11
- Fincham, O. E., Gasevic, D. V., Jovanovic, J. M., & Pardo, A. (2018). From Study Tactics to Learning Strategies: An Analytical Method for Extracting Interpretable Representations. *IEEE Transactions on Learning Technologies*, 1–13. <https://doi.org/10.1109/TLT.2018.2823317>
- Gatta, R., Lenkiewicz, J., Vallati, M., Rojas, E., Damiani, A., Sacchi, L., ... Valentini, V. (2017). pMineR: An Innovative R Library for Performing Process Mining in Medicine. In A. ten Teije, C. Popow, J. H. Holmes, & L. Sacchi (Eds.), *16th Conference on Artificial Intelligence in Medicine, AIME 2017*. Vienna. <https://doi.org/10.1007/978-3-319-59758-4>
- Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34(1), 18–29. <https://doi.org/10.1016/j.cedpsych.2008.05.006>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81.
- Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, 33, 74–85. <https://doi.org/10.1016/j.iheduc.2017.02.001>
- Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Does time-on-task estimation matter? Implications for the validity of learning analytics findings. *Journal of Learning Analytics*, 2(3), 81–110. <https://doi.org/10.18608/jla.2015.23.6>
- Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, 27, 74–89. <https://doi.org/10.1016/j.iheduc.2015.06.002>
- Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. *Computers & Education*, 60(1), 385–395. <https://doi.org/10.1016/j.compedu.2012.09.001>
- Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J., & Clarebout, G. (2011). Tool-use in a blended undergraduate course: In Search of user profiles. *Computers & Education*, 57(3), 2135–2144. <https://doi.org/10.1016/j.compedu.2011.05.010>
- Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2017). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, 0(0). <https://doi.org/10.1111/bjet.12592>
- Pardo, A., & Mirriahi, N. (2017). Design, Deployment and Evaluation of a Flipped Learning First-Year Engineering Course. In C. Reidsema, L. Kavanagh, R. Hadgraft, & N. Smith (Eds.), *The Flipped Classroom: Practice and Practices in Higher Education* (pp. 177–191). Springer Singapore. <https://doi.org/10.1080/14739879.2015.1109809>
- Saint, J., Gasevic, D., & Pardo, A. (2018). *Detecting Learning Strategies Through Process Mining: 13th European Conference on Technology Enhanced Learning, EC-TEL 2018, Leeds, UK, September 3-5, 2018, Proceedings*. https://doi.org/10.1007/978-3-319-98572-5_29
- Siadat, M., Gašević, D., & Hatala, M. (2016a). Associations between technological scaffolding and micro-level processes of self-regulated learning: A workplace study. *Computers in Human*

- Behavior*, 55, 1007–1019. <https://doi.org/10.1016/j.chb.2015.10.035>
- Siadaty, M., Gašević, D., & Hatala, M. (2016b). Measuring the impact of technological scaffolding interventions on micro-level processes of self-regulated workplace learning. *Computers in Human Behavior*, 59, 469–482. <https://doi.org/10.1016/j.chb.2016.02.025>
- Siadaty, M., Gašević, D., & Hatala, M. (2016c). Trace-Based Microanalytic Measurement of Self-Regulated Learning Processes Moray House School of Education and School of Informatics. *Journal of Learning Analytics*, 3(1), 183–214.
- van der Aalst, W. (2016). *Process Mining: Data Science in Action. Process Mining* (Second). Berlin: Springer-Verlag. <https://doi.org/10.1007/978-3-642-19345-3>
- Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, 8(4), 327–353. [https://doi.org/10.1016/S1041-6080\(96\)90022-9](https://doi.org/10.1016/S1041-6080(96)90022-9)
- Winne, P. H. (2017). Learning Analytics for Self-Regulated Learning. In C. Lang, G. Siemens, A. F. Wise, & D. Gašević (Eds.), *The Handbook of Learning Analytics* (1st ed., pp. 241–249). Alberta, Canada: Society for Learning Analytics Research (SoLAR). Retrieved from <http://solaresearch.org/hla-17/hla17-chapter1>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as Self-Regulated Learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, New Jersey: Lawrence Erlbaum Associates. <https://doi.org/10.1016/j.chb.2007.09.009>
- Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, 25(1), 3–17. <https://doi.org/10.1207/s15326985ep2501>
- Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). San Diego: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-012109890-2/50031-7>

Innovation to Improve Learning: A Study of Content Modalities in Universal Design for Learning

Author: Catherine A. Manly
University of Massachusetts Amherst
cmanly@umass.edu

ABSTRACT: Educational researchers have surprisingly little empirical evidence showing how practices facilitated by new technological capabilities, such as those recommended by Universal Design for Learning (UDL), translate to learning success. I address this by investigating student success in courses using a learning-analytics-infused system facilitating UDL implementation. I investigate the causal effect of using multiple content modalities (i.e., text, video, audio, or interactive content) on learning outcomes for nontraditional undergraduates. I aim to identify any benefit when a student utilizes more than one modality while learning. I incorporate innovative practice in theory, data, and methods. Theoretically, use of multiple modalities deserves deeper investigation to confirm its effect. Data-wise, studying UDL with learning data gathered in 20-minute intervals across more than 50 courses is new. Methodologically, using multiple causally-oriented methods in educational research is unusual, particularly using learning analytics to help identify recommendations for improving student learning. My project combines these research innovations.

Keywords: Postsecondary education, universal design, modality, online education

1 PROBLEM AND GOALS

While it seems reasonable that designing higher education learning experiences intended to be universally accessible to students of all abilities would lead to improved student success, potentially improving course completion and graduation rates, well-designed empirical research corroborating this intuition is surprisingly sparse. Universal Design for Learning's (UDL) empowering frame views all individuals as capable learners given a supportive environment that does not disable their capacity (UDL; Burgstahler, 2015). Recognizing that varied student learning needs too frequently are not adequately addressed through course design, UDL posits that students benefit from multiple means of engagement, representation, action and expression. Educational researchers have surprisingly little empirical evidence to show how specific educational practices facilitated by new technological capabilities, such as UDL practices, translate to learning success and therefore how to improve the affordability of higher education's core educational mission by reducing failure rates and the need to repeat courses. I address this gap by investigating student success in courses that use a learning-analytics-infused system which facilitates implementation of UDL, a framework arising from disability studies that is grounded in cognitive science. While targeted at improving the experience of students with disabilities, UDL is posited to extend beyond addressing students with disabilities to hold relevance for all students.

Specifically, I investigate the causal effect of using multiple content modalities (i.e., text, video, audio, or interactive content) on student learning outcomes for nontraditional undergraduates. These older women students, juggling work and family, need affordable education (most are low-

income), and may benefit from an individually tailored educational approach. Combining data from an adaptive learning system and multiple campus support systems, I aim to discover whether the multiple content representation of UDL is beneficial for these students. I do this because the efficacy of using multiple content modalities as proposed by UDL still needs to be rigorously and empirically investigated (Rao, Ok, & Bryant, 2014). I use multiple quasi-causal analytic approaches, seeking to improve the internal validity of my findings through the confirmation of multiple indications. The goal of my study is to identify any beneficial effect that exists when a student utilizes more than one modality when learning course content. (See Table 1 for my research questions.)

2 SIGNIFICANCE OF THE WORK

By investigating a key aspect of UDL, this study offers a needed contribution to both the UDL and higher education literature. Research about how to support the academic success of students with disabilities needs to be extended in higher education (Kimball, Wells, Ostiguy, Manly, & Lauterbach, 2016). By isolating one aspect of UDL, I intend to advance our understanding of a practice that has the potential to help not just students with disabilities, but all students.

From a practical standpoint, this work's primary contribution is to provide guidance for faculty development efforts regarding the effectiveness of UDL's incorporation. This aim will be achieved by extending prior research in important ways, including: 1) examining the effects of use of multiple modalities on student learning outcomes in ways that have not yet been explored, 2) using multiple campus systems to gather more comprehensive data about students than has typically been studied, 3) seeking confirmation for the effectiveness of an aspect of UDL theory, and 4) providing proof-of-concept analyses employing rigorous, advanced analytical methods in a study that could be extended to other circumstances to investigate, predict, and present analysis results about the connection between elements of UDL and student success. Overall, my inquiry will work toward addressing systemic inequality in higher education, particularly for students who have been traditionally underserved, by improving understanding of providing appropriate options and guidance for all our students as they make choices about how they will engage with course content.

3 CURRENT KNOWLEDGE

The educational implications of natural learning variations have been explored for several decades by a community of scholars and practitioners interested in universally designing educational experiences (Burgstahler, 2015; Silver, Bourke, & Strehorn, 1998). Despite this interest, universal design frameworks still struggle to gain acceptance in academic culture (Archambault, 2016) and remain understudied in postsecondary education (Rao et al., 2014). My study delves into one UDL aspect proposed to be important in addressing the variety of perceptive and processing abilities that students bring to their education.

I do not claim my study will answer all pertinent questions about the efficacy of providing options for perception, but in agreement with Crevecoeur and co-authors (2014), I believe that rigorous investigation of various tenets of UDL, both separately and together will be necessary for the field to gain a deep understanding of what aspects of UDL are key and why. Likewise, McGuire, Scott and Shaw (2006) conclude that more rigorous research is necessary "to allow this potentially powerful model to be developed and proven *before* [emphasis in original] it is widely—and possibly

ineffectively–implemented” (McGuire et al., 2006, pp. 173–174). From a practical standpoint, given that faculty are frequently advised to begin implementation of UDL in tractable pieces, knowing which aspects of UDL offer substantial benefits for student learning by themselves holds importance, as those would be the most appropriate places to encourage faculty to begin course redesign.

The universal design literature remains notable for its lack of effectiveness-oriented peer-reviewed research (Kimball et al., 2016). Multiple literature reviews spanning several decades have uncovered relatively few articles pertaining to universal design (Rao, Ok, & Bryant, 2014). Mangiatordi and Serenelli’s (2013) review of 80 ERIC abstracts between 2000 and 2012, found 19 mentioning research results (all positive), including 4 that demonstrated academic improvement by students, supporting an expectation of positive learning outcomes for my study.

I explore the boundaries of UDL’s applicability since fully universally designed instruction remains a high aspiration that can be difficult to achieve in practice. Tutoring can be considered a necessary augmentation to instructional methods when particular students have individualized learning needs that are not sufficiently addressed through existing course design. This is consistent with Edyburn’s concern that, “we need to renew our commitment to equitably serving all students in the event that our UDL efforts fall short” (Edyburn, 2010, p. 40). Tutors individualize instruction of content, customizing presentation to an even greater degree than is possible otherwise, even with adaptive learning technology, as is used in the present study. Given the demonstrated benefits for students both in course outcomes and persistence, tutoring benefits are expected. I posit providing content through multiple modalities combined with tutoring may provide the additional assistance that struggling students need to succeed if the presentation of material does not address sufficient learning variability. I view tutoring as augmenting the design of the course in ways that have the potential to address gaps in the universality of content presentation, since a tutor will explain the material in a highly interactive and personalized way that goes beyond other ways of presenting the content.

4 CONTRIBUTIONS OF THIS PROJECT’S APPROACH

My study incorporates three areas of innovative educational practice, including theory, data, and method. Regarding theoretical innovation, UDL challenges higher education institutions to design students’ learning experiences intentionally paying attention to including multiple means of achieving key elements for facilitating learning. However, the connection between college student learning outcomes and presenting course content through multiple modalities is not yet well understood. For instance, we do not know how much learning outcomes would improve if students were guided to the most appropriate modality, combination of modalities, or modalities plus additional tutoring support. Very few researchers have studied the connection between multiple content modalities and student course outcomes, even as part of larger research on UDL (Rao et al., 2014). While guidance for faculty exists, more attention to the development of universal design practices based in extensive, high-quality, empirically-based research remains needed.

With regard to data innovation, technological advances make it increasingly straightforward to gather data about student interactions with course material, particularly in online courses where most student activity and interactions leave recordable and analyzable traces. These traces may include logs with date/time stamps, duration of activity, and modality utilized, for example. Such

automatically collected electronic data makes investigating the connection between content representation and student success more tractable than before. This research utilizes data traces recorded as students progress through online courses to investigate the effectiveness of representing content to students through multiple modalities. Given the relative newness of extensive learning data availability, few prior studies have empirically investigated the efficacy of providing multiple modalities for presenting content.

Methodologically, the literature includes multiple calls for better research designs and evidence of UDL's efficacy, as well as investigation of learning outcomes. There has been a notable lack of assessment of UDL's components through causally-oriented investigation (Crevecœur et al., 2014). In general, the approach used in this study of using directed acyclic graphs (DAGs) to study causal mechanisms has seen only sparse use to date in higher education, and that use has been more in educational technology journals (e.g., Xenos, 2004) than directly in higher education. However, use of DAGs can help make alternative models explicit, facilitating identification of which models are more likely. I aim to extend knowledge of the efficacy of offering multiple means of representation by investigating how use of multiple modalities connects causally to student learning measures. The challenges associated with designing high quality, causally-oriented quantitative studies of educational outcomes make this study particularly significant as a guide for future research efforts.

5 RESEARCH METHODOLOGY

The novelty of this area of study suggests combining exploratory and confirmatory approaches. In an exploratory sense, I look descriptively at use of multiple modalities, investigating both variation and patterns in the data. In a confirmatory sense, I use associative statistical methods to investigate hypothesized relationships among the variables as indicated by theory. I also probe causal connections in these relationships as allowed by the data, with the intent of indicating areas of interest for future experimental or quasi-experimental research. I aim to identify and communicate potential tutoring intervention points within a course using a learning analytics approach.

5.1 Data and Ethics

The data come from a single institution, including 55 undergraduate online courses in a variety of disciplines. Each six-week course's design has been broken down into 20-minute learning activities, with approximately 5-15 activities per week. A student's prior knowledge of the material to be covered is assessed at the beginning of each week and this knowledge score is updated to reflect their evolving understanding at the completion of each activity. Data is being collected in two phases. Data from Spring 2018 will be analyzed in an instrumental variables analysis, using the randomization from a randomized controlled trial (RCT) as the instrument. Additional data from Fall 2018 (after the RCT ended) through February 2019 will be incorporated for other analyses, including a change score panel data analysis and propensity score analysis.

Regarding ethical considerations, implementing recommendations for tutoring implies students willing to share learning data. A hallmark of the support advertised to students at the institution I study is ongoing individualized support throughout their online degree experience. This mitigates potential ethical concerns regarding use of student data, because the students know information about their online activities is being collected and analyzed expressly for their educational benefit.

5.2 Method

Beyond describing these data both through tabulations and graphs, I use associational and causally-oriented approaches to statistical inference. Investigating causal effects offers a particularly important and too often overlooked direction for higher education research (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). I intentionally order my analyses to build upon each other to the extent possible in order to facilitate analysis and interpretation of complex models. In sum, I begin my inquiry with regression and structural equation modeling, then move to instrumental variable, propensity score, and panel data analyses in a coordinated fashion, providing greater internal validity together than any individual analysis separately. The research questions in Table 1 guide my inquiry through the corresponding analyses and data.

Table 1: Correspondence between research questions, analysis method, and data.

Research Question	Method	Data
1.VARIATION. To what extent do students vary in their usage of multiple content representations?	Descriptive statistics.	Spring 2018, Fall 2018
2.PATTERN. What are the most common patterns of student use of content representations?	Descriptive statistics and graphical representation.	Spring 2018, Fall 2018
3.GAIN. What is the relationship between use of multiple content representations and student learning gains (i.e., knowledge state assessed at entry and exit for an activity)?	Regression analysis.	Spring 2018, Fall 2018 (Aggregate at the activity level)
4.BYSUBJECT. Within individual subjects (e.g., English, History, Humanities, Psychology), how does <i>variation</i> in use of multiple content representations across classes relate to differences in student learning outcomes? Similarly, what about <i>patterns of use</i> of multiple content representations?	Structural equation modeling.	Spring 2018, Fall 2018 (Aggregate at the weekly level (i.e., multiple activities per week))
5.EFFECT. What are the effects of choosing a second modality (either text, video, audio, or interactive) for learning course material on subsequent learning outcomes?	a. Instrumental variable. Instrument = assignment to RCT treatment group Treatment = using any second representation Outcome = knowledge gain b. Propensity score analysis. Treatment = using any second representation Control = using only one c. Change score panel analysis.	a. RCT data from Spring 2018 (Aggregate at the activity level) b. Spring 2018, Fall 2018 (Aggregate at the weekly level) c. Spring 2018, Fall 2018 (Aggregate at the activity level)
6.COMBINATION. What combination of modality switches and tutoring	Panel data analysis. Visualization of predictions.	Spring 2018, Fall 2018 (Individual courses, no

maximize later activity, module, and
course success for struggling students?

aggregation)

6 CURRENT STATUS OF THE WORK

Preliminary results from Spring 2018 data indicated a meaningful benefit from the use of multiple modalities, with a medium effect size (Cohen's h). Specifically, preliminary regression analysis accounting for clustering by student found a significant percentage point improvement of 0.072, averaged over all students in the sample (i.e., when calculating the average marginal effect, or AME). A change score panel data analysis accounting for clustering by student indicated a statistically significant percentage point change of 0.078 in a student's knowledge score across an activity, averaged over all students, explaining almost $R^2 = 3\%$ of the variance in the outcome. A benefit of this type of change score analysis was that it accounted for time-invariant student characteristics that could otherwise confound an effect estimate, thus coming closer to a desired causal estimate than the associational regression-based analysis.

Given the six-week nature and non-traditional timing of these online courses, data collection will end in February 2019. I am preparing to conduct full analyses once data collection is complete. I will gain access to these data in late February or March and will be conducting data cleaning and analyses this spring. If these preliminary results are confirmed with new data from Fall 2018 and additional analyses, my results will provide guidance for faculty development efforts by showing that demonstrable benefit to student learning comes from straightforward design elements following a UDL principle.

In conclusion, my study searches for multiple indications of a relationship between usage of multiple content representations and student outcomes. I aim to better understand UDL's proposition that providing multiple means of content representation benefits student learning.

REFERENCES

- Archambault, M. (2016). *The diffusion of universal design for instruction in post-secondary institutions* (Ed.D. dissertation). University of Hartford, Hartford, CT.
- Burgstahler, S. E. (2015). *Universal design in higher education: From principles to practice* (2nd ed.). Cambridge, MA: Harvard Education Press.
- Crevecoeur, Y., Sorenson, S., Mayorga, V., & Gonzalez, A. (2014). Universal design for learning in K-12 educational settings: A review of group comparison and single-subject intervention studies. *The Journal of Special Education Apprenticeship*, 3(2). Retrieved from <http://scholarworks.lib.csusb.edu/josea/vol3/iss2/1>
- Edyburn, D. L. (2010). Would you recognize universal design for learning if you saw it? Ten propositions for new directions for the second decade of UDL. *Learning Disability Quarterly*, 33(1), 33–41. <https://doi.org/10.1177/073194871003300103>
- Kimball, E. W., Wells, R. S., Ostiguy, B. J., Manly, C. A., & Lauterbach, A. (2016). Students with disabilities in higher education: A review of the literature and an agenda for future research. *Higher Education: Handbook of Theory and Research*, 31, 91–156.

- Mangiatoridi, A., & Serenelli, F. (2013). Universal design for learning: A meta-analytic review of 80 abstracts from peer reviewed journals. *Research on Education and Media*, 5(1), 109–118.
- McGuire, J. M., Scott, S. S., & Shaw, S. F. (2006). Universal design and its applications in educational environments. *Remedial and Special Education*, 27(3), 166–175.
- Rao, K., Ok, M. W., & Bryant, B. R. (2014). A review of research on universal design educational models. *Remedial and Special Education*, 35(3), 153–166. <https://doi.org/10.1177/0741932513518980>
- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.
- Silver, P., Bourke, A., & Strehorn, K. C. (1998). Universal instructional design in higher education: An approach for inclusion. *Equity & Excellence in Education*, 31(2), 47–51. <https://doi.org/10.1080/1066568980310206>
- Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, 43(4), 345–359. <https://doi.org/10.1016/j.compedu.2003.09.005>

Designing a Teacher Dashboard for Interactive Simulations

Diana B López Tavares

Instituto Politécnico Nacional
Calz. Legaria 694, 11500 Mexico City,
diana@cicata.edu.mx

ABSTRACT: A dashboard with data from student interactions with interactive simulations (sims) can give teachers information to help them improve activities and plan lessons following. Sims are open-ended environments that can be manipulated in many complicated ways, generating fine-grained, non-linear process data. Identifying the types of data that can provide useful information for teachers and how to present that information in an easily digestible way is the focus of our research. First, the opinions and needs of the teachers were identified. Using several approaches to visualize student activity data, the dashboard shows individual student interaction patterns with the sim as well as the aggregated information of an entire group. To test the prototype dashboard, data from homework activities in college physics classes using the PhET sim *Capacitor Lab: Basics* was collected and analyzed. Future research will examine dashboard design, use and interpretation across diverse teachers, simulations, and contexts.

Keywords: Educational dashboard, Student engagement, Interactive simulations, User centered design

1 INTRODUCTION

The use of interactive simulations (sims) is increasing in science and math education classrooms (Perkins, et al., 2014; Velasco & Buteler, 2017). How students interact with sims is connected to the level and type of guidance teachers choose in activities coupled with such sims. (Adams, Paulson, & Wieman, 2008; Chamberlain et al., 2014; Moore et al., 2013; Podolefsky et al., 2010; Salehi et al., 2015). Research shows that appropriate guidance can help center the attention of students on sim elements that are important for achieving specific learning goals. However, excessive guidance can lead students to follow directions rather than achieving the style of deep exploration associated with high-quality engagement. Student engagement with sims has proven crucial in reaching meta-goals such as self-questioning, exploring, making predictions, testing ideas, designing experiments, monitoring their own understanding, and authentic scientific inquiry (Salehi et al., 2015).

Designing sim-centered activities that strike the optimal balance between generating engagement and focusing on specific elements can be challenging for teachers. For homework activities or online courses, for example, teachers appreciate knowing the extent to which the sim was used to answer questions, if the time assigned is appropriate for completing the activity, or if students interact with specific sim elements. With a dashboard for interactive sims, teachers can get information about how their students interact with the sim and answer these questions.

Significant research focusing on learning analytics dashboards has been done (Corrin Linda, 2014; Dyckhoff, Zielke, Bültmann, Chatti, & Schroeder, 2012; Greller & Drachsler, 2012; Holstein, McLaren, & Aleven, 2018; Klerkx,

Verbert, & Duval, 2017; Verbert et al., 2013; Xhakaj, Alevan, & McLaren, 2016). Focused on a variety of educational tools and contexts, such as Learning Management Systems (LMS), Massive Open Online Courses (MOOCs), and Intelligent Tutor Systems (ITS), this research provides insight into the types of information and visualizations teacher find most useful and actionable for these contexts, as well as the research methodologies used to investigate these questions. The resulting dashboards have used a range of metrics, such as logins, student products, material that students review, scores and grades, to generate their visualizations (Schwendimann et al., 2017). No prior research on dashboard design and use with interactive simulations was found in our literature review.

Open-ended educational environments, like interactive simulations, pose unique challenges for characterizing and communicating useful and actionable information to teachers. In these environments, individual students can manipulate and explore the simulation in many complicated and unique ways. These environments often do not include any concrete measures of achievement. The student data produced by such interactions is thus quite different from other educational web services, requiring new research. Identifying the kinds of data that can generate useful information for teachers and how to present that information in an easily digestible way is the focus of our research.

Some research studies about student experiences and engagement in the classroom have made use of student interaction data (Borek, et. al, 2009; Chamberlain et al., 2014; Moore et al., 2013) Other research with sims analyzed the interaction patterns of students to predict student learning (Käser, Hallinen, & Schwartz, 2017; Perez et al., 2017), demonstrating that student behaviors while trying to solve a challenge correlated with improved learning. These studies used the time of interaction, clicks per minute, sim elements used, evolution of clicks in time, and elements used from the aggregated information of the group of study. While this information was used only for researcher interpretation, the work informs approaches to the design of a tool for teachers. In the current work, the dashboard is intended to be used after the student activity is complete.

2 RESEARCH QUESTION

The central research question addressed by this research is: How can student interaction data from open-ended exploratory environments, such as interactive simulations, be organized and presented in ways that teachers find useful and actionable for instruction?

The goals of the research are to:

- Identify teachers' instructional challenges and questions that can be informed by collection of student interaction data from simulations.
- Design and develop a teacher dashboard that collects, organizes, and presents the information that teachers need to inform their sim-based instruction using accessible, interpretable visualizations.
- Evaluate teacher's interpretation of the dashboard and how they can use that information in their instructional practice.

It is important to emphasize that this research focuses on characterizing aspects of student interaction and engagement with a simulation that are interpretable and actionable by teachers. The data collected and presented in the dashboard does not directly give insight about learning or understanding. Rather, this teacher dashboard aims to impact student learning by improving teacher pedagogical actions with simulations, similar to other teacher dashboard efforts (Holstein et al., 2018).

3 METHODS

Teacher needs survey: As an initial step in the design process of the dashboard, we used information gathered in a survey of teachers regarding the collection and usefulness of student interaction data with interactive sims. A total of 816 teachers responded, representing K12 and college environments.

Simulation student interaction data: The sims used in this work are from the PhET Interactive Simulations project of the University of Colorado Boulder (<https://phet.colorado.edu>). PhET sims that have been instrumented with the PhET-iO extension provide a back-end data stream, in a JSON format, that logs mouse or touch interactions, interactions with specific sim elements, and model response events.

Simulation selected: PhET Interactive Simulations have a wide variation of manipulation complexity. For this prototype dashboard, the sim selected needed to have enough content and interaction to test visualizations ideas for the dashboard but with appropriate constraints. The selected sim was an adaptation of *Capacitor Lab: Basics*. In this sim, students can explore the physics of a parallel plate capacitor. Users can change the plate area, the separation distance, and the capacitor can be connected and disconnected from a battery. In addition, students can change battery voltage, display numeric data and use a voltmeter.

Dashboard design and development process: Initial mock-ups were guided by teachers' survey responses. Some indicators and visualizations were adapted from visualizations useful in prior research studies on engagement with sims (Chamberlain et al., 2014; Moore et al., 2013) and website use (Atterer, Wnuk, & Schmidt, 2006; Navalpakkam & Churchill, 2012). Iterative improvements were made over a series of meetings where expert designers, developers, and teachers from the PhET team reviewed mockups and low-fidelity prototypes.

Classroom Data Collection and Testing: To date, three classroom studies have been conducted in college physics classrooms, with over 3,000 logs of student sim interaction collected. In each class, students were assigned a homework activity. Across the classes, the simulation, the student population, and the instructional design of the homework activity was varied to test the design of the dashboard, its ability to describe student interactions, and its ability to meaningfully compare across sim-based instructional conditions.

4 PROGRESS

4.1 Teacher Survey Analysis

Analysis of the teacher surveys established several areas where a significant fraction of teachers noted the student data need important or crucial. These areas included: information about the state of the sim (e.g. number of students that create a saturated solution...), information about the controls used in the sim and the settings (e.g. range of values used in a slide bar), information about time (e.g. duration of sim use, how interaction changed over time) and the possibility of comparing the interactions of different students.

The open-response survey questions further probed teachers' perspective on the types of information that would be useful and how they would use this data. Qualitative analysis and coding techniques, including "Interpretation Sessions" where quotes that represent the key issue/necessity are extracted from teachers answers and "Affinity diagramming" that is used to summarize patterns across the interviewed population by organizing interviews quotes – are being used to identify teacher necessities and prioritize them (Holstein, McLaren, & Aleven, 2017).

4.2 Dashboard Design and Development

The current dashboard prototype is shown in the Figure 1. The dashboard separates the visualizations into two screens – a screen with information related with the events and time and a screen with information about the elements used.

Users can select the group they want to analyze and get some median values for the group, like total time of interaction, and events (Fig 1-A). The main visualization is the events map (Fig 1-B); here red dots overlayed on a screenshot of the sim represent all the recorded events (clicks, mouse-up and mouse-down). Users can select the time interval of interest (Fig 1-C). In Fig. 1-D each dot represents one student, the coordinates represent the total time of interaction and the total number of events. The red lines mark the median values of total time and events for that group. The user can access individual student data by clicking dots in this graph. This graph in Fig 1-E helps to visualize how student interaction rate evolved over time, with steepness showing moments with more and less interaction with the sim.

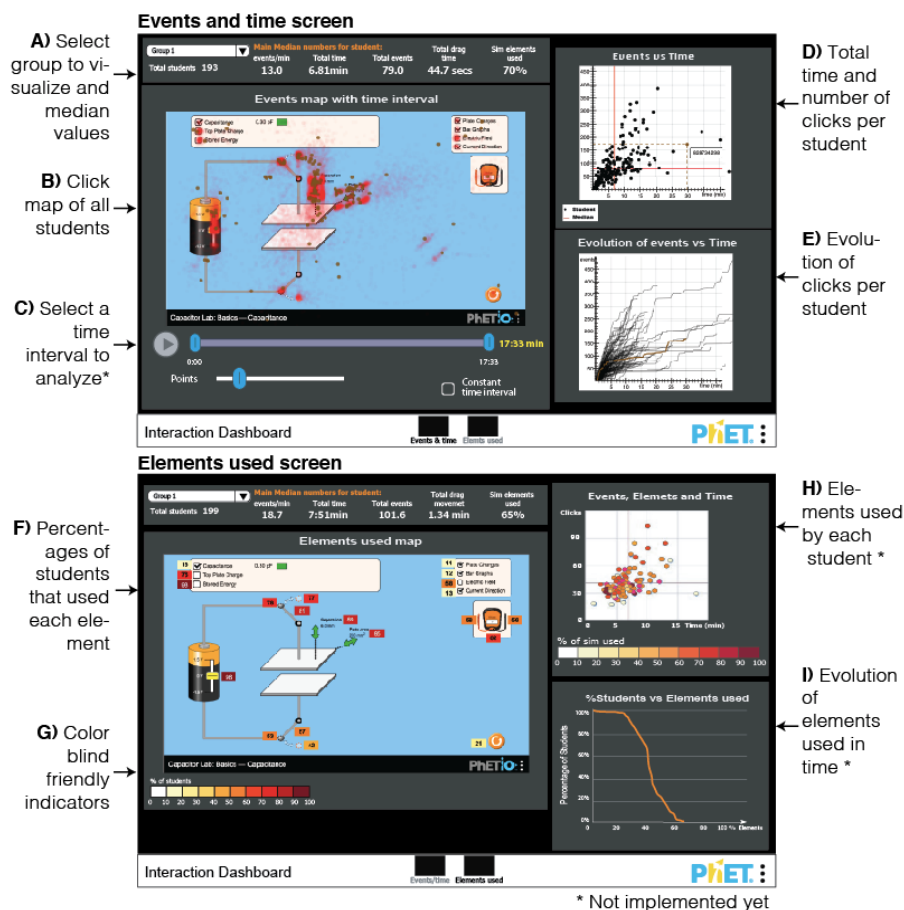


Figure 1: Dashboard design and its components

In the second screen, the elements used map (Fig. 1-F) displays the percentage of students that used each interactive element, and also provides access to the percentage of elements and which elements were used by each individual student. Evolution of the interaction with the elements over time is also planned (Fig. 1-I).

4.3 Preliminary Classroom Results

The first classroom study was conducted in Sprint 2018 and used a variant of *Capacitor Lab: Basics*. Initial data analysis is completed, with results presented at the 2018 Physics Education Research Conference and a peer-reviewed proceedings paper accepted. The dashboard visualizations are shown in Figure 2, and demonstrate early evidence of the instructional insights around student engagement provided by the dashboard tool.

We can immediately see the most used elements are the battery and the controls for plate area and separation. A more detailed analysis provides more insights (matching letters in Fig. 2): (A) The activity required the activation of the Stored Energy checkbox, yet 7% of the students did not activate this element. (B) While most students (96%) change the battery voltage, the pattern of events show few students tested negative voltage values. (C) Slightly less than 90% of the students modify the area and separation of the plates. Most explored the extremes (min and max values) but also a significant number of events in intermediate values for the separation of the plates. This information suggests that several students tested different values of this variable with some combination of plate area values, and separation of the plates had more student interaction than plates area. (D) The activity shows student did connect and disconnect the battery. (E) Both maps show that the elements active by default in the sim (checkboxes selected) have low interaction. (F) The voltmeter has low interaction (62% students touch it and less than 60% use the probes), despite the fact that the lesson would benefit from interaction with it.

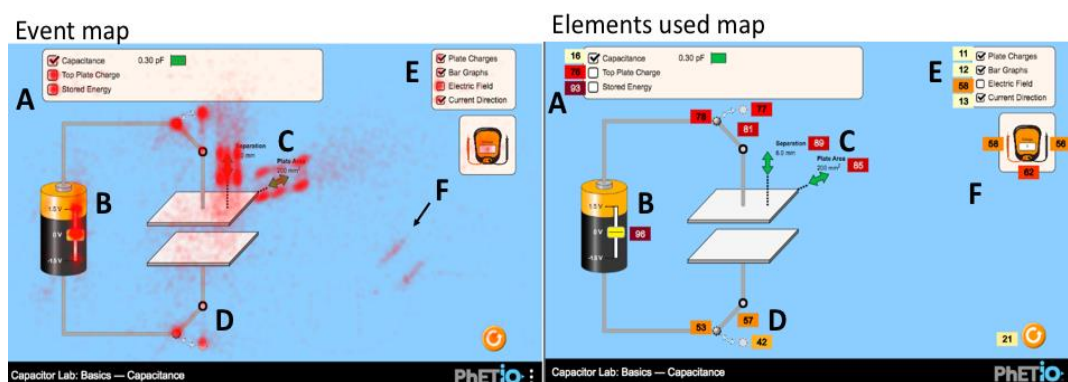


Figure 2. Visualizations in the dashboard results of a homework about stored energy. Events map shows the pattern of common events (B and C). Elements used map shows in dark red the elements more used in the activity (A, B and C), and in light colors the elements less used (E).

The events versus time graphs (not shown) provide insight into student patterns of interaction. Evidence of short pauses in interaction is observed, which is often time spent by the student to take notes, reflect, or analyze what is happening in the sim (Perez et al., 2017). Interesting patterns arise; for instance, we observed one student with over 10 minutes of very active interaction with little time between interactions (steep slope in the curve), but followed by almost 10 minutes with slowed interaction, perhaps indicative of a more planned manipulation or potentially a time moving between interaction and answering the homework (Perez et al., 2017).

4.4 Improvements and Next Steps

While the current version of the dashboard provides some information that teachers and researchers find interesting and useful (personal communications at GIREP'18, AAPT'18 and PERC'18), further work is planned. The open questions in the teacher survey are still under analysis to have a deeper understanding of teacher needs regarding data collection. The next step in the project is to do interviews with teachers to explore how they interpret the visualizations in the dashboard, what extra information they would like to have, how they can use

that data, and any information they feel is not useful. Data collection with other sims to test the visualizations are also in process.

5 CONCLUSIONS

Simulations, and the fine-grained, non-linear interaction data produced that require novel approaches to facilitate teacher visualization and understanding of data about student's interaction. In this work, we present the work in process of the design, development, and evaluation of a teacher dashboard for interactive simulations. The design leverages teacher surveys, prior work around student engagement and web usability, but a closer collaboration with teachers is needed to improve the design and analyze how teachers interpret the visualizations in the dashboard and how that data is used in classroom.

ACKNOWLEDGMENT

The author is grateful to all the teachers who responded to our surveys, and to the PhET team for their support. This work was funded by the Gordon and Betty Moore Foundation as well as the CONACYT and CICATA-IPN (Mexico).

REFERENCES

- Adams, W. K., Paulson, A., & Wieman, C. E. (2008). [What levels of guidance promote engaged exploration with interactive simulations?](#) In *AIP Conference Proceedings* (Vol. 1064, pp. 59–62).
- Atterer, R., Wnuk, M., & Schmidt, A. (2006). [Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction.](#) In *Proceedings of the 15th Int. Conf. on World Wide Web* (pp. 203–212).
- Borek, A., McLaren, B. M., Karabinos, M., & Yaron, D. (2009). [How Much Assistance is Helpful to Students in Discovery Learning?](#) In U. Cress, V. Dimitrova, & M. Specht (Eds.), *Proceedings of the Fourth European Conference on Technology Enhanced Learning, Learning in the Synergy of Multiple Disciplines (EC-TEL 2009)* (Vol. LNCS 5794, pp. 391–404). Nice, France: Springer-Verlag. Retrieved from
- Chamberlain, J. M., Lancaster, K., Parson, R., & Perkins, K. K. (2014). [How guidance affects student engagement with an interactive simulation.](#) *Chem. Educ. Res. Pract. Chem. Educ. Res. Pract.*, 15(15), 628–638.
- Corrin Linda, de B. P. (2014). [Exploring students' interpretation of feedback delivered through learning analytics dashboards.](#) In & S.-K. L. B. Hegarty, J. McDonald (Ed.), *Rhetoric and Reality: Critical perspectives on educational technology* (pp. 629–633). Proceedings ascilite Dunedin 2014. Retrieved from
- Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). [Design and Implementation of a Learning Analytics Toolkit for Teachers.](#) *Educational Technology & Society*. International Forum of Educational Technology & Society.
- Greller, W., & Drachsler, H. (2012). [Translating Learning into Numbers: A Generic Framework for Learning Analytics.](#) *Educational Technology & Society*, 15(3), 42–57. Retrieved from
- Holstein, K., McLaren, B. M., & Aleven, V. (2018). [Student Learning Benefits of a Mixed-reality Teacher Awareness Tool in AI-enhanced Classrooms.](#) *Aied*, 1–14. Retrieved from
- Holstein, K., McLaren, B. M., & Aleven, V. (2017). [Intelligent tutors as teachers' aides.](#) In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17* (pp. 257–266).
- Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017). [Modeling exploration strategies to predict student performance within a learning environment and beyond.](#) In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17* (pp. 31–40).
- Klerkx, J., Verbert, K., & Duval, E. (2017). [Learning Analytics Dashboards.](#) In *Handbook of Learning Analytics* (pp. 143–150). Belgium.

- Moore, E. B., Herzog, T. A., & Perkins, K. K. (2013). [Interactive simulations as implicit support for guided-inquiry](#). *Chem. Educ. Res. Pract.*, 14(3), 257–268.
- Navalpakkam, V., & Churchill, E. (2012). [Mouse tracking: measuring and predicting users' experience of web-based content](#). In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (pp. 2963–2972).
- Perez, S., Massey-allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., & Roll, I. (2017). [Identifying Productive Inquiry in Virtual Labs Using Sequence Mining](#). *Proceedings of Artificial Intelligence in Education*, 10331, 287–298.
- Podolefsky, N. S., Perkins, K. K., & Adams, W. K. (2010). [Factors promoting engaged exploration with computer simulations](#). *Physical Review Special Topics - Physics Education Research*, 6(2).
- Salehi, S., Keil, M., Kuo, E., & Wieman, C. E. (2015). [How to structure an unstructured activity: Generating physics rules from simulation or contrasting cases](#). In *2015 Physics Education Research Conference Proceedings* (pp. 291–294). American Association of Physics Teachers.
- Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Shirvani Boroujeni, M., Holzer, A., ... Dillenbourg, P. (2017). [Understanding learning at a glance: A systematic literature review of learning dashboards](#). *IEEE Transactions on Learning Technologies*, 10(1), 148–157. Retrieved from
- Velasco, J., & Buteler, L. (2017). [Simulaciones computacionales en la enseñanza de la física: una revisión crítica de los últimos años](#). *Enseñanza de Las Ciencias*, 35(2), 161–178.
- Verbert, K., Govaerts, S., Duval, E., Santos, J., Assche, F., Parra, G., & Klerkx, J. (2013). [Learning dashboards: an overview and future research opportunities](#). *Personal and Ubiquitous Computing*, 1–16.
- Xhakaj, F., Aleven, V., & McLaren, B. M. (2016). [How Teachers Use Data to Help Students Learn: Contextual Inquiry for the Design of a Dashboard](#) (pp. 340–354). Springer, Cham.

Improving research students writing with writing analytics

Sophie Abel

University of Technology Sydney

sophie.abel@student.uts.edu.au

ABSTRACT: High level literacy and written communication skills are essential for Higher Degree Research (HDR) students. There is increased pressure on research students to write about their research effectively and quickly while also conducting research. However, most students find writing difficult. Issues of argument, expression and organization have been reported as key problems in research students' writing. Writing Analytics (WA), is one approach that could be leveraged to help students improve their research writing. WA supports student writing practices by providing formative feedback on their writing. This feedback allows the user to reflect on what they have written and revise their writing. Therefore, the aim of my research is to integrate WA tools in research writing programs to help develop and improve research writing. The outcome of this research is a writing analytics tool that helps improve student writing and a learning design framework that integrates WA in research writing programs. My research will also document an innovative approach to teach research writing in the Australian research training context. The findings from my research will demonstrate how to better implement WA tools in research writing pedagogy to better support research students learn research writing so that they can produce quality writing.

Keywords: Learning Analytics, Writing Analytics, Research Writing, Genre, Learning Design

1 INTRODUCTION

Writing effectively is critical for research students. Effective written communication skills are not only necessary to complete the dissertation and therefore a core graduate outcome, writing effectively is also essential post dissertation. Effective written communication skills are necessary for publishing research, applying for research grants and employability, making them one of the core skills identified by employers as necessary for research graduates (McGagh et al., 2016). Research students are expected to not just conduct research, but to also write about it effectively. However, most students find writing difficult and supervisors have also reported that writing is a challenge for research students (Aitchison, Catterall, Ross, & Burgin, 2012).

Quality research writing involves more than just understanding and applying grammar rules. Quality writing involves rhetoric; understanding the audience and providing appropriate cues to facilitate understanding. Rhetorical insight into the disciplinary discourse community is necessary for creating and disseminating knowledge. However, understanding this rhetorical nature of research writing has been reported as one of the writing challenges that research students face (Paltridge & Starfield, 2007). The rhetorical complexity of the dissertation is a challenge for students (Thompson, 2016), as they are now expected to write for their discipline's discourse convention. Despite this expectation, most research students do not have the expertise in applying the discipline discourse conventions in their writing, and few students have the experience of writing for an academic audience (Torrance, Thomas, & Robinson, 1992). While there are numerous studies on undergraduate writing practices

and writing pedagogy, there is little information on the writing practices of research students or how they learn research writing. While, there is literature on research writing pedagogy, limited research exists on how it is implemented in doctoral programmes (Lee & Danby, 2012). Understanding the writing practices and the writing approaches of Higher Degree Research (HDR) students could help educators develop better research writing pedagogy, writing tools, and interventions to help students with research writing.

One approach to help students to improve their research writing could involve the use of Learning Analytics (LA), specifically, writing analytics. Writing Analytics (WA) derives from LA with an emphasis on supporting students writing practices (Buckingham Shum et al., 2016). WA measures and analyses written text through a Natural Language Processing (NLP) tool and parser. The parser can be designed to detect specific patterns or parts of a text. WA tools can provide formative feedback to students about their writing, for example, on rhetorical and structural features.

2 CURRENT KNOWLEDGE & EXISTING SOLUTIONS

Current WA tools exist in the form of Automated Writing Evaluation (AWE) tools. AWEs are used in classrooms to provide students with formative feedback. Using similar computational techniques AWEs analyse student writing and generate instant feedback on students' texts. Different AWEs apply different feedback forms, from reports to visualisations. The feedback provided aims to help students improve their writing. Students receive feedback on their text and then revise their text, encouraging the drafting and revision process of writing. These systems have been primarily employed in primary and secondary schools and undergraduate university classrooms to analyse students' essays. Examples include Criterion (Burststein, Chodorow, & Leacock, 2004) and Writing Pal (Roscoe, Allen, Weston, Crossley, & McNamara, 2014). Both systems identify the writing constructs of grammatical and mechanical errors, discourse structure, and style but also provide individual diagnostic feedback to improve the quality of writing. While these tools help students revise and think more about their writing they are designed for essays which is not appropriate to deal with the complexity of research writing, where students are required to understand the rhetorical nature of research writing and write for a discourse community that have specific writing conventions.

Few tools exist that help research students with their writing needs. One such tool is Mover (Anthony & Lashkia, 2003), a text analysis software that annotates research article introductions and abstracts. Mover analyses research introductions based on the Swales (1990) Create A Research Space (CARS) model. It has been experimented in a classroom setting to determine if Mover helps develop HDR students' research writing (Anthony & Lashkia, 2003). Their results are promising; students were able to annotate the discourse features of published research articles quicker with the help of Mover vs. doing it by hand without Mover, and students were able to analyse structural and discourse features of their own abstracts quicker with the help of Mover. However, the experiment was only conducted with six students. Another limitation of Mover is that it does not provide actionable feedback for its users. While, the tool shows students the moves they have written, it does not provide feedback on the moves that are missing nor how to achieve those moves in their writing.

One tool that does provide formative feedback on research writing is Research Writing Tutor (RWT) (Cotos, 2014). This tool has been developed specifically to help graduate students develop their

research writing skills. RWT uses NLP to compare student writing against a corpora of published research articles. Machine learning was used to train a classifier to identify the CARS moves in research article introductions from 30 disciplines. Like Mover, it also detects the CARS rhetorical moves, but RWT provides formative feedback on the rhetorical moves. For example, students are shown to what percentage their moves corresponds with research article introductions in their discipline. RWT also analyses other sections of the research article, such as the discussion and conclusion sections (Cotos, Huffman, & Link, 2015). Studies on RWT reveal promising results. One study found that students rhetorical composition improved from their first draft to last draft (Cotos, Link, & Huffman, 2017). Other studies report that students found RWTs feedback useful (Cotos & Huffman, 2013) and it made them think critically about their writing (Ramaswamy, 2012). These studies demonstrate that RWT does indeed help research students with their writing. However, RWT's corpus only contains research articles from 30 disciplines. This means that if the students' discipline is not in the corpus the tool may not be useful for them. In addition, as doctoral programmes are changing and interdisciplinary fields emerge the machine learning approach is not sustainable as new articles need to be added and trained.

Another tool that detects rhetorical moves in students writing is AcaWriter, developed by the Connected Intelligence Centre, UTS. AcaWriter detects writing patterns that signpost rhetorical moves and then highlights the move for the user (Knight, Shum, Ryan, Sándor, & Wang, 2016). AcaWriter has been used to help civil law students with essay writing (Knight et al., 2016) and assist pharmacy students with reflective writing tasks (Gibson et al., 2017). In both studies AcaWriter was used by students to analyze their written work which then provided students with feedback that prompts them to reflect on their writing and then revise it. In their studies Knight et al. (2016) and Gibson et al. (2017) found that students did reflect on their writing. However, to date AcaWriter has not been used to assist research students with their writing.

3 PROBLEM STATEMENT

Theoretically: limited literature exists on the doctoral writing and how research writing pedagogy is implemented in doctoral programmes. While there is literature on research writing pedagogy (Carter & Laurs, 2014), others argue that the implementation of research writing pedagogy is undocumented and undertheorized (Lee & Danby 2012).

Empirically: while WA tools has been used to improve high school students and undergraduates' student writing skills as seen above, limited research exists on how WA can be used to support research students writing skills and how these tools impact research students writing process and improve the quality of their writing.

Methodologically: for WA tools to be successful in developing research students writing skills more information is needed on how to implement these tools in the classroom and online. While, WA tools are implemented in classrooms and online, how they are implemented, their learning designs and how they are evaluated are rarely mentioned. There are few learning design frameworks or models that implement WA tools within a course. An approach to develop, implement and evaluate a writing analytic tool and intervention is through Design Based Research (DBR), this approach will be adopted to implement and evaluate the writing tool and intervention.

4 RESEARCH GOALS AND QUESTIONS

My research aims to develop an intervention and learning design that embeds AcaWriter in the teaching and learning of research writing. A specific focus will be upon the introductory section and abstracts of research articles. I will then investigate how AcaWriter impacts students' writing process and the effectiveness of the tool and intervention. The learning design and writing analytic tool will be evaluated to determine the effectiveness of the approaches developed. I will direct my study by considering the following **research questions**:

1. How do HDR students learn research writing and what are their research writing experiences?
 - a) What deficits or barriers do HDR students face in their writing?
2. What impact does the writing analytic tool have on students' writing process?
 - a) How does the writing analytic tool's feedback help students improve their understanding of rhetorical moves?
 - b) To what extent does the writing analytic tool help students improve their writing?

5 CONTRIBUTION: WRITING ANALYTICS A NOVEL APPROACH

Theoretically, my research provides a deeper understanding on the writing challenges faced by research students, how students currently learn research writing and how best to support them. *Empirically*, I am using this deeper understanding of students' challenges and approaches to learning research writing and incorporating this knowledge in developing a writing analytic tool that is specifically designed to support research students writing. The WA tool takes a rules based approach where new rules can added and created without training a large corpus of text. The tool allows research students to submit their writing for feedback so that they then can reflect and revise their writing. *Methodologically*, my research documents how to apply DBR in the implementation and evaluation of WA tools. In addition, to ensure that WA tools are used effectively to assist and develop students writing a learning design framework is being developed to effectively embed such tools in research writing pedagogy.

6 METHODOLOGY, CURRENT STATUS & RESULTS

As my research aims to improve the teaching and learning of research writing using writing analytic tools, I will adopt DBR to investigate, implement and evaluate the learning design framework and intervention strategy. DBR strives to enhance "the impact, transfer, and translation of education research into improved practice" and "stresses the need for theory building and the development of design principles that guide, inform and improve both practice and research in educational contexts" (Anderson & Shattuck, 2012, p.16). I will follow DBR's four phases and apply a mixed method approach using both quantitative and qualitative research methods:

- *Phase one: Analysis of practical problems by researchers and practitioners*

A literature review has been conducted to identify and explore the educational problem. I have also interviewed supervisors and students to gain understanding of the problem from their perspective. An online survey was administered to research students to gain insight on how they learn research writing and their approaches and perceptions to research writing. Preliminary data analysis shows that students use a variety of resources to learn research writing and some wanted more writing support.

- *Phase two: Development of solutions informed by existing design principles and technological innovations*

The writing analytic tool has been developed and the learning design of the intervention was created to fit the research student context. A genre-based pedagogical approach was taken to develop the tool and the intervention see (Abel, Kitto, Knight, & Buckingham Shum, 2018) for more information. The writing analytic tool AcaWriter was extended to include a parser that analyses research articles and introductions using the Create a Research Space (CARS) model developed by Swales (1990). The AcaWriter CARS parser highlights the CARS rhetorical moves see appendix figure 1. Feedback was also designed to align with the CARS model see appendix figure 2. The learning design was designed to help students understand, identify and apply rhetorical moves in their writing. The intervention consisted of two sessions where the first session introduced students to CARS and rhetorical moves, while the second session focused on applying the rhetorical moves learned to their own writing (see appendix for the learning design pattern and sequence of learning activities).

- *Phase three: Iterative cycles of testing and refinement of solutions in practice*

The first iteration of the intervention was conducted with 12 participants. The intervention was evaluated via an online survey, a focus group and interviews. Results from the online survey reveal that students found the intervention useful, they learned new skills and knowledge, and that they felt confident they could apply the new skills learned in their own writing. The focus group and interview data showed that all students found the highlighting and automated feedback messages useful. The automated feedback helped them think about structure when writing and focus on rhetorical moves. But, some students reported that they needed more time to become familiar with tool and the CARS model.

7 NEXT STEPS

The work presented here is the first iteration of the AcaWriter CARS parser. While, the results from the first iteration of testing are promising and show that AcaWriter has the potential to help develop HDR students' research writing skills, more iterations need to be conducted to determine the effectiveness of AcaWriter and how it impacts students' writing process and if the quality of students' texts improve. Additional parsers will be developed for other sections of the research article and an online course will be developed with AcaWriter embedded. For more information on the development of AcaWriter head to <http://heta.io/> and here <http://acawriter-demo.utscic.edu.au/> to demo the tool.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Kirsty Kitto, Simon Knight and Simon Buckingham Shum for their support and advice in this research.

REFERENCES

Abel, S., Kitto, K., Knight, S., & Buckingham Shum, S. (2018). Designing personalised, automated feedback to develop students' research writing skills. In *Proceedings ASCILITE2018: 35th*

International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education. Melbourne.

- Aitchison, C., Catterall, J., Ross, P., & Burgin, S. (2012). 'Tough love and tears': learning doctoral writing in the sciences. *Higher Education Research & Development*, 31(4), 435–447.
- Anderson, T., & Shattuck, J. (2012). Design-Based Research: A Decade of Progress in Education Research? *Educational Researcher*, 41(1), 16–25.
- Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication*, 46(3), 185–193. <https://doi.org/10.1109/TPC.2003.816789>
- Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics (pp. 481–483). ACM Press.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *Ai Magazine*, 25(3), 27.
- Carter, S., & Laurs, D. (Eds.). (2014). *Developing generic support for doctoral students: practice and pedagogy*. London ; New York, NY: Routledge.
- Cotos, E. (2014). *Genre-Based Automated Writing Evaluation for L2 Research Writing*. London: Palgrave Macmillan UK. <https://doi.org/10.1057/9781137333377>
- Cotos, E., & Huffman, S. (2013). Learner Fit in Scaling Up Automated Writing Evaluation: *International Journal of Computer-Assisted Language Learning and Teaching*, 3(3), 77–98.
- Cotos, E., Huffman, S., & Link, S. (2015). Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes*, 19, 52–72. <https://doi.org/10.1016/j.jeap.2015.05.004>
- Cotos, E., Link, S., & Huffman, S. R. (2017). Effects of DDL technology on genre learning. *Language Learning & Technology*, 21(3), 104–130.

- Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback (pp. 153–162). ACM Press. <https://doi.org/10.1145/3027385.3027436>
- Knight, S., Shum, S. B., Ryan, P., Sándor, Á., & Wang, X. (2016). Designing Academic Writing Analytics for Civil Law Student Self-Assessment. *International Journal of Artificial Intelligence in Education*, 1–28. <https://doi.org/10.1007/s40593-016-0121-0>
- Lee, A., & Danby, S. (Eds.). (2012). *Reshaping doctoral education: changing approaches and pedagogies*. Milton Park, Abingdon, Oxon ; New York: Routledge.
- McGagh, J., Marsh, H., Western, M., Thomas, P., Hastings, A., Mihailova, M., ... Australian Council of Learned Academies (ACOLA). (2016). *Review of Australia's research training system: final report*.
- Paltridge, B., & Starfield, S. (2007). *Thesis and Dissertation Writing in a Second Language: A Handbook for Supervisors*. Abingdon, Oxon: Routledge.
- Ramaswamy, N. (2012). *Online tutor for research writing*. IOWA State University. Retrieved from <http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=3750&context=etd>
- Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition*, 34, 39–59. <https://doi.org/10.1016/j.compcom.2014.09.002>
- Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.
- Thompson, P. (2016). Genre approaches to theses and dissertations. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 378–391). London: Routledge.
- Torrance, M., Thomas, G. V., & Robinson, E. J. (1992). The writing experiences of social science research students. *Studies in Higher Education*, 17(2), 155–167. <https://doi.org/10.1080/03075079212331382637>

8 APPENDIX

Analytical Report	Feedback	Resources
Move 1: Establishing a research territory		
<p>E Emphasis of a significant or an important idea</p> <p>B Background information and reviewing previous work</p>		
Move 2: Establishing a Niche		
<p>C Contrasting idea, tension, disagreement or critical insight</p> <p>Q Question or gap in previous knowledge</p>		
Move 3: Occupying the Niche		
<p>N Novelty and value of your research</p> <p>S Summary of the author's goal or nature of the research, or structure of the paper</p>		

E B ABSTRACT:

It is now widely accepted that timely, actionable feedback is essential for effective learning. In response to this, data science is now impacting the education sector, with a growing number of commercial products and research prototypes providing "learning dashboards", aiming to provide real time progress indicators. **E C** From a human-centred computing perspective, the end-user's interpretation of these visualisations is a critical challenge to design for, with empirical evidence already showing that 'usable' visualisations are not necessarily effective from a learning perspective. Since an educator's interpretation of visualised data is essentially the construction of a narrative about student progress, we draw on the growing body of work on Data Storytelling (DS) as the inspiration for a set of enhancements that could be applied to data visualisations to improve their communicative power. **S** We present a pilot study that explores the effectiveness of these DS elements based on educators' responses to paper prototypes. **S** The dual purpose is understanding

It looks like you are missing Move 1 – Establishing a research territory (E or B sentences). Here you should show how your research topic is relevant and important by introducing & reviewing previous research on your topic. For example, recent research indicates that the effects of climate change have.... (for more examples head to the resources tab)

Figure 1: AcaWriter CARS

Figure 2: AcaWriter CARS feedback

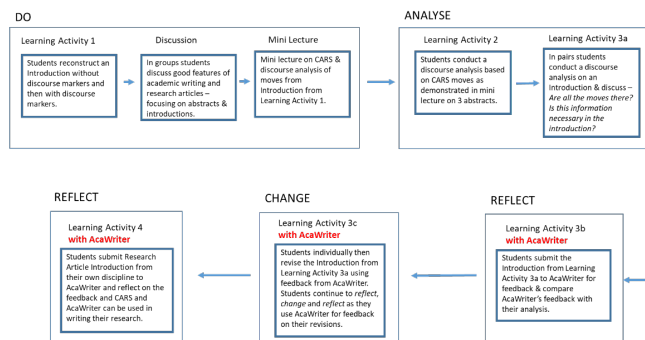


Figure 3: Session 1 learning design pattern

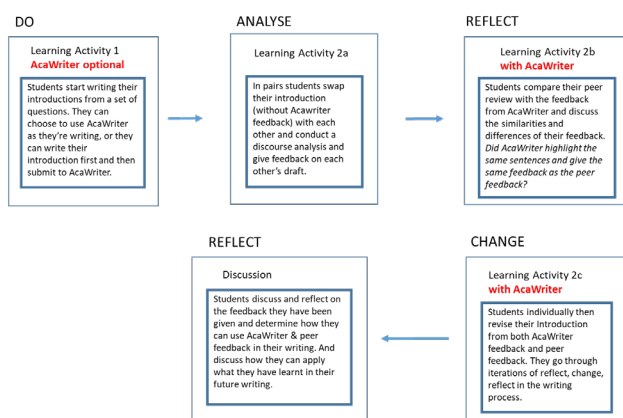


Figure 3: Session 2 learning design pattern

Trace Data: How to Improve a Method to Measure Self-regulated Learning in Online Courses

Heeryung Choi
University of Michigan
heeryung@umich.edu

ABSTRACT: As a self-regulated learning (SRL) process is critical for learners to become motivated, independent goal-achievers, various ways to measure SRL process have been developed. While aptitude-approach methods such as self-reported survey have been most frequently used, there have been recent criticisms of aptitude methods due to their theoretical detachment and potential biases. Therefore, there are researchers who have carefully refined and applied trace measures, which is a type of event-approach methods to capture SRL process through more accurate and richer data. This doctoral consortium paper proposes a method measuring SRL process through trace data which particularly focuses on SRL process in an online programming course.

Keywords: Self-regulated learning, trace data, Computer-based Learning Environments (CBLEs)

1 INTRODUCTION

As the amount of data on learning increases with the growing of usage of computer-based learning environments (CBLEs), there have been discussions on how to utilize a large dataset from CBLEs in learning analytics. In particular, CBLEs are claimed to offer unprecedented opportunities to expand understanding of self-regulated learning (SRL) (Azevedo & Aleven, 2013; Greene & Azevedo, 2010; Winne, 2010). SRL is the process by which a learner monitors and controls metacognition, cognition, motivation, affect, and contextual factors to achieve goals (Boekaerts & Niemivirta, 2000; Greene & Schunk, 2017; Pintrich, 2000; Winne, 2001; Zimmerman, 2000). One major efforts to measure SRL in CBLEs is through self-report measures. While self-report measures have been the most frequently used method in SRL studies, recent criticism points out that self-report measures are theoretically detached from SRL models and are not free from questions of bias and inaccuracy of data produced (Winne, 2010). While a trace measure has been suggested as another powerful measure to study SRL (Azevedo et al., 2013; Winne & Perry, 2000), the trace measure for studying SRL require in-depth understanding of SRL models and cautious approach to capturing useful and accurate data. In my dissertation, I will develop methods for measuring SRL processes through trace data. In particular, as the first study of my dissertation, I will focus on systems for studying how to use metacognitive prompts in order to help learners make better use of hints which is described more fully in this paper.

2 RELATED WORK

2.1 Self-regulated Learning Theories

Self-regulated learning (SRL) theories explain how learners activate and sustain their cognition, metacognition, motivation, affect, and contextual factors to achieve their learning goals (Greene & Schunk, 2017; Zimmerman, 2000). The SRL models depict the SRL process as a recursive cycle of phases: preparatory phase, performance phase, and evaluation phase (Boekaerts & Niemivirta, 2000; Panadero, 2017; Pintrich, 2000; Winne, 2001; Zimmerman, 2000). The preparatory phase is the first phase where learners identify tasks and goals. The performance phase is where learners work on a given task based on their identification of contexts and goals from the previous phase. The evaluation phase is where learners assess their operations and decide what to keep and what to change in the next cycle of phases. That is, this evaluation from the current cycle will influence how learners monitor and control contexts and execute operations.

The cycle of phases in SRL models is composed of sets of contexts and operations. As Winne (2010) states, during SRL processes learners identify and understand their internal contexts (e.g., self-efficacy) and external contexts (e.g., task difficulty, availability of peers for help-seeking, and noise in classroom), and metacognitively think and decide how they are going to work on a task to manage the given contexts. Because each operation is a reaction to a given context, it is important to understand and analyze operation with the corresponding context. Thus, a measure of SRL should be able to capture both context and operation.

2.2 Problems of the Aptitude-approach SRL Measures

2.2.1 *Aptitude approach and event approach*

How researchers design and apply a measure of SRL reflects how the researchers perceive SRL theories. Researchers should carefully align measures they are going to use and SRL models for accurate data collection (Greene & Azevedo, 2010; Winne, 2010; Winne & Perry, 2000). Through discussions on how to measure SRL components such as metacognition and motivation, Winne and Perry (2000) conceptualize SRL processes as aptitudes and as events.

An aptitude is an interpretation of a set of SRL events from a person's view point (Winne, 2014; Winne & Perry, 2000). After an aggregation of events is generalized over time and through an individual's belief and personality attributes, this aggregation of events is shaped as an aptitude. Therefore, an instrument measuring aptitude often includes ratings such as "most of the time" or "typically" to earn aggregated responses on SRL across different contexts (Winne, 2010; Zimmerman, 2008). To get such aggregation as responses, aptitude-approach measurements often collect learners' *interpretation* of SRL instead of referring to one event.

An event is "a snapshot that freezes activity in motion (p. 534)" (Winne & Perry, 2000), which has an obvious beginning and ending and therefore can be easily separated from a prior event and a subsequent event. A size of an event could be differently defined from every particular time span (e.g. every 5 seconds) to a distinguishing context. Since an event is a record of an actual action, an event is not a description of an action or a consequential mental state after an action (Winne, 2014; Winne & Perry, 2000).

2.2.2 *Issues in self-report measures*

Self-report measures such as surveys have been most frequently adopted as SRL measures. A data format of self-report measures is inevitably an interpretation of SRL activity because the data format is a result of a sample of multiple data points chosen by each learner, not an objective observation. Learners subjectively interpret, or sample as Winne (2010) said, their various past behaviors and thoughts to give a response which meets the restraint format for responses. For example, learners are asked to choose 1 to 7 out of 7 Likert scales or to answer in a few sentences instead of showing their entire activities during learning – which is impossible considering the constraint on memory.

Concerning that responses are sampled as described above, Winne (2010) questions how reliable self-report methods are. This is because learners do not go through a statistically valid sampling process to come up with responses representing their entire behavior and cognition. Furthermore, data generated through self-report methods can be biased and might not reflect an actual behavior and cognition of learners. Learners might respond not based on the actual action but based on their knowledge about which action is recommended for effective learning (Pintrich, 2000). These issues of reliability are inevitable for aptitude measures.

Furthermore, Researchers have pointed out the fundamental distance between aptitude-approach measure and SRL theories. SRL theories are based on the belief that learners' operations such as monitoring and control keep changing dynamically between and during tasks in response to internal and external contextual factors such as self-efficacy and task environments (Greene & Azevedo, 2010; Zimmerman, 2008). Since each SRL action is made with respect to a certain context, it is crucial to report both context and subsequent reaction and to distinguish a set of context and reaction from other sets in order to understand SRL processes of learners with higher accuracy. Considering that, aptitude measure, which is an aggregation of multiple contexts and operations, is inevitably inaccurate in measuring SRL because of its assumption that SRL is static and can be measured operations without corresponding contexts.

2.3 **Trace Data as Event Methods**

Trace is data that a learner produces concurrently with the cognitive operations that the learner adopts to process information in working memory. For example, Winne (2010) suggests learners' highlights of words in a text as an example of a trace. In CBLEs examples of trace data are log data, clickstream data, and eye-tracking data generated by learners interacting with learning materials. As a definition of trace data suggests, trace data are event-approach measures. They are event-approach since each operation generated a piece of trace data without going through learners or researchers' interpretation; there is no aggregation, sampling from learners' end, or interpretation.

Because trace data measure SRL as a sequence of events, this measurement holds multiple advantages. Firstly, treating SRL as a sequence of events allows researchers to capture and model the dynamic nature of these processes as they are continually deployed and adjusted during learning as students attempt to regulate various aspects of the context including internal cognitive conditions (e.g., prior knowledge) and external conditions (e.g., affordances of the CBLE such as access to help-seeking features to facilitate problem-solving). Secondly, trace data are highly accurate and precise because they are based on the use of objective data collection methods which do not ask students

for their perceptions regarding their ability to regulate their cognitive and metacognitive processes (Winne, 2010).

Yet, trace data generation process is not completely free from a potential possibility of generating inaccurate data. This is because tracing cognitive events sometimes requires learners' additional action which learners do not usually make during learning and have to be trained to perform. This "unnatural (p. 272)" trace generation can be either barely adopted by learners or affect learning experiences by working as interventions (Winne, 2010). For example, if researchers ask learners to draw a flowchart of their thought while learning, it might be considered as another task with an extra cognitive load on learners who are unfamiliar with a flowchart could. In this case, researchers cannot measure the usual learning experience of learners. Even weblogs can be intervening if a researcher designs a system requiring learners to do additional, unnatural behavior in order to acquire certain type of logs as evidence of a certain SRL process. Regarding this issue effect of trace-generating process, Winne (2010) claims that every measurement inevitably intervenes learning, and therefore, issues are not how to create a non-intervening measurement but how to build an SRL measure that (a) aligns well with SRL theory, (b) produces data that can verify researchers' interpretation on observed events, and (c) additionally helps learners' learning experiences. With these considerations, unnatural intervention could become a part of learning more easily as highlighting terms in a text.

There have been few studies on analyzing, building, and evaluating trace measure for SRL (Azevedo et al., 2013; Winne & Hadwin, 2013; Winne, Nesbit, & Popowich, 2017). Azevedo et al. (2013) introduced MetaTutor system which was built to collect multi-channel data while supporting students. Winne et al. (2017) have built and evaluated nStudy platform collecting data of note taking and highlighting words. While these previous works can be adopted across subjects, these works are not designed for domain-specific contexts and operations. In this work, I propose a platform and data analysis framework both for learners and researchers focusing on online computational data science education.

3 METHODOLOGY

Prompting has been suggested as a way to support SRL process especially in CBLEs. In this study, metacognitive prompts, which support learners' monitoring and control of information processing, are used in this study design to encourage learners effective usage of information from hints.

3.1 Participants

Participants will be recruited from a Coursera course on data analysis in Python taught by (Removed for the blind review). The number of participants is expected to be two thousand.

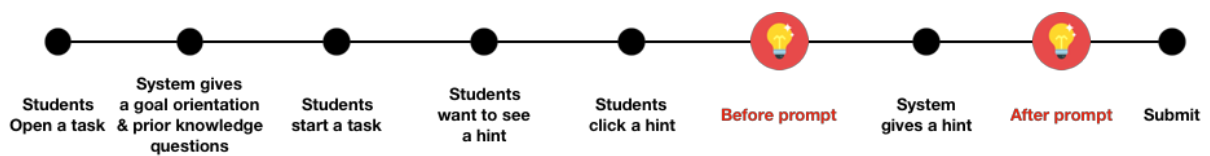
3.2 Learning Task

There will be three types of pre-prompts to be shown before a hint, and the other three types of post-prompts to be shown after a hint. Therefore, there will be nine prompt conditions (i.e., 3 pre-prompts X 3 post-prompts) (Table 1). Each learner will be randomly assigned to one prompt condition. In this experiment (Figure 1), learners will be asked to finish a Python coding task on Jupyter Notebook. When learners open a Jupyter Notebook, they will be assigned randomly to one

of nine prompt conditions, and they will see mandatory survey questions respectively on a goal orientation and a prior knowledge. After that, learners will proceed to solve the given question set composed of multiple questions such as sorting a data frame or finding the ten highest values. Each coding question will have a hint button and whenever learners are looking for a hint for the question during the task, they can click the hint button to get the hint. When learners click the hint button, they will see one of three different pre-prompts based on their condition. After they textually respond to the given prompt, they will receive a hint. When they click "Continue" button on the pop-up windows with the hint, learners will see one of three different post-prompts based on their condition. Once learners give textual responses to the prompt they are assigned, learners will be able to proceed to try to solve their problem sets with the given hint. Log data will be recorded during learners' attempts to solve the problem sets to trace the following learners' behaviors which will be considered as dependent variables: how long learners spend reading the hint, how long it takes for learners to solve the task, whether learners use other materials like videos after receiving hints, whether learners use ideas from the hint in writing their code in the Jupyter notebook, whether learners do better on subsequent problems.

Table 1: Pre-prompts and Post-prompts with examples

Timing	Prompt Type	Definition	Example
Pre-prompt	Reflection	Reflection on current knowledge/confusion type	What are you confused about right now?
	Planning	Explanation on plan to solve problems and what they expect from hints	What is the step you need to take next?
	None	No prompt	
Post-prompt	Reflection	Reflection on what learners learn from hints	Does anything in this hint conflict with the way you understood the problem? If yes, what is it?
	Planning	Planning about of use of hint type	How did the hint lead you to rethink your initial plans to solve the task?
	None	No prompt	Cell Value

**Figure 1: Overall study design**

4 CURRENT STATUS OF THE WORK

A Jupyter notebook extension giving prompts and hints has been built. Prompts, problem sets, and hints for problem sets are currently being refined. Discussion on how to interpret log data is ongoing.

REFERENCES

- Azevedo, R., & Aleven, V. (2013). Metacognition and Learning Technologies: An Overview of Current Interdisciplinary Research. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 1–16). New York, NY: Springer New York.
- Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F., & Landis, R. (2013). Using Trace Data to Examine the Complex Roles of Cognitive, Metacognitive, and Emotional Self-Regulatory Processes During Learning with Multi-agent Systems. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 427–449). New York, NY: Springer New York.
- Boekaerts, M., & Niemivirta, M. (2000). Chapter 13 - Self-Regulated Learning: Finding a Balance between Learning Goals and Ego-Protective Goals. In *Handbook of Self-Regulation* (pp. 417–450). San Diego: Academic Press.
- Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, 45(4), 203–209.
- Greene, J. A., & Schunk, D. H. (2017). Historical, Contemporary, and Future Perspectives on Self-Regulated Learning and Performance. In *Handbook of Self-Regulation of Learning and Performance* (pp. 17–32). Routledge.
- Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology*, 8, 422.
- Pintrich, P. R. (2000). Chapter 14 – The Role of Goal Orientation in Self-Regulated Learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 451–502). San Diego: Academic Press.
- Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, 2, 153–189.
- Winne, P. H. (2010). Improving Measurements of Self-Regulated Learning. *Educational Psychologist*, 45(4), 267–276.
- Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, 9(2), 229–237.
- Winne, P. H., & Hadwin, A. F. (2013). nStudy: Tracing and Supporting Self-Regulated Learning in the Internet. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 293–308). New York, NY: Springer New York.
- Winne, P. H., Nesbit, J. C., & Popowich, F. (2017). nStudy: A System for Researching Information Problem Solving. *Technology, Knowledge and Learning*, 22(3), 369–376.
- Winne, P. H., & Perry, N. E. (2000). Chapter 16 - Measuring Self-Regulated Learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 531–566). San Diego: Academic Press.
- Zimmerman, B. J. (2000). Chapter 2 – Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of Self-Regulation* (pp. 13–39). Elsevier.
- Zimmerman, B. J. (2008). Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal*, 45(1), 166–183.

***IntVisRep*: An Interactive Social Learning Analytics Tool**

Fan Ouyang
Zhejiang University
fanouyang@zju.edu.cn

ABSTRACT: This **demo** paper devised an interactive social learning analytics tool named *IntVisRep*, to demonstrate three representations of online discussion data: interaction networks, keyword flows, and temporal online engagements. This tool *IntVisRep* aimed to help learners become aware of their interaction, discourse, and cognition processes, and further adjust their interaction, participation, and collaboration accordingly during online learning processes.

Keywords: Online discussions; Learning analytics; Student-facing learning analytics tools

1 BACKGROUND

Inspired by the social perspective of learning (Vygotsky, 1978), researchers design and implement learning analytics tools, representations, and dashboard systems to demonstrate interactive, dialogic, collaborative aspects of online learning. These learning analytics tools usually aim to increase student awareness, improve engagement, and facilitate social behaviors in online learning processes (Bodily & Verbert, 2017). However, most existing tools merely provided relation-related information such as social interaction information (e.g., network visualizations, centrality metrics), or behavior-related information (e.g., time spent online, number of messages); they did not provide students with a holistic picture of learning processes which may increase social comparisons, and discourage further student engagement (Ouyang & Chang, 2018). To provide students with richer information, a student-facing learning analytics tool named *CanvasNet* was devised to offer both social interaction networks and conceptual lexical information in order to increase student social and conceptual engagement; yet results showed the use of *CanvasNet* did not have significant effects on increasing students' social and cognitive engagement (Chen, Chang, Ouyang, & Zhou, 2018). Given the complication implied by previous studies, I devised an interactive social learning analytics tool named *IntVisRep* to demonstrate three types of interactive, visualized representations: interaction networks, keyword flows, and temporal online engagements. The goal of this student-facing tool is to provide students with a holistic picture of their learning processes and help students become aware of their interaction, discourse and cognition. I hope in the future the use of *IntVisRep* can help facilitate student social, interactive, collaborative aspects of online learning.

2 INTRODUCTION of *IntVisRep*

Together, *IntVisRep* demonstrated three representations: interaction networks, keyword flows, and online engagement changes (see Figure 1). First, interaction networks demonstrated students' interactions (i.e., replies and comments). Second, keyword flows demonstrated sequential relations of frequently-used keywords (i.e., one word followed by the other). Third, the temporal representation demonstrated changes of participants' social, cognitive, facilitative engagement over time. Specifically, social interaction network - showing the structure of interaction network and participant position - was designed to help participants become aware of the individual and class interaction processes. The

keyword flow representation - showing the sequential relations between concepts or ideas - was designed to help students become more aware of their discourse in inquiry. The online engagement representation - showing social, cognitive, facilitative contributions from a temporal perspective - was designed to help students become aware of their engagement dimensions during different time points in discussions.

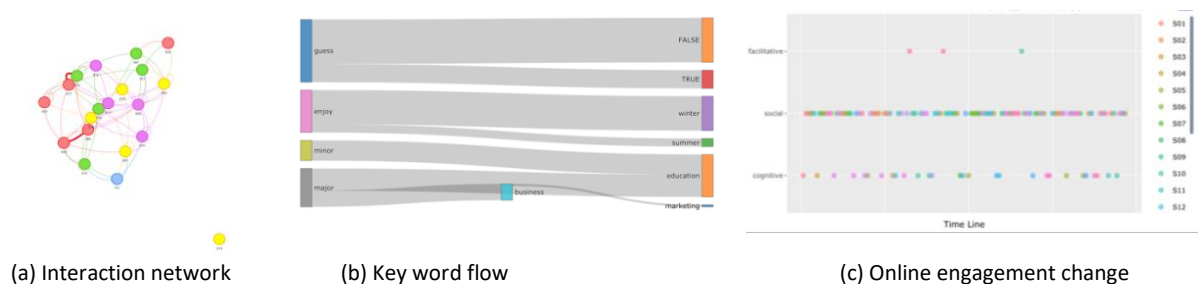


Figure 1: Three representations of *IntVisRep*

3 USE of *IntVisRep*

The data used in the demo originated from an authentic online, undergraduate-level course titled “*Foundations of Computer Applications for Business and Education*”, offered at a midwestern university in US. The demo of *IntVisRep* was deployed through a [ShinyApp](#); a brief [video](#) introduced the use of this tool; and selected data, descriptions, and relevant codes can be accessed through my [Github repository](#).

4 FUTURE DESIGN AND RESEARCH GOALS

Student-facing learning analytics tools, representations and reports have potentials to aid information navigation, sense-making, and decision-making. Since the initial version of *IntVisRep* included some intensive post-analyses which can only provide delayed information, in the future design, I will use Canvas or Moodle API to capture real-time data from learners and generate interactive representations directly. Moreover, since previous studies indicated a complication of the effect of a social learning analytics tool on student learning (e.g., Chen et al., 2018), I will further examine whether and how the use of *IntVisRep* would influence students’ social interaction, topic contribution, and online engagement.

ACKNOWLEDGEMENT

I appreciate the assistance from Yuhang Li and Yixuan Chen during data cleaning process.

REFERENCES

- Bodily, R., & Verbert, K. (2017). Trends and issues in student-facing learning analytics reporting systems research. *Proceedings of LAK 2017*, pp. 309-318. ACM.
- Chen, B., Chang, Y. H., Ouyang, F., & Zhou, W. Y. (2018). Fostering discussion engagement through social learning analytics. *The Internet and Higher Education*, 37, 21–30.
- Ouyang, F. & Chang, Y. H. (2018). The relationship between social participatory role and cognitive engagement level in online discussions. *British Journal of Educational Technology*. [Online Version of Record]
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

A stage-based matrix model of student progression

Amelia Brennan

RMIT Studios, RMIT University,
amelia.brennan@rmit.edu.au

Pablo Munguia

RMIT Studios, RMIT University
pablo.munguia@rmit.edu.au

ABSTRACT: No course exists in isolation, so examining student progression through courses within a broader program context is an important step in integrating course-level and program-level analytics. Integration in this manner allows us to see the impact of course-level changes to the program, as well as identify points in the program structure where course interventions are most important. This poster highlights the significance of program-level learning analytics, where the relationships between courses become clear, and the impact of early-stage courses on program outcomes such as graduation or drop-out can be understood. We present a matrix model of student progression through a program as a tool to gain valuable insight into program continuity and design. We demonstrate its use in a real program, and examine the impact upon progression and graduation rate if course-level changes were made early on. We also extend the model to more complex scenarios such as multiple program pathways and simultaneous courses. Importantly, this model also allows for integration with course-level models of student performance.

Keywords: Program analytics, matrix model, program pathways.

1 THE MATRIX MODEL

The progression of students through a program can be modeled as a series of stages, where at the end of each stage some proportion of the cohort will progress, others will drop out, and some will repeat the stage (fig. 1). This flow can be represented as a Lefkovitch matrix (Lefkovitch, 1965; Caswell, 2001), where the populations of each year (the N_i) at each timestep are found through matrix multiplication:

$$\begin{bmatrix} a & 0 & 0 & 0 & 0 \\ 1 & r_1 & 0 & 0 & 0 \\ 0 & p_1 & r_2 & 0 & 0 \\ 0 & 0 & p_2 & r_3 & 0 \\ 0 & 0 & 0 & g & 0 \end{bmatrix} \begin{bmatrix} N_{in} \\ N_1 \\ N_2 \\ N_3 \\ N_g \end{bmatrix}_t = \begin{bmatrix} N_{in} \\ N_1 \\ N_2 \\ N_3 \\ N_g \end{bmatrix}_{t+1} \quad (1)$$

Options such as taking multiple courses simultaneously, choosing between available pathways, and studying part-time, have also been built into a more complex formulation of this model to represent more realistic student behaviours. Course-level models of student outcomes abound in the literature (e.g. Nghe et al., 2007), and can be integrated by building functions of, say, student ability, engagement, teacher capability and course design, that predict rates of p_i and r_i .

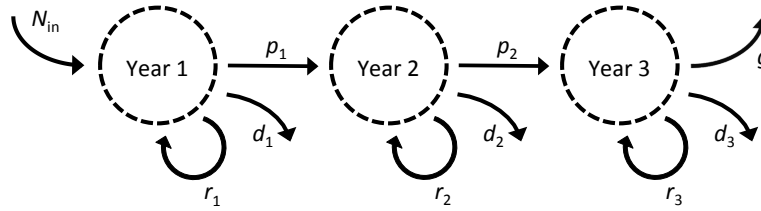


Figure 1: Progression of students through a three-year program. At the end of each year, a student can either progress (or graduate), drop out, or repeat the year, with probabilities $p_i(g)$, d_i and r_i respectively. The incoming cohort is N_{in} .

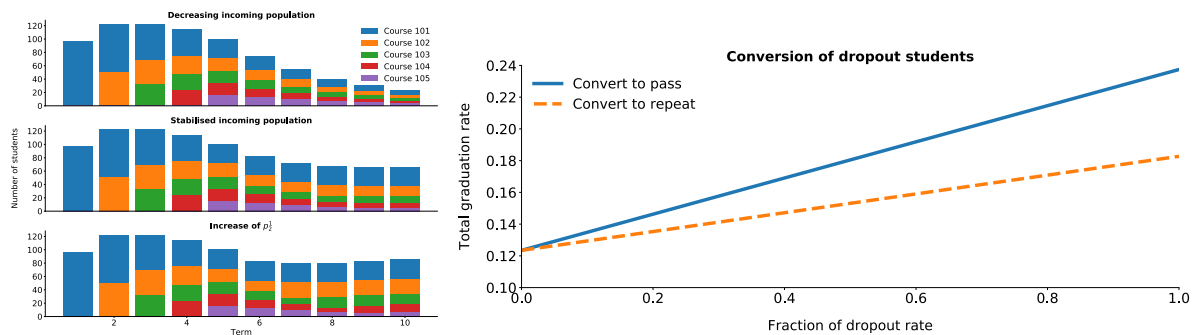


Figure 2: The impact of potential changes to matrix values such as the year 1 progression rate can be modeled, showing the effect on program populations and their distributions across courses (taken consecutively) over time (left). If students who dropped out were instead somehow enabled to either pass the subject and progress (solid) or repeat the subject (dashed), the resulting change to the overall graduation rate can be projected (right).

Thus, models of course-level outcome can be situated within the broader program context, and by measuring, modelling and reporting the values within the matrix, the impact of course variations on the overall program flow can be identified (fig. 2). Program managers are able to understand how their cohorts are increasing or decreasing, where to direct resources to have the greatest impact, and where bottlenecks are located in a program, allowing for better planning. Course coordinators are able to see where their course sits within a program, where their students are coming from, and where they are likely to go. To see how the pathways taken by successful students differ from those taken by students who perform poorly, the progression rates of a program matrix could be recalculated following a filtering of the input data to particular groups of students. Further understanding the choices behind these alternative pathways would then assist in developing recommended pathways or other options for the latter group.

REFERENCES

- Caswell, H. (2001). *Matrix population models*. Wiley Online Library.
- Lefkovich, L.P. (1965). *The study of population growth in organisms grouped by stages*. Biometrics, 1-18.
- Nghe, N.T. et al. (2007). *A comparative analysis of techniques for predicting academic performance*. Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual (pp. T2G-7). IEEE

Learning analytics adoption – approaches and maturity

Yi-Shan Tsai

University of Edinburgh
yi-shan.tsai@ed.ac.uk

Vitomir Kovanović

University of South Australia
Vitomir.Kovanovic@unisa.edu.au

Dragan Gašević

Monash University
Dragan.Gasevic@monash.edu

ABSTRACT: The poster presents the change of prioritised approaches to learning analytics (LA) among higher education as their experience of adoption increases. The study examined 27 UK and European higher education institutions using the Epistemic Network Analysis technique. Results show that institutions with one or more years of experience with LA put more emphasis on understanding learning or teaching phenomena, whereas institutions with less experience of LA focused more on measuring the phenomena. This implicates a change of conceptualisation among institutions as their experience with LA increases.

Keywords: Learning analytics, higher education, adoption, strategy, approach

1 INTRODUCTION

A strategic vision that responds to the needs of an organisation is critical for long-term impact and the development of institutional capability for LA. While existing studies of LA adoption have shed light on policies and strategies targeted at institutional or national level of implementation (Colvin, Dawson, Wade, & Gašević, 2017), there is limited understanding of the change of priorities when institutions' experience with LA increases. The current study seeks to bridge the gap and highlight the need for a strategy that evolves based on evaluations of short-term objectives for LA (Kotter, 2006). This work explores an overarching question: *what is the state of adoption among UK and European HEIs in terms of learning analytics?* The poster focuses on identifying the prioritised goals and approaches to LA.

2 METHODOLOGY

To answer the research question, we carried out 29 semi-structured interviews with institutional leaders from 27 HEIs. The interview data was first transcribed and coded before subsequently analysed using the Epistemic Network Analysis (ENA) technique (Shaffer et al., 2009). ENA works by examining the co-occurrence of *codes* (representing concepts) within a set of *stanzas*, which are text excerpts (e.g., conversation utterances) where co-occurrence represents a meaningful relationship for each of the *units of analysis* (e.g., institutions). For this poster, we present the interwoven networks of eight codes under two themes – goals (institutional, teaching, and learning levels) and approaches (measuring, exploratory, data-led, problem-led, and experimental). The institutions were put in two groups by adoption experience: less than one year of experience (n=9) and one or more years of experience (n=18). One year was chosen as a threshold due to the fact that only two institutions had adopted LA for more than 3 years.

3 RESULTS

To understand what institutional adoption looks like when the experience of LA increases, we plotted two mean ENA networks of institutions by their experience of adopting LA (Figure 1). The X-axis corresponds to the first singular value and explains 9.1% of the variability in the study subjects' networks, while the Y-axis, corresponds to the second singular value that explains additional 15.0% of the variability in subjects' networks. The thickness of the lines between nodes represents the frequency of their co-occurrence across stanzas, which indicates the strength of connections.

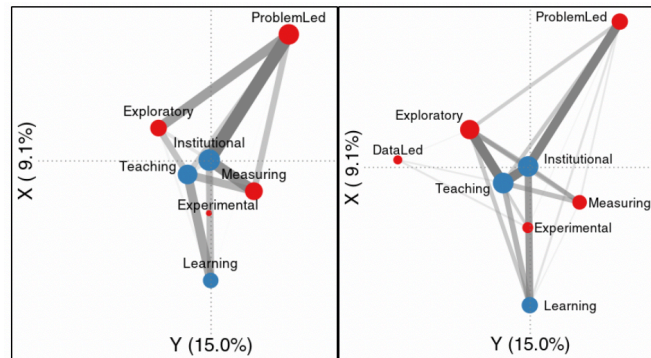


Figure 1. Mean ENA networks for novice institutions (left) and experienced institutions (right)

The results showed both groups having strong connections between institutional goals and a problem-led approaches. This suggests that problem-solving approaches were usually adopted for improving institutional performance. The network of novice institutions also displays a strong connection between institution-level goals and measuring approaches. This suggests that LA was often adopted as a measuring tool for institutional performance, e.g., student retention rate. By contrast, more experienced institutions showed strong connections between teaching-level goals and exploratory approaches. This suggests that as institutions' experience with LA increased, there was a growing interest in understanding a teaching or learning phenomenon to enhance teaching.

4 CONCLUSION

The strong connection between 'institution goal' and the 'problem-led' approach alludes to the political pressure that HEIs are currently under – providing evidence to demonstrate and enhance excellence and quality. Nevertheless, the study showed a movement among the institutions from measuring the phenomena of learning and teaching to exploring them for a better understanding of factors that contributed to the phenomenon. This suggests a need for a strategy that evolves based on evaluations of short-term objectives for LA, as institutions' experience matures.

REFERENCES

- Colvin, C., Dawson, S., Wade, A., & Gašević, D. (2017). Addressing the Challenges of Institutional Adoption. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (First, pp. 281–289). Society for Learning Analytics Research.
- Kotter, J. P. (2006). Leading Change: Why Transformation Efforts Fail. *Harvard Business Review*, 1–10.
- Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., ... Mislevy, R. (2009). Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media*, 1(2), 33–53. <https://doi.org/10.1162/ijlm.2009.0013>

What Can We Learn About Learner Interaction When One Course is Hosted on Two MOOC Platforms?

Yuanru Tan, Rebecca M. Quintana
University of Michigan
{yuanru, rebeccaq}@umich.edu

ABSTRACT: Since the inception and adoption of MOOCs, pedagogues have criticized the quality of social learning within centralized platforms. Learning analytics researchers have investigated patterns of forum use and their relationship to learner performance. Yet, there are currently no cross-platform comparisons that explain how technical features of MOOC platforms may impact social interaction and the formation of learner networks. To address this issue, we analyzed MOOC discussion forum data from a single data science ethics course that ran concurrently on two different MOOC platforms (edX and Coursera). Using Social Network Analysis methods, this study compares networks of active forum posters using “Direct Reply” and “Star” tie definitions. Results show that the platforms afforded formation of different networks, with higher connectedness and higher network centralization seen on edX. This study presents preliminary results, discusses limitations inherent within the current analysis, and sets further directions of research investigating design features of centralized discussion platforms.

Keywords: MOOCs, discussion forums, social network analysis

1 INTRODUCTION

In Massive Open Online Courses (MOOCs) discussion forums are a principal means of enabling peer-to-peer interactions, a key aspect of social learning. Existing MOOC forum research has examined how learners interact with each other, such as by specifying the number of contacts a learner makes (i.e., either sending or receiving a comment). Yet, no existing research has examined the potential reciprocal effects between (1) learner activity on discussion forums and (2) technological platform affordances. Typically, institutions offer a course on a single platform, which has prevented cross-platform comparisons that allow us to understand the influence of technical affordances on social interaction. However, this lack of insight is problematic, given that technological affordances are an integral part of socio-technical online learning environments (Skrypnik et al., 2015). To better understand these dynamics, we used data from two instances of the *same* ethics of data science course, offered on both edX and Coursera. This study’s goal is to define the structure of learner-to-learner interactions—as captured through social network analysis (SNA)—and to hypothesize the impact of various platform-specific factors, such as user interface design. Such preliminary work allows us to hypothesize the relationship between centralized platform features and structural features of learner-to-learner networks, which can be further tested in future work.

2 METHODS

Our study examined discussion forum data from a data science ethics MOOC created by a large U.S. Midwestern university. The course used a case-based approach, offering multiple opportunities for discussion in the course forums around these cases. We used data from the first six months that the

MOOC ran on edX and Coursera. On edX, 168 out of 6,058 learners (2.78%) posted in the forums, creating a total of 452 posts. On Coursera, 193 learners out of 1204 learners (16.03%) posted in the forums, creating a total of 724 posts. Using SNA methods, we constructed networks to align study indicators with previous work (Wise & Cui, 2018): we created the edge-list following *Direct Reply* ties (i.e., the author of each post was connected with the author of its parent post) and *Star* ties (i.e., the author of each reply and reply-to-reply was connected with the author of the starting post). We extracted the node-list from the case study posts after excluding the posts of logistical questions (e.g., assignment submissions). We constructed one network using the two ties for each platform. Both networks were undirected and weighted.

3 FINDINGS AND FUTURE WORK

The edX network demonstrated higher connectedness and higher network centralization, implying that a greater number of central learners were critical to facilitate frequent interaction. In contrast, the Coursera network had a lower density and lower centralization, implying lower interaction overall among learners using the forums, and lower cohesion within the network as a whole. We also observed that edX learners interacted with more learners on average. This finding was reflected in higher numbers of edges, average degree, density, and centrality on edX, see Table 1. While the number of forum participants on edX (n=159) was lower than on Coursera (n=187), a higher standard deviation of average degree was observed on edX.

Table 1: SNA measures of learner interaction network in each platform

	Nodes	Edges	Density	Average degree (SD)	Betweenness	Closeness	Centrality
Coursera	93	166	0.036	2.18 (1.45)	0.45	0.004	0.18
edX	145	419	0.421	8.17 (2.81)	0.88	0.06	0.56

We conclude that these SNA methods show promise for understanding the impact of platform features on discussion forum interactions. These findings show that edX discussion forums have greater potential to engender interaction among learners than Coursera forums, when the course design is identical. We hypothesize that some features of centralized platforms may contribute to these behaviors. For instance, on edX, pre-existing posts are visible to learners before they respond to a prompt, while on Coursera, learners must respond to the prompt without seeing historic posts. This may account for the higher participation rate on Coursera, though a reduction in learner-to-learner interaction. Hence, platform affordances may facilitate different types of discussion forum behavior (e.g., initiating a new thread, replying to a new thread). Future work will therefore further compare post *types* in order to better understand the role of platform features in promoting social interaction among learners. Additionally, we will employ user-experience testing methods to better understand the impact of user interface design on social interaction.

REFERENCES

- Skrypnik, O., Joksimović, S., Kovanović, V., Gašević, D., & Dawson, S. (2015). Roles of course facilitators, learners, and technology in the flow of information of a cMOOC. *International Review of Research in Open and Distributed Learning*, 16(3). doi: 10.19173/irrodl.v16i3.2170.
- Wise, A. F. & Cui, Y. (2018). Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education*, 122(18), 221–242.

Dynamic Feedback System supporting self-regulation, adaptive teaching and program level curricular development

Ville Kivimäki
Aalto University, Finland
ville.kivimaki@aalto.fi

Joonas Pesonen
University of Helsinki, Finland
joonas.pesonen@helsinki.fi

ABSTRACT: Teaching in higher educational institution (HEI) is often based on courses or modules that are small fractions of the full degree-level curriculum. Several methods for visualizing the curriculum have been developed and tested from an institutional point of view. In this demonstration, we introduce a method for creating a curricular concept map that is distributed to all learners. As they enrich the template based on their experiences, the template turns into a structured learning diary, aimed at supporting the development of self-regulated learning skills. After this, the data is aggregated and visualized on a shared dashboard. This dynamic feedback loop creates opportunities for adaptive learning and teaching during the course and curriculum development on the program level.

Keywords: dynamic, feedback, adaptive teaching, dashboard, curriculum, concept map, self-assessment, universal design

1 DEMONSTRATION OF THE DYNAMIC FEEDBACK SYSTEM

Dynamic Feedback System (DFS) is a toolset that integrates curriculum-level thinking to everyday teaching and learning activities (Kivimäki et al, 2018). Having teachers and learners situate the learning topic at hand within the course syllabus and degree-level curriculum can stimulate seeing that the curriculum as being open to change and improvement (Wijngaards-de Meij & Merx, 2018). In DFS, students self-monitor their learning process by enriching a curricular concept map template, which thus becomes a structured learning diary. Various self-monitoring items can be integrated to this process, including structured items, open items and topic relationship items (concept mapping). While the self-monitoring exercise aims for the development of the learner's self-regulatory skills, sharing the data collected with the tool as a dashboard for the teacher and the learners generates a culture of dynamic feedback for the entire degree program.

This demonstration is based on a methodology that is under development. However, DFS has been tested in various forms in over 20 courses and one full degree program. We started with commercial mind-mapping software and are now working on a mobile app, a web app and a Moodle plugin. Interactive demonstration contributes to DFS development and opens possibilities for open-source-based collaboration. The DFS video presentation consists of two parts: a motivation video (<https://www.youtube.com/watch?v=zAmFmfNaWYo>) and a demonstration of the tool itself (https://drive.google.com/file/d/1MzE_G-lhB8xF_oluHKwHhXX5KCetLaZ/view?usp=sharing).

REFERENCES

- Wijngaards-de Meij, L. & Merx, S. (2018). Improving curriculum alignment and achieving learning goals by making the curriculum visible, *International Journal for Academic Development*, 23(3), 219-231, DOI: 10.1080/1360144X.2018.1462187
- Kivimäki, V., Pesonen, J., Romanoff, J., Remes, H., & Kauppinen, T. (2018). Supporting understanding of students' learning via visual self-assessment, in *Proceedings of EUNIS 2018 – Coming of Age in the Digital World*, 2018.

Comparing Interaction Activity Patterns of Different Achievement Learner Groups in MPOCs

Di Sun Syracuse University dsun02@syr.edu	Gang Cheng The Open University of China chenggangouc@163.com	Pengfei Xu Beijing Normal University xupf@bnu.edu.cn	Qinhua Zheng Beijing Normal University zhengqinhua@bnu.edu.cn
---	--	--	--

ABSTRACT: Comparing interaction activity patterns of different achievement learner groups in MPOCs has been paid little attention in online learning research. This study used hidden Markov models to identify activity interaction patterns of two different achievement groups in MPOCs settings. The results demonstrated high-achievement learners especially spent time on content learning, assessment, and discussion to consolidate their knowledge construction, while low-achievement learners did not perform the same. Although all the learners were interested to check learning statements; however, low-achievement learners spent 80% of their time on it, and ignored other learning activities.

Keywords: Comparing, interaction, patterns, different achievement groups, MPOCs

1 INTRODUCTION AND RELATED WORK

Different learners conduct their learning with different interaction characteristics (Chen, 2004). Some researchers have employed learning analytics techniques such as clustering, sequential pattern mining, and hidden Markov models (HMMs) to compare the interaction patterns between the high and the low performers in online settings (Jeong, Biswas, Johnson, & Howard, 2010; Kinnebrew, Loretz, & Biswas, 2013; Martinez, Yacef, Kay, Al-Qaraghuli, & Kharrufa, 2011; Perera, Kay, Koprinska, Yacef, & Zaïane, 2009). Their studies indicated that strong learners did perform more effectively than weaker learners. However, these efforts limited within a small scale of learners conducting collaborative learning, and did not tackle the context of massive online environments, especially on massive private online courses (MPOCs) (Guo, 2014). Thus, to improve all MPOCs learners' achievements, this study aims to investigate: In MPOCs, are the interaction activity patterns for the high-achievement learner group and the low-achievement learner group the same?

2 METHODOLOGY

In an open university of the south China, 1,481 out of 1,560 learners finally finished an 18 weeks online course. There were 12 course activity modules: Introduction, Announcement, Content, Resource, Assignment and Quiz, Forum, Frequently asked questions, Experiment guide, Group learning, Learning Statement, Exam, and Course evaluation. The final grade C was used to divide these learners into High (grade $\geq C$, $n=1,025$) and Low (grade $< C$, $n=456$) achievement groups. HMMS with three sets of parameters are the key to identify hidden activity states of two groups (Jeong et al., 2008; Rabiner, 1989): 1) Initial probability vector π , initial probabilities for hidden

activity states, each of which represents the proportion of the engagement time of a hidden activity state in the entire course process. 2) Transition probability matrix, A , transition probabilities between each of the hidden activity states. 3) Output probability matrix, B , probabilities to detect particular observable activities in a hidden activity state; each output probability represents the proportion of the engagement time of an observable activity in a given hidden activity state. The 12 learning modules are the observable interaction activities in this study. Bayesian information criterion (BIC), Baum-Welch method, and the Viterbi algorithm were used to identify π , A , and B . Sequences of hidden activity states were interpreted as interaction activity patterns.

3 RESULTS

The comparison between two groups were based on Figure 1 and Figure 2 (values lower than 0.08 in matrix A and values lower than 0.2 in matrix B did not show in the figures).

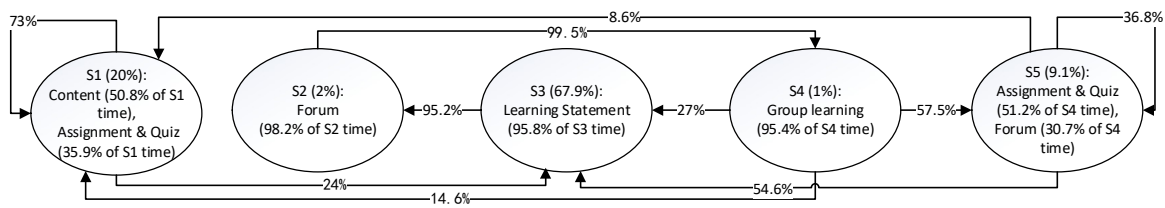


Figure 1: HMM of High-achievement group

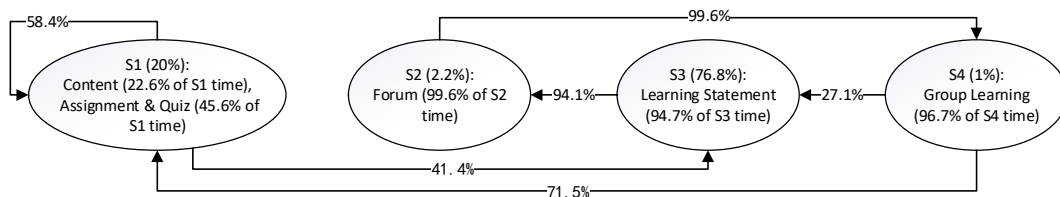


Figure 1: HMM of Low-achievement group

Except S5 (Assignment & Quiz, and Forum) in High-achievement group, the two groups performed similarly in S1 (Content, and Assignment & Quiz), S2 (Forum), S3 (Learning statement), and S4 (Group learning), especially the observable activities in each state and the transition patterns among the four states. In S1, S3 and S4, two groups spent similar time proportions (less time on S1 and S4, 1~2%; most time on S3, over 67%). Further, the transition patterns and probabilities of two groups in the circle of S2, S3, and S4 were almost the same: S3 to S2, and S2 to S4 were almost 100% of probability, S4 to S3 was around 27%.

The most noted difference was that High-achievement learners spent 9.1% of their time on S5 (Assignment & Quiz, and Forum), while Low-achievement learners did not have S5 because they spent (76.8%) more than around 9% of time on checking learning statements in S3 than High achievement ones (67.9%). In addition, the transition probabilities in S4 was different: except similar 27% of transition from S4 to S3, High-achievement learners transited to S1 with 14.6% of probability, to S5 with 57.5% of probability, but Low-achievement learners only transited to S1 with 71.5% of probability. Further, the time distributions of observable activities in S1 were different as well as the transition probabilities of S1: High-achievement learners spent more time on learning content and

less time on taking assessments, while Low-achievement learners performed totally oppositely. Moreover, High-achievement learners kept learning in S1 with 73% of probability, while Low achievement learners only stayed with 58.4% of probability because they transited more to check learning statements in S3 (41.4%) than High-achievement ones (24%).

One interesting discovery was that learners in both of the groups spent over 65% of learning time in S3 to check their learning statements. Except transition from S3 to S2, learners in S1, S4 and S5 all transited to S3 with more than 20% of probability. Especially High-achievement learners in S5, they transited to S3 with 54.6% of probability (the highest probability), while stayed in S5 with 36.8% of probability and transited to S1 with 8.6% of probability.

4 CONCLUSIONS

Two different interaction activity patterns of High and Low achievement learners in MPOCs were generated by HMMs, and compared based on three sets of HMMs parameters (π , A, and B). The results indicated that certain interaction patterns distinguish strong learners from weak ones in MPOCs settings. Further, focus on content and assessment is the basis of effective learning, taking assessment and discussion is much helpful to consolidate learners' knowledge construction and gain the final achievement in MPOCs. We also discovered that all learners cared about their own and peers' learning statements, which enlightens researchers to design more effective functions in learning statement module in LMSs to guide learners to focus on their learning and gain high achievement.

REFERENCES

- Chen, L. (2004). An investigation into "Interactivity" and the related concepts. *China Distance Education*, 3, 12-19.
- Guo, W. (2014, October). *From SPOC to MPOC--The effective practice of Peking University online teacher training*. Paper presented at the International Conference of Educational Innovation through Technology, Queensland, Australia.
- Jeong, H., Biswas, G., Johnson, J., & Howard, L. (2010). *Analysis of productive learning behaviors in a structured inquiry cycle using hidden Markov models*. Paper presented at the 3rd International Conference on Educational Data Mining, Pittsburgh, USA.
- Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008). *Using hidden Markov models to characterize student behaviors in learning-by-teaching environments*. Paper presented at the International Conference on Intelligent Tutoring Systems.
- Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM-Journal of Educational Data Mining*, 5(1), 190-219.
- Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., & Kharrufa, A. (2011). *Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop*. Paper presented at the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands.
- Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759-772.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in Speech Recognition*, 77(2), 267-296.

Academic Quality Data Landscape: Establishing a Sustainable Process to Measure Learner Performance

Authors: Mamta Saxena & Melanie Kasparian

Northeastern University,
College of Professional Studies
m.saxena@northeastern.edu

ABSTRACT: Poster submission. Student learning and academic quality are central to higher education. Nonetheless, colleges and universities spend more time and resources on data collection and reporting of metrics such as enrollment or graduation rates relative to student learning. The Academic Quality Assurance (AQA) team at Northeastern University's College of Professional Studies ventured to expand the data landscape to find out more about learning and performance data. While the immediate goal was to show measurable impact on learning based on an annual assessment cycle, the end goal was to promote a culture of assessment and a model of continuous improvement for programs by using learning analytics to inform planning and implementation. While other initiatives may be focused on stories based on the data, this is the story about the data: This poster will share a team's journey to collect and report the academic quality data and the challenges faced in the context of people, process, and tools.

Keywords: outcomes, student, learner, performance, Tableau, academic quality, rubrics, grades, assessment, faculty, dashboards, assignments, competencies

Reference

Maki, P. L. (2017). *Real-Time Student Assessment: Meeting the Imperative for Improved Time to Degree, Closing the Opportunity Gap, and Assuring Student Competencies for 21st Century Needs*. Sterling: Stylus.

1 PURPOSE

In order to measure learning and academic quality, the College of Professional Studies began to formalize its curriculum assessment practices in 2012 when the AQA unit was formed. After defining program learning outcomes based on the Degree Qualification Profile framework, faculty selected signature assignments as authentic demonstrations of learner performance (capstone, case studies, projects with employers, field exams, etc.) to assess if the learners achieved the expected outcomes. Then came the challenges, operational and technical, related to people, process, and tools of data collection and reporting. Key questions included: How to obtain organizational buy-in; how to create a streamlined process for data collection; and how to best utilize existing tools to analyze and report the data real-time. Data collection only occurred in 2016 after the AQA unit worked with faculty to develop consistent assignments within programs and use rubrics in the Learning Management System (LMS). The AQA unit then established a process to report on this recently acquired data to make it meaningful for faculty—drawing from methods of data visualization and effective storytelling.

2 METHODOLOGY

AQA relies heavily on program faculty leads to identify key data points about learner performance and to implement assessment frameworks. AQA helps define the assignments to ensure it measures the outcomes, and then collects and analyzes the data (grades and rubric scores) against benchmarks. Data collection is as reliable as the assumptions listed below. Hence, each term, AQA works with the program leads to address identified gaps to improve data reliability in the subsequent terms.

- Alignment: Rubrics are aligned to the learning outcome and assignment
- Consistency: The same assignments are used across class sections
- Accuracy: Rubrics are calibrated so the data is the same (Inter-rater reliability)

The AQA unit is also responsible for the oversight of the completion of an **annual program evaluation cycle**, where faculty leads review the performance data along with graduation and retention rates, survey data, and participation in experiential activities. These reports include a narrative section where the faculty can tell a story of the program based on the data and set goals for the subsequent year based on the gaps and success for continuous improvement. AQA and the Deans track the status of the goals to help close the assessment loop and to share measurable impact on learning based on the changes made by the programs.

3 FINDINGS

The college now has a process where data is collected after each term, then compiled, analyzed, and updated in Tableau and distributed to faculty. Faculty enjoy seeing the results of the data for their program in a clean and organized dashboard view where they can look at their results and then make informed decisions such as addressing issues related to alignment, consistency, accuracy, and updating the curriculum. The image below shows sample data on learning outcomes for a program, where the results are displayed as percent of learners who met the benchmark set by the program:

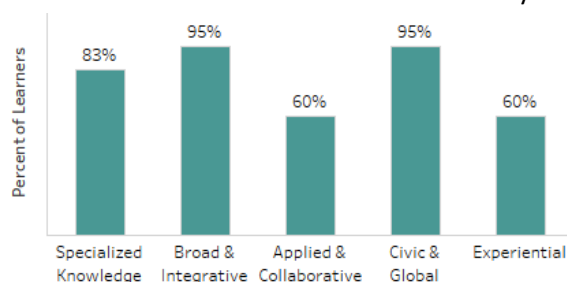


Figure 1: Sample Tableau Results on Program Outcomes

Ultimately, it took several years to define and launch data on learner performance across 45 programs and hundreds of course sections. As frameworks and goals change, the notion of using data to inform teaching and learning remains the same. In order for the process to work and be sustainable, several challenges were overcome:

3.1 Technical Challenges

The metrics to measure success and learning for program evaluation live in different systems. While there are many points of direct connection with Operational Databases such as the Student Information System, not everything is connected, such as survey data. There is still a need to cross-reference different dashboards and create high level views that would include the majority of the metrics in one place.

3.2 Building Capacity and Obtaining Executive Support

In order to maintain momentum, it was necessary to schedule regular monthly meetings with program faculty leads. Obtaining the buy-in and promoting data literacy has been a significant challenge. Learning analytics became a priority for leadership but is yet to be explicitly included as a KPI. Therefore, it has been difficult to make this a priority for programs given the limited faculty time and resources.

3.3 Changing Frameworks

In 2012, AQA selected the Degree Qualifications Profile (DQP) framework and the AACU Value rubrics to help drive the direction of our outcomes and rubrics. The university moved toward a different framework based on industry domains and demands for new interdisciplinary skills such as technology and data literacy. This posed yet another challenge of aligning assessments with a competency-based model unique for each program.

3.4 Alignment and Consistency

There were several data collection roadblocks encountered, such as different assignments across sections, inconsistent naming conventions, or rubrics were not built in the LMS. With LMS data, consistency was key for data reliability and automation.

3.5 Structuring the Team for Strategic Action

AQA is part of the Strategic Research and Analysis Team, which provides resources and support to move the initiatives forward. Without the analytics resources and tools, assessment data collection would not be perceived as a priority for the college.

4 NEXT STEPS

This process has helped inform existing learning analytics initiatives, such as analyzing and distributing survey data or LMS data. However, there are several challenges we still face. There are a number of programs that have only started to engage with AQA. Obtaining buy-in and improving the data literacy of the college is still an issue. AQA has several trainings, including sessions on creating rubrics, storytelling, and interpretation of data to assist in data literacy. We are also building out Academic Quality Insight Sheets that provide a snapshot of a number of data points all in one place to help faculty connect the dots between this work and other initiatives, such as LMS data, retention rates or course grades. Ultimately, AQA continues to strive to find ways of using learning analytics to:

1. Inform the teaching and learning at the program and college level (improvement); and
2. For reporting to respective external accrediting bodies (accountability).

Finding the At-Risk Online Learners: Development of the Online REadiness Screener (ORES)

Oi-Man Kwok, Yu-Chen Yeh, Hsiang-Yu Chien, Noelle Wall Sweany, Eunkyeng Baek,
William McIntosh

Texas A&M University

{omkwok, yuchen188, johnny.chien, nsweany, baek, w-mcintosh}@tamu.edu

ABSTRACT: The goal of this study was to develop a screener that could identify at-risk online learners. With the use of both stepwise regression and exploratory factor analysis, we created an 8-item (3-facet) screener that was validated with two different samples.

Keywords: At-Risk Online Learners, Online Learning Readiness, Stepwise Regression, Exploratory Factor Analysis, Path Model

1 INTRODUCTION

Online learning is one of the drivers of the learning analytics development (Ferguson, 2012) and has become a popular option for students to complete course requirements and pursue a college degree. According to the latest *Distance Education Enrollment Report* (Digital Learning Compass, 2017: <http://digitallearningcompass.org>), more than six million students had taken at least one online course in 2015. However, despite the increasing popularity of online learning in U.S. higher education, online courses are often associated with higher dropout rates, presenting a major concern for many universities and higher education institutes (Moody, 2004). Therefore, it is important to determine who is at risk so that instructors can provide appropriate assistance and adjustments to help these students successfully complete their online coursework. Funk (2005) defined at-risk online learners as those who are not expected to succeed and who drop out early. Successful online learners are likely to demonstrate the following characteristics and skills: high learning motivation and self-regulated learning habits, effective time-management skills, and low multitasking self-efficacy (Cohen & Baruth, 2017). The aim of this study was to create a short screener that could identify at-risk online learners based on a set of questions measuring different psychological and behavioral aspects that closely related to online learning readiness.

2 METHODS AND RESULTS

Two cohorts of participants (1st cohort: 93 students; 2nd cohort: 46 students) were recruited from a large public university (funded by the T3 initiative; Texas A&M University). Based on the first cohort data we developed a screener, and then validated it with the data of both cohorts separately. The questionnaire used with the first cohort consisted of 92 items from a set of online learning related questionnaires. Students' expected grade and academic expectations were the major outcomes. With the use of the first cohort data and the expected grade as the outcome, eight items (see Table 1) were selected through stepwise regression analysis. We subsequently ran an exploratory factor analysis on these eight items, which led to a 3-factor solution: social media notification (Q6),

learning strategy (Q2, Q3, Q4, Q5, and Q8 with an average factor loading equal to .494), and failure-avoidant motivation (Q1 and Q7 with factor loadings equal to .912 and .375, respectively).

Table 1. Stepwise Regression Summary Table.

Item	Description	R ²	F	ΔR^2	ΔF
Q1	Sometimes I am afraid that I may not understand the content of this online class as thoroughly as I'd like.	.082	8.172	.082	8.172**
Q2	I can learn by working independently.	.148	7.833	.066	6.958**
Q3	I am capable of solving problems alone.	.226	8.674	.078	8.970**
Q4	I need faculty to remind me of assignment due dates.	.295	9.225	.069	8.643**
Q5	I know my resources.	.336	8.810	.041	5.334*
Q6	When I see or hear notifications from social media (e.g., Twitter, Instagram, Facebook), I cannot wait to check them.	.374	8.549	.037	5.144*
Q7	I just want to avoid doing poorly in this online class.	.411	8.466	.037	5.364*
Q8	I find it hard to stick to a study schedule for this online course.	.438	8.184	.027	4.071*

Based on the 3-factor solution, we then created three composite scores and used them to predict the outcome variable within each cohort. For the first cohort data (Figure 1), both learning strategy ($\beta = .23, p < .01$) and failure-avoidant motivation significantly predicted the students' expected grade ($\beta = -.32, p < .01$). For the second cohort data (Figure 2), only learning strategy significantly predicted students' academic expectations ($\beta = .29, p < .05$). This difference may be a result of the low statistical power due to a smaller sample size in the second cohort (only 46 students).

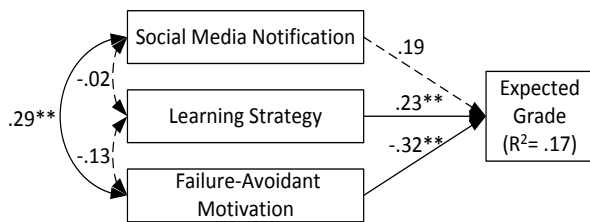


Figure 1: Path analysis for the cohort 1 data

Note. All the coefficients were standardized. Dashed lines represent no significant association.

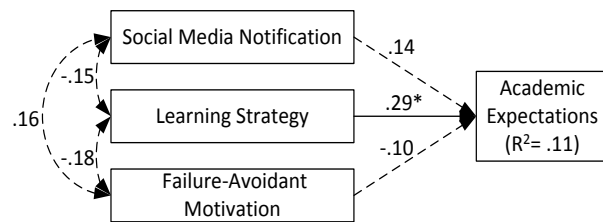


Figure 2: Path analysis for the cohort 2 data

3 DISCUSSION

The goal of this study was to create a screener that could identify at-risk online learners. With the use of stepwise regression and exploratory factor analysis (classic data exploratory techniques), we created an 8-item screener that predicted students' expected academic outcome. Based on these findings, we intend to apply the screener to a larger sample with the goal of ultimately providing additional helpful assistance to college students who are identified as at-risk online learners.

REFERENCES

- Cohen, A., & Baruth, O. (2017). Personality, learning, and satisfaction in fully online academic courses, *Computers in Human Behavior*, 72, 1-12.
- Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304-317.
- Funk, J. T. (2005). *At-risk online learners reducing barriers to success*. Retrieved from <https://elearnmag.acm.org/featured.cfm?aid=1082221>
- Moody, J. (2004). Distance education: Why are the attrition rates so high? *The Quarterly Review of Distance Education*, 5(3), 205-210.

Examining the Effects of Adaptive Task Selection on Students' Motivation in an Intelligent Tutoring System

Micah Watanabe¹, Kathryn S. McCarthy², Danielle S. McNamara¹

¹Arizona State University, ²Georgia State University

micah.watanabe@asu.edu; kmccarthy12@gsu.edu; dsmcnama@asu.edu

ABSTRACT: Adapting tasks to the individual has been shown to improve learning, and improve learners' experiences in the learning process. This study investigated how adaptive task selection affected learners' experiences in iSTART, an intelligent tutoring system for improving reading comprehension. Participants (n = 59) engaged with iSTART for 7 hours across three sessions. Participants read and self-explained texts that were presented in random order or adaptatively based on participants' performance. Adaptive task selection did not increase engagement, but did enhance participants' judgments of learning.

Keywords: Intelligent Tutoring System, Scaffolding, Motivation.

1 INTRODUCTION

Interactive Strategy Training for Active Reading and Thinking (iSTART; McNamara et al., 2007) is an intelligent tutoring system (ITS) that provides reading comprehension strategy instruction through videos lessons, and game-based *self-explanation* practice. Recently, we have implemented adaptive text selection to increase individualization of instruction. An algorithm selects the difficulty of texts that the learner reads and self-explains based on their average self-explanation (SE) score (0-3). When average SE score is above a threshold (2.0), the algorithm selects a subsequent text that is more difficult. When the SE score is below the threshold, the algorithm selects an easier text. This adaptive task selection has effectively improved student learning (McCarthy et al., 2018).

Adaptivity may also benefit student experiences with the system. Scaffolding tasks to a learner's *zone of proximal development* (ZPD) has been shown to improve learners' motivation and engagement (Murry & Arroyo, 2002). Thus, this study used students' survey responses to investigate how adaptive task selection affected learners' experience (e.g., engagement, motivation, metacomprehension) during training. It was hypothesized adaptive text selection better targets students' ZPD, which may in turn enhance learners' experiences with iSTART.

2 METHOD

Participants (n = 59) engaged in 3 sessions (~7 hours) of iSTART training in which they are presented science texts and asked to write self-explanations during reading. In the random condition, participants received texts in random order. In the adaptive condition, an algorithm selected texts based on the participants' average self-explanation scores. At the end of each session, participants' answered 5-point Likert scale items about that day's training (Table 1). To account for differences in trait-level motivation, participants also completed the Learning Orientation (LO) and Performance Orientation (PO) scales (Jha & Bhattacharyya, 2013).

3 RESULTS

T-tests indicated no differences across conditions in LO, $t(57) = .21$, $p = .83$, or PO, $t(57) = .24$, $p = .81$. Nonetheless, LO was included as a covariate in subsequent analyses.

A series of 2(condition: random, adaptive) x 3(session: 1, 2, 3) ANCOVAs revealed no changes across sessions (all $F_s < 1.00$). Adaptive task selection had no significant effect on learners' overall experience or enjoyment of iSTART, nor did adaptivity increase negative experiences (boredom, frustration). However, participants in the adaptive text selection condition more strongly agreed that they learned the material presented in the texts that they had read (Table 1).

Table 1: Likert Item Scores (EMMs and SE) and ANCOVA results as a Function of Condition

	Today's session was: (1-6)	I was bored (1-5)	I was frustrated (1-5)	I had problems with the program (1-5)	I felt like I learned the material (1-5)	I feel like my reading skills improved (1-5)	I enjoyed today's session (1-5)
	<i>EMM (SE)</i>	<i>EMM (SE)</i>	<i>EMM (SE)</i>	<i>EMM (SE)</i>	<i>EMM (SE)</i>	<i>EMM (SE)</i>	<i>EMM (SE)</i>
Random	4.08 (.17)	3.39 (.15)	2.32 (.16)	1.91 (.14)	3.05 (.11)	2.99 (.16)	2.93 (.17)
Adaptive	4.39 (.18)	3.05 (.16)	2.48 (.16)	2.20 (.15)	3.59 (.12)	3.37 (.17)	3.27 (.18)
ANCOVA (F_s)							
Condition, $F(1,56)$	1.70	2.46	< 1.00	1.90	10.63**	2.46	1.90
Condition x Session, $F(1,112)$	< 1.00	1.09	< 1.00	1.79	1.22	2.05	< 1.00

** $p < .01$; Note: For session, all $F_s < 1.00$

4 DISCUSSION

Consistent with the theory of learners' ZPD, adapting the difficulty of a task to the learner increased their sense of learning. However, adaptivity had no significant effect on learners' self-reported enjoyment or engagement. Future work will explore log data to examine how task adaptivity impacted more moment-to-moment experiences in iSTART and how the effects of adaptivity may depend on learners' individual differences in skills.

REFERENCES

- Jha, S., & Bhattacharyya, S. S. (2013). Learning orientation and performance orientation: Scale development and its relationship with performance. *Global Business Review*, 14(1), 43-54.
- McCarthy, K. S., Johnson, A.M., Watanabe, M., & McNamara, D. S. (2018). Implementing the Outer Loop in iSTART: Adapting Text Difficulty as Feedback. Poster presented at the Society for Computers in Psychology (SCiP), New Orleans, LA.
- McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. *Reading comprehension strategies: Theories, interventions, and technologies*, 397-421.
- Murray, T., & Arroyo, I. (2002, June). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In S.A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.) *International Conference on Intelligent Tutoring Systems* (pp. 749-758). Springer, Berlin, Heidelberg.

Know Your Students : Empowering Educators to Improve Learning Using Data

Matthew Steinwachs

University of California, Davis – Center for Educational Effectiveness
mksteinwachs@ucdavis.edu

Marco Molinaro

University of California, Davis – Center for Educational Effectiveness
mmolinaro@ucdavis.edu

ABSTRACT: Know Your Students (KYS) is a web application designed to improve inclusive instruction across undergraduate courses on the UC Davis campus by raising awareness of key characteristics of a class; helping instructors gain a deeper understanding of their students through a centralized repository of instructional support materials, analysis tools, and expertise to guide action that improves inclusive instruction; and providing a place for reflection on this process and its outcomes to facilitate the continual improvement of courses and to document these efforts for appropriate recognition.

KYS is being developed by the Center for Educational Effectiveness (CEE) as part of a five-year project funded by the Howard Hughes Medical Institute. Currently we are piloting KYS with a limited number of trained instructors with the goal of making it available to all faculty in the coming years.

The core of KYS is collection of charts and statistics which describe aggregate characteristics of the students in a course. We believe that giving an instructor aggregate information about the students in their course can encourage actions that improve learning and reduce achievement gaps, but exposure to this information should be paired with resources that highlight positive actions that can be taken. Each of the charts shown in KYS is linked to relevant education research and best practices by a tagging system. Instructors can submit best practices and links to research to be shared with other instructors. These resources are voted on by users so that the most relevant or useful information appears most prominently.

After deciding on a course of action, an instructor can access a suite of tools within KYS that allow the collection of data to measure how a course has changed and what effect those changes have on learning and achievement gaps.

KYS provides tools that can aid in course reflection. For example, a text analysis of test questions or course learning outcomes can categorize questions and outcomes according to Bloom's Taxonomy and/or reading level which may provide insight into observed outcomes or facilitate planning the next course offering.

KYS allows instructors to take notes and record questions, ideas, planned actions, and outcomes. These notes can be useful reminders for instructors who may not teach a course frequently. Furthermore, KYS can compile these notes along with relevant charts and data into a Teaching Portfolio which documents their pedagogical or curricular innovation for merit and promotion.

Keywords: inclusive excellence, inclusion excellence, instructional innovation and research, data visualization, performance gaps, student demographics, professional development, faculty, instructors, analytics

Examining Procrastination Behavior in Academic Settings Using a Mobile App

Semih Bursali, Majed Ali, Reza Feyzi Behnagh

State University of New York, Albany

{sbursali, mmali, rfeyzibehnagh}@albany.edu

ABSTRACT: This poster describes a mobile app designed to help students plan tasks, monitor progress, and gather trace and self-report data on their procrastination behavior. The app can be used for planning and managing individual and group tasks. As a research tool, the app administers surveys at predetermined time points and gathers time-stamped trace data of all student interactions with the app at all times. A progress bar and page has also been implemented as a dashboard to provide feedback to students on their progress on individual and group tasks. Data from the app will be used to (a) examine the underlying processes of procrastination, (b) model procrastination behavior in individual and group settings, and (c) upon implementation in a real classroom, to automatically detect and ultimately predict procrastination to understand student disengagement and dropout.

Keywords: procrastination, self-regulated learning, trace data

1 BACKGROUND

Procrastination, refers to the voluntary, irrational delay of beginning or completing an intended action despite expecting negative consequences for the delay (Pychyl et al., 2000). Procrastination is identified as failure of self-regulated learning (SRL), or failure to effectively plan (set goals), monitor (monitor progress toward goals, time management), and self-reflect (revise plans based on feedback from monitoring) (Pychyl & Flett, 2012). Procrastination has been linked to negative academic performance as well as poor health and wellbeing (Kim & Seo, 2015). Existing studies have predominantly examined underlying processes of procrastination using only self-report surveys; very few have employed ‘trace’ data of student actions to infer regarding procrastination antecedents and corollaries. The use of surveys underlie the ‘trait’ view of procrastination, whereas a variety of factors (e.g., value, interest, difficulty) could influence how one procrastinates on particular tasks (van Eerde, 2003). Additionally, the state-of-the-art in procrastination research focuses on how individuals procrastinate, however, the underlying processes of procrastination in groups or collaborative tasks has not been investigated. The goal of this poster is to introduce a mobile app we have developed as a research tool to collect trace and self-report data on students’ procrastination behavior.

2 DESIGN OF THE APP

We have been developing a mobile application to scaffold students in accomplishing academic tasks individually and in collaboration with their peers. This app is primarily a research tool to collect self-report and trace data. As a research tool, the app tracks all student actions along with the corresponding timestamps (e.g., goals and subgoals set, study times, progress) and responses to

administered in-app surveys on a secure online cloud server allowing us to draw informed inferences about students' procrastination behavior and factors influencing that. In order to understand the underlying processes of procrastination, at pre-determined times, we administer short pop-up surveys to measure different cognitive, metacognitive, motivational, and affective processes. These data will be used in conjunction with online generated trace data. When the app is opened, the student is asked to create an individual or group task (e.g., project), propose a goal and one or more sub-goals, along with deadlines, level of work involved, and who will take over the sub-goal (in group work). In order to keep track of when and for how long students engage in task-related behaviors, we have integrated a tool by which students can indicate they've started working on the task, chat with their peers, and mark the end time. An adjustable 25-minute work followed by 5-minute rest Pomodoro style timer is also made available. In order to track students' progress, a color-coded progress bar along with a progress page for main goal and the sub-goals have been developed that incorporate information gathered from the user(s) to mark the progress of the task, goals met, time spent, and time remaining to the deadline. This progress feedback tool acts as an 'open learner model' or a learner-facing-dashboard where the user can see a summary of their performance visually (Bull & Kay, 2013). Such feedback dashboards have been indicated in the literature to improve students' self-regulated learning and monitoring skills. Trace and self-report data from the app will be analyzed for (1) modeling individual and group procrastination, (2) understanding how peers in a group negotiate goal setting and sub-goal assignment, (3) modeling procrastination (individual and group) in accordance with performance (e.g., assignment grade), and (4) understanding how students utilize (e.g., visitations, revision of behavior after) the progress bar and feedback dashboard. Findings will contribute to the existing theories of procrastination in (a) enhancing the description of underlying and corollary cognitive, metacognitive, motivational, and affective processes, and (b) understanding individual-in-group procrastination and differences with individual procrastination. Furthermore, models developed based on data from students in real classes will help in detecting and early prediction of procrastination, allowing interventions to prevent student disengagement and dropout.

3 REFERENCES

- Bull, S. & Kay, J. (2013). Open learner models as drivers for metacognitive processes. In *International handbook of metacognition and learning technologies*. Springer: New York, NY.
- Kim, K., & Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality and Individual Differences*, 82, 26-33.
- Pychyl, T., & Flett, G. (2012). Procrastination and self-regulatory failure: An introduction to the special issue. *Journal of Rational-Emotive and Cognitive Behavior-therapy*, 30, 203-212.
- Pychyl, T., Lee, J., Thibodeau, R., & Beck, A. (2000). Five days of emotions: An experience sampling study of undergraduate student procrastination. *Journal of Social Behavior and Personality*, 15, 239-254.
- van Eerde, W. (2003). A meta-analytically derived nomological network of procrastination. *Personality and Individual Differences*, 35, 1401-1418.

MOOC Effort Dashboard: An Interactive Web Dashboard Built in R

Jason Baik

Carnegie Mellon University
joonwoob@andrew.cmu.edu

John Stamper

Carnegie Mellon University
jstamper@cs.cmu.edu

Huzefa Rangwala

George Mason University
rangwala@cs.gmu.edu

Massive Open Online Courses (MOOCs) offer accessible education to students around the world. However, a major issue is the extremely high dropout rate. MOOC literature reveals that it ranges from 90% to 92%. Our research seeks to identify factors of dropouts. We developed a visual dashboard that targets educators and course administrators to understand student course behaviors and identify students at risk of dropping out.

We used R packages for data visualization and integrated these in an interactive web service with shinydashboard (Winston Chang and Barbara Borges Ribeiro, 2018) and shiny (Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson, 2018). Shiny provides interactive features in R and shinydashboard is a template for building dashboards in R.

The Stanford Lagunita Online course, Statistics in Medicine, enrolled approximately 9000 students, 7659 of whom were examined. We estimate that around 5000 students dropped out, which is a dropout rate of 65.3%. The landing page introduces the developer of the MOOC Effort dashboard, its target audience, a brief explanation of the other pages and short motivation for using our product. Next, the Overview of Class tab provides a bird's-eye view of the class with the number of students, dropout rate, distributions of module usage and final grade. The Final Grades Table livestreams data from a csv file by automatically updating changes every second. The third tab is the Student Selector, a customizable search query. When the user selects a table, corresponding filters appear. Then, the user chooses desired columns in the data and the output is a personalized view of the raw data. The Effort Level tab shows visualizations of students' effort by completion of course and effort level (high, medium, low). For interactive purposes, we added graphs of each student's effort with the plotly package (Carson Sievert, 2018). These graphs enable the user to pick students. Then, they illustrate each student's effort level over the course period by highlighting. The last tab lays out results of our K-Means Clustering analysis. We present our analysis with plotly graphs for user interaction.

The MOOC Effort Dashboard is certainly unique in the learning analytics world. It is coded in R, an open source programming language. In the near future, we will host the MOOC Effort Dashboard on our servers so that other researchers have access. Furthermore, we have made our application reproducible for other researchers by hosting the source code on GitHub: https://github.com/jasonbaik94/mooc_project_lak19. To view the demo video, please visit: <https://www.youtube.com/watch?v=9YZMc9x164o&feature=youtu.be>

References

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Winston Chang and Barbara Borges Ribeiro (2018). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.1. <https://CRAN.R-project.org/package=shinydashboard>

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.2.0. <https://CRAN.R-project.org/package=shiny>

Carson Sievert (2018) plotly for R. <https://plotly-book.cpsievert.me>

Development of a Real Time Viewing Status Feedback System and Its Impact

Yasuhiro MORI^{a*}, Komei SAKAMOTO^a & Takahiko MENDORI^b

^a Graduate School of Engineering, Kochi University of Technology, Japan

^b School of Information, Kochi University of Technology, Japan

225129c@gs.kochi-tech.ac.jp

ABSTRACT: Our research aim is to improve learner's activities by feedback based on learning history. We developed a real time viewing status feedback system on LMS. The system collects page transition of the teaching materials during the lectures. The system gives the collected information visually to teachers and students. The students can confirm how many students are viewing the previous or subsequent pages or same page as the teacher. Through this study, we confirmed to give the collected information affect students learning activities.

Keywords: Learning Analytics, real time, viewing status, feedback, LMS.

1 INTRODUCTION AND RELATED WORK

In recent years, Learning Analytics catch a great deal of attention because it can analyze learning histories to figure out the achievement level, problems of learner. The study of Shimada (2017) gave feedback to teachers and provided the following potential benefits teachers can adjust the lecture speed based on the real-time visualization of students' activities and teachers can slow down to allow students to catch up. In this study, we give feedback to students and research the affect students learning activities.

2 REAL TIME VIEWING STATUS FEEDBACK SYSTEM

We developed a real time viewing status feedback system on Moodle. The system collects page transition information of teaching materials during lectures and gives the collected information visually. Figure 1 shows our real time viewing status feedback system. The belt graph shows page number of students in color. The table shows the number of students browsing each page. Students can confirm how many students are viewing the previous or subsequent pages or same page as the teacher.

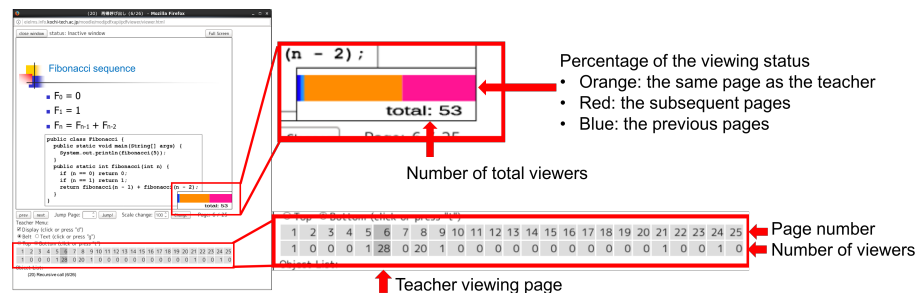


Figure 1: Our real time viewing status feedback system.

3 EXPERIMENT

We conducted experiments at our university classes (10 lessons, about 60 students) with two different condition. One was a control phase without the feedback (6 lessons) and the other was an experimental phase with feedback (4 lessons). In the experimental phase, students can confirm the collected information. Figure 2 compare of two heat maps in the lectures. The redline is the teacher's page transition. The cell color represents the number of students browsing each page. The left heat map shows control phase. The right heat map shows experimental phase. In the experimental phase, students tended to view the same page as the teacher. In the control phase, students tended to view pages before they were explained by the teacher. We compared difference between the teacher's viewing time and student's average viewing time with per page to the control phase. As the result, we use Mann-Whitney U test and found out that the difference time of the experimental phase was significantly different from the score of the control phase.

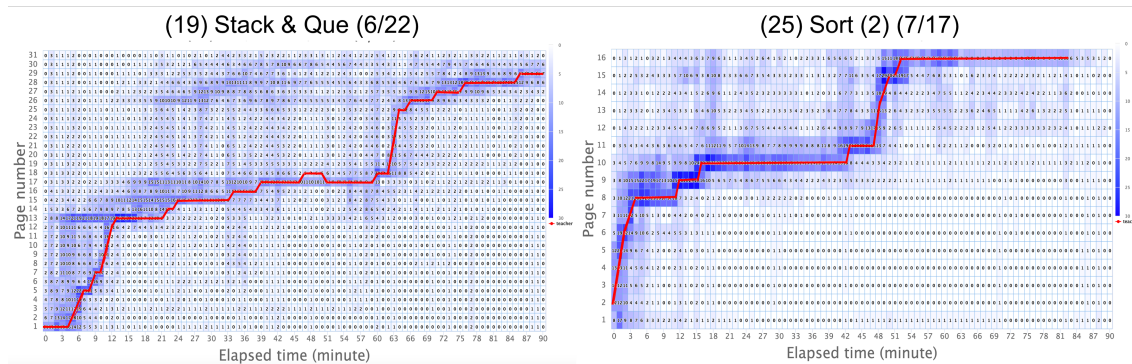


Figure 2: Comparison of two heat maps in the lectures.

4 CONCLUSION

This paper described our real time viewing status feedback system and its experimental result. As the result, we found out that the difference time of the experimental phase was significantly different from the score of the control phase and affect students learning activities. In our future work, we hypothesize that students expand an understanding of the lecture by to keep same page as the teacher. We will evaluate the control phase and the experimental phase with the learning results.

REFERENCES

- Horikoshi, I., Yamazaki, K., Tamura, Y. (2015). Learning Style Verification with use of Questionnaire and Page Flip History. Proceedings of the 23rd International Conference on Computers in Education. Hongzhou, China, pp.627-636
- Ogata, H., Yin, C., Mouri, K., Oi, K., Shimada, A., Okubo, F., Kojima, K. (2016). Learning Analytics toward the Usage of Educational Big Data. The Journal of Information and Systems in Education, Vol.33, No.2, pp.58-66. (in Japanese)
- Shimada, A., Mouri, K., Ogata, K. (2018). Real-time Learning Analytics of e-Book Operation Logs for On-site Lecture Support. Interactive Technology and Smart Education, 2018.

Multimodal Learning Analytics: Society 5.0 Project in Japan

Shizuka Shirai Osaka University shirai@ime.cmc.osaka-u.ac.jp	Noriko Takemura Osaka University takemura@ids.osaka-u.ac.jp	Yuta Nakashima Osaka University n-yuta@ids.osaka-u.ac.jp
Hajime Nagahara Osaka University nagahara@ids.osaka-u.ac.jp	Haruo Takemura Osaka University takemura@ime.cmc.osaka-u.ac.jp	

ABSTRACT: Multimodal learning analytics is expected to provide informative insights to support teaching and learning. This paper introduces our research project regarding multimodal learning analytics which recently started as a part of the Society 5.0 research project in Japan. We present the ideas of using multimodal data for blended learning and collaborative learning environments.

Keywords: multimodal learning analytics, blended learning, collaborative learning

1 INTRODUCTION

From November 2018, Osaka University has started “the initiative for Life Design Innovation (iLDI)” project to develop core technologies as a part of the MEXT (the Ministry of Education, Culture, Sports, Science and Technology) project for the realization of “Society 5.0”. Society 5.0 is proposed by the government as a future society that Japan should aspire to. In Society 5.0, a huge amount of information from various sensors in the physical space is accumulated and analyzed by artificial intelligence (AI). The analysis results are fed back to humans helping them to solve various social issues (Cabinet Office, 2018). In this context, the iLDI project will address prevention and resolution of some social issues using a large amount of multimodal personal life record data. The iLDI project consists of 4 research subprojects. One of them is “the school of the future assistance project.” It aims to develop a knowledge base for detecting students who showing signs of dropping out from university education, and for assisting students’ learning and school life. This paper introduces the latter project in which we are engaged. We present the ideas of collecting multimodal data of learners and analyzing them for the development of an in-class real-time adaptive learning system.

2 STUDENTS’ EDUCATION AND SCHOOL LIFE ASSISTANCE PROJECT

This project will address issues in two major types of learning styles: blended learning (a combination of face-to-face and online learning) and collaborative learning. In recent years, blended learning and collaborative learning have gained popularity in higher-education institutions. However, teachers and students have faced several challenges. For instance, lectures in higher education generally include a large number of students, and it is difficult for teachers to check individual student’ progress such as students’ comprehension, concentration, and confidence with learning content. Some students may also feel difficulties in asking something about they don’t understand in front of a large number of students whether the lectures are face-to-face or online.

Various approaches using logs of learners' interactions (e.g. video interaction such as number of pauses, number of backward seeks and, so on) have been studied to solve issues mentioned above. However, clickstreams per se cannot detect students' state, especially in real time. In addition, logs of learners' interactions are not available when learning face to face. To address these issues, we focus on multimodal data such as learner's eye movement, heart rate, physical body movement, and so on. Figure 1 summarizes our approach. In blended learning, students' state is estimated based on multimodal data accumulated during face-to-face learning and online learning. Then personalized feedback is provided to each student in real time. In collaborative learning, achievement of each group is analyzed using multimodal data for supporting teachers serving as facilitators. Most previous multimodal learning analytics studies have been conducted as experimental studies under controlled conditions (Blikstein & Worsley, 2016). In contrast, this project examines the adaptive learning framework in realistic scenarios from the same perspective as Martinez-Maldonado et al. (2018). First of all, we will conduct a preliminary experiment to estimate students' state in January 2019. From April 2019, we will start blended learning in information literacy courses with an approximate enrollment of 3,500 freshmen and will accumulate data for developing an adaptive learning paradigm with real-time support based on multimodal sensor data.

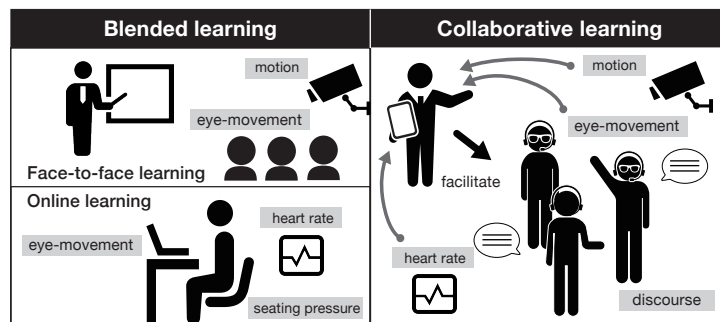


Figure 1: Overview of our projects.

3 CONCLUSIONS

In this paper, we presented an overview of our project that aims to assist students' learning and living. Though the project has just started in November 2018, we hope that the poster which includes the result of an experiment scheduled in January 2019 will provide a basis for future discussion regarding multimodal learning analytics.

REFERENCES

- Cabinet Office. (2018, Nov.). *Society 5.0*. Retrieved from https://www8.cao.go.jp/cstp/english/society5_0/index.html
- Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks. *Journal of Learning Analytics* 3, 2, 220-238.
- Martinez-Maldonado, R., Echeverria, V., Santos, O. C., Dos Santos, A. D. P., & Yacef, K. (2018). Physical Learning Analytics: A Multimodal Perspective. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 375-379.

Exploring Medical Education Learning Analytics from the Use of Electronic Health Record Systems

Yancy Vance Paredes¹, Sarada Panchanathan², Pamela Carol Garcia-Filion², I-Han Hsiao¹
Arizona State University¹, University of Arizona²
yvmparedes@asu.edu, spanchan@email.arizona.edu, pgarciafilion@email.arizona.edu,
sharon.hsiao@asu.edu

ABSTRACT: Poster. Patient simulations may be used to teach medical students necessary clinical reasoning skills. More recently, simulated patients in electronic health record (EHR) systems have been used for specific diagnosis and management tasks. Such systems are capable of producing audit trails of all the actions performed by the user. This sequence of events can be used to gain insight about how medical students process patient information. In this preliminary analysis, we made use of transition matrices to model the sequence of events performed by the users. We then compared the transition matrices of the experts and the students. Differences can be seen depending on the specific patient diagnosis. Analysis of audit trails in electronic health records can provide medical educators with new insights in the development of clinical reasoning in their students.

Keywords: sequence analysis, electronic health records, transition matrix

1 BACKGROUND

Patient simulation has become a mainstay of medical education. However, the use of simulated patients within electronic health records (EHR's) for medical education rather than for specific EHR training is more recent (March et al., 2016). This kind of patient simulation can be used to model patient diagnosis and management and to assess the development of clinical reasoning. In addition, these systems are capable of logging audit trails, allowing for events to be replayed or reviewed when needed. When used as an adjunct to clinical exercises, there is great potential of gaining insight on how students learn complex medical tasks. Several approaches can be utilized to analyze the sequence of interactions performed by the students in the system. One is through Markov chains and some visualization tools as done by Ozkaynak et al. (2015). In another study, clinical events have been used to build predictive models to identify which event is more likely to happen based on prior actions (Choi, et al., 2016). However, to the best of our knowledge, exploring how medical students use EHR's for specific patient care tasks has not been previously done. This opens a new area of learning analytics within medical education. This work explores the learning traces of medical students performing clinical diagnoses based on EHR operations.

2 METHOD

Medical students were asked to complete a pre-rounding exercise on an intensive care unit patient, reviewing all data about a fictional patient admitted overnight. They were then expected to be able to determine the patient's status and plan further management. In this analysis, we looked at a class which had students exploring the system with 6 different fictional patients. Furthermore, the

instructor of the class (who was not previously familiar with the specific patients) performed the same set of activities. We classified the instructor as expert. The sequence of logs captured by the system were extracted. In this initial analysis, we wanted to compare how the expert differed from the students who used the system. Students were not naïve to the system. However, they were given simulated patients whose complexity they had not previously encountered. The objective of the exercise was for students to demonstrate their clinical reasoning skill, which is an advanced skill that is difficult to assess. A total of 43 system events were identified. These events were grouped into specific areas where the event took place. Eight event areas were identified by the system, namely Clinical Notes, Documents, Flowsheets, Medications, Orders, Patient Clinical Info, Patient Demographics, and Problems. For simplicity, we used these as states in building the transition matrix. The transition matrix (Figure 1) is obtained by computing the normalized frequency of a state transition (A->B) in a sequence. This quantifies the probability of transitioning from a current state (A) to another state (B). The transition matrix for both the expert and the student groups were computed. To be able to compare the behavior of the expert and the novice students, we obtained the average of the matrices of the 6 student groups.

	Clinical Notes	Documents	Flowsheets	Medications	Orders	Patient Clinical Info	Patient Demographics	Problems
Clinical Notes	0.03	0.00	0.00	0.00	0.06	0.00	0.00	0.00
Documents	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Flowsheets	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
Medications	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Orders	0.00	0.00	0.03	0.00	0.06	0.11	0.03	0.00
Patient Clinical Info	0.06	0.00	0.00	0.00	0.11	0.43	0.03	0.00
Patient Demographics	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
Problems	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 1: Transition matrix of an expert for a certain patient visualized using a heatmap

3 DISCUSSION, LIMITATIONS AND FUTURE WORK

In this work, we attempted to model the sequences of actions using a transition matrix. We then compared the transition matrix of the expert and the students for the 6 different patients. Interestingly, there were different trends that emerged. For certain patients, the expert would refer to the Clinical Notes more often than the Patient Clinical Info. For some patients, it is the opposite. This raises questions why such phenomenon exists. It could be due to the initial impression of the user on the patient or it could be due to the circumstance of the patient being observed. A further analysis on this approach should be done.

The current analysis only looked at the sequences of a few students. Some state transitions may not have been captured. To be able to come up with a reliable result, more students should be considered. Another is the limited number of expert users. More experts should be considered in the study to validate the result. Their agreement must be measured. As mentioned earlier, assessing clinical reasoning skill requires an expert to assess based on the presentation of the patient and/or the plan. Our approach could potentially provide an objective view of where potential deficiencies are. Finally, in terms of educational implication, the ability to identify a sequence of actions as effective or not would be beneficial to the learning experience of the students. It could allow us to design certain personalized intervention to help students who may be struggling.

REFERENCES

- Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016, December). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference* (pp. 301-318).
- March, C. A., Scholl, G., Dversdal, R. K., Richards, M., Wilson, L. M., Mohan, V., & Gold, J. A. (2016). Use of electronic health record simulation to understand the accuracy of intern progress notes. *Journal of graduate medical education*, 8(2), 237-240.
- Ozkaynak, M., Dziadkowiec, O., Mistry, R., Callahan, T., He, Z., Deakyne, S., & Tham, E. (2015). Characterizing workflow for pediatric asthma patients in emergency departments using electronic health records. *Journal of biomedical informatics*, 57, 386-398.

Development of a Visualization Tool for Student Characterization using Mobility Data

Minh Hieu Nguyen¹, Hyoungjoon Lim², Sung Bum Ju³, Joon Heo^{4*}

^{1,2,3,4} Yonsei University, Seoul, South Korea

{nguyenminhhieu, joony729, yunsb33, jheo}@yonsei.ac.kr

ABSTRACT: In Yonsei University, undergraduate students whom are issued with a student card are notified by the school that their card-entrance transactions including accessing the services and buildings will be collected in order to improve education service to students. For the privacy, the student IDs are displayed in an anonymous form. By this agreement, the data center have received an average 20,000 card transactions of 4,000+ students per day. This large amount of data can be helpful for academic performance analysis if links between students' behavior and their learning outcome is indicated. Based on the results of the study by Park et al. (2015), this paper presents the development of a visualization tool for not only the objective as mentioned but also for the characterization of other objects from other datasets. Initially, through the Temporal View of this tool, we could see the different patterns between student groups categorized by academic performance.

Keywords: Visualization Tool, Smart Card Data, Student Trajectory, Big Data.

1 MOTIVATION

Spatial data has been also considered as a potential data in educational data mining. In modern campuses, log data are generated when students use their electronic card to access services and buildings within the school area. This data is a kind of mobility data which contains building's location, contextual information, and access time. "Does regular visit to library help improve students' performance?", "Is there any link between mobility habit and learning outcome of the student?", and many other questions can be given by the educators. However, answering these questions can take a long process while the data could be collected automatically, and the combination of statistic and visualization is very promising to address these questions (Wu et al. 2016). This research does not merely satisfy the questions as mentioned but also is a part of the education innovation project in Yonsei University, South Korea.

2 THE VISUALIZATION TOOL

In this tool, the view types are designed to express students' behavior based on three statistic methods: 1) The temporal statistics (at a certain location, "what did the object do?", "for how long or for how often?". This group view is named Temporal View); 2) The spatial statistics (within a certain period of time, "where did the object appear?", "in what frequency or in what order?". This group view is named Location View); 3) Multi-criteria statistics (this is a mix of contextual information, e.g. a criterion can be created based on an assessment of frequency "F" of action "A" of the object at location "P", for a period of time "T"). The views are designed to enable display the results in parallel or stack, which are convenient to recognize different patterns between student groups. When the context is complex and difficult to express within a single view (e.g. in the library area, students can

enter there for many purposes such as lending books, using public computers, or using seminar rooms), the comparison is further clarified through the Chart View including commonly used charts such as “column chart”, “scatter chart” or “radar chart”. In a limited space of the article, Temporal View is selected to clarify the difference between the two groups of students which are G1 (High GPA) and G2 (Low GPA) in the two groups of location (Dormitory and Library). Temporal View is designed to visualize the frequency of the appearance in a given period time and a given location. The statistical values have been averaged based on the number of students per group, normalized to [0-1], and visualized corresponding to the color intensity. This view can visualize data at different levels of detail (year, month, week, and hour), some specific statements could be drawn as follows.

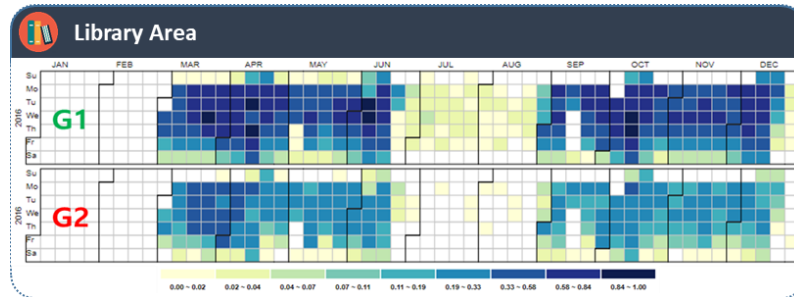


Figure 1: Temporal view at large scale (year, month, week, and day)

In Figure 1, the group G1 went to the library more frequently than the group G2 even it was vacation time (July and August). Through Figure 2, the group G2 did not show their appearance in the library area too late according to the survey time (5 pm - 12 pm), however, they showed their unusual patterns in the dormitory area (2am-7am). In the dorm area, the students only use their card to access the entrances of the building and their bedroom. In a time frame that students should take for rest, however, the group G2 showed that their travel is more frequently compared to the group G1.

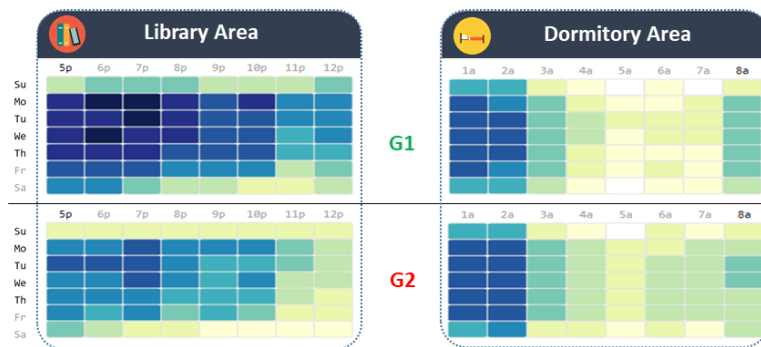


Figure 2: Temporal view at small scale (week, day, and hour)

REFERENCES

- Park, Y., & Jo, I. H. (2015). Development of the Learning Analytics Dashboard to Support Students' Learning Performance. *J. UCS*, 21(1), pp. 110-133.
- Wu, Y., Gong, R., Cao, Y., Wen, C., Teng, Z., & Pu, J. (2016). eduCircle: Visualizing Spatial Temporal Features of Student Performance from Campus Activity and Consumption Data. In *International Conference on Cooperative Design, Visualization and Engineering*, Springer, Cham, pp. 313-321.

Learning Activity Analytics across Courses

Atsushi Shimada, Takuro Owatari, Tsubasa Minematsu, Rin-ichiro Taniguchi

Kyushu University, Japan
atsushi@ait.kyushu-u.ac.jp

ABSTRACT: In this paper, we focus on e-Book operation logs across courses conducted by different teachers. The target courses are conducted by different teachers using the same syllabus, course design, and lecture materials. More than 1,300 students are assigned to one of ten courses taught by different teachers. We extract learning activities and quiz scores from each course. Statistical summaries of e-book operations, the browsing time for each page, and the distribution of the quiz scores for each lecture are analyzed to gather the characteristics of the courses.

Keywords: Learning activity analysis, e-Book logs, analytics across courses

1 INTRODUCTION

Due to the widespread use of digital learning environments in education, collecting large-scale educational data has become easier in recent years. In this paper, we focus on e-Book operation logs across courses conducted by different teachers, and perform learning analytics to investigate the following research question: “Are learning activities common among courses or characterized by each individual course?”.

2 METHOD

The dataset used in this study was collected from e-Learning and e-Book systems (Ogata 2015). The target courses were a series of lectures that constitutes the “Primary Course of Cyber Security,” which commenced in university of Kyushu University in April 2018. Overall, 1,354 students were mechanically assigned to one of the 10 courses in advance. The lectures were conducted by six teachers (four teachers were assigned to give two lectures) in face-to-face style. Teachers followed the same syllabus and used the same lecture materials in the courses.

First, we extract learning activities and quiz scores from each course. Statistical summaries of e-book operations, the browsing time for each page, and the distribution of the quiz scores for each lecture are analyzed to gather the characteristics of the courses.

Second, we perform similarity analysis among courses. For the investigation, we calculate a learning activity feature for each student in each course. The feature vector consists of page-wise features including how long the student browsed each page, how often he/she utilized the operations of bookmark, highlight and memo in each page. Due to the page limitation, the detailed representation cannot be explained in this paper, but the feature vector becomes high dimensional features so that we apply t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten 2008), which converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler

divergence between the joint probabilities of low-dimensional embedding and high-dimensional data. We investigate the similarity and dissimilarity of feature vectors within the course and among the courses through the visualization of feature vectors.

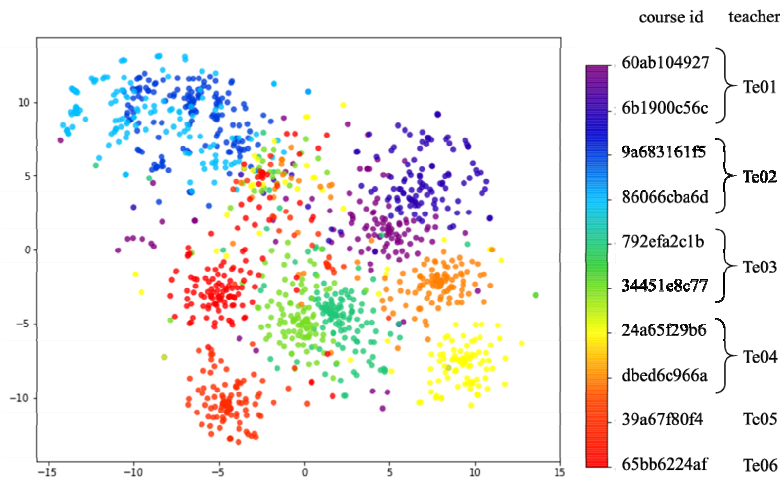


Figure 1: Visualization of feature vectors by t-SNE

3 RESULTS

Figure 1 shows the visualization result of feature vectors in two-dimensional space. Courses are marked by color. From the micro perspective, we can see that the feature vectors distribute closely in the same course, while those of other courses make distinguishable clusters. On the other hand, quiz scores differ among students who belong to the same course. Therefore, if a good match between a learning activity and a course is realized, the learning performance would become better, resulting in better quiz scores.

In our future work, we will conduct further analytics of learning behaviors and teaching behaviors to investigate the successive research question: “Does better matching students and teachers improve students' performance?”.

We will tackle an optimization issue to find better match between students and teachers using learning activity logs. Besides, we will demonstrate the effectiveness of better matching via simulation experiments as the first step, followed by the realistic experiments.

ACKNOWLEDGEMENT

This work was supported by JST PRESTO Grant Number JPMJPR1505, and JSPS KAKENHI Grand Number JP18H04125, Japan.

REFERENCES

- Maaten, L., Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, pp.2579-2605.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., Yamada, M. (2015). E-Book-based learning analytics in university education. *Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015)* pp.401-406.

Determining reading comprehension of domain texts

David Quigley*, Donna Caccamise, Peter Foltz, Eileen Kintsch, John Weatherley, Holly Kurtz

University of Colorado Boulder - Institute of Cognitive Science

*david.quigley@colorado.edu

ABSTRACT: This poster presents a new approach to real-time measurement of reading comprehension of expository texts in the classroom. Our approach combines traditional comprehension questions with a temporal clickstream analysis to build an understanding of both the student's current comprehension of a text along with their experiences with reading supports over the course of the unit.

Keywords: Reading Supports, Comprehension, Clickstream Analysis

1 INTRODUCTION

While great national and international effort has gone into supporting students' reading skills (National Assessment of Education Progress, 2013; e.g. Likens et. al. 2018), a significant gap exists in supporting the reading of domain-specific text. According to the comprehension-integration (CI) model of text comprehension (Kintsch, 1998), successful readers form a representation of text content through coherence building process at multiple levels, notably the local level, macro level and situation model. These mental operations have been implemented in a curriculum (BRAVO, Caccamise et al, 2014) designed to teach students how to deeply understand what they read. eBRAVO represents an expansion of this approach to an e-reader with adaptive, individualized reading comprehension supports. As students complete activities within eBRAVO, the system measures for deep comprehension of the text and determines which reading skill lessons the system should present to the user. measuring deep comprehension

2 APPROACH

Our approach to measuring deep comprehension uses a parallel approach of traditional questions and clickstream analytics. We then use this information to categorize the student on a level of deep comprehension to determine the need for additional reading support. These two approaches work together to draw on information not previously available for real-time systems to recommend interventions targeted at various aspects of reading comprehension from the CI model.

2.1 Comprehension Questions

Our approach builds on the long history of prior work measuring reading comprehension using traditional student questions (Kintsch, 2005). Our efforts emphasize the need for short questions to implement analyses, primarily focused on a series of multiple-choice questions that are mapped to the different areas of the CI model (local cohesion, global cohesion, inference-making, and integration). These formats allow students to engage with the comprehension questions in class,

directly in response to the reading. This approach provides rich details related to which aspect of comprehension the student may need to explore further using eBRAVO's guided instruction.

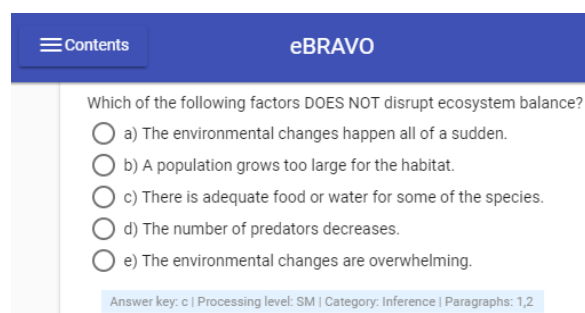


Figure 1: The teacher view of a reading comprehension question in eBRAVO

2.2 Clickstream Analytics

Our analysis also draws from extensive work in the LA community (e.g. Quigley et al, 2017) to extract behaviors from a user's clickstream. These patterns include individual action information, such as the time spent reading the source text, as well as temporal information such as the student's action sequence through the system as well as their engagement with interventions in previous chapters.

2.3 Comprehension Prediction

Similar to other work in the field (e.g. Quigley et al, 2017), the clickstreams and student responses are fed to a predictive algorithm to classify students according to their understanding. We use both streams of information to counteract the shortcomings of each approach, such as guessing at questions and overfitting clickstream data. This approach uses an iterative process, building our models of deep comprehension by comparing active users to previous students.

REFERENCES

- Caccamise, D., Friend, A., Groneman, C., Littrell-Baez, M., & Kintsch, E. (2014). Teaching struggling middle school readers to comprehend informational text. Boulder, CO: International Society of the Learning Sciences.
- Kintsch, E., 2005. Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, 25 (1), 48-61
- Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.
- Likens A.D., McCarthy, K.S., Allen, L.K., and McNamara, D.S. 2018. Recurrence quantification analysis as a method for studying text comprehension dynamics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18)*. ACM, New York, NY, USA, 111-120.
- National Center for Education Statistics. 2013. *The Nation's Report Card: A First Look: 2013 Mathematics and Reading* (NCES 2014-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.
- Quigley, D., Ostwald, J., & Sumner, T. 2017. Scientific modeling: using learning analytics to examine student practices and classroom variation. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)* pp. 329-338. ACM

A Platform for Image Recommendation in Foreign Word Learning

Mohammad Nehal Hasnine¹, Brendan Flanagan¹, Masatoshi Ishikawa², Hiroaki Ogata¹,
Kousuke Mouri³, and Keiichi Kaneko³

Kyoto University¹, Tokyo Seitoku University², Tokyo University of Agriculture and
Technology³

hasnine.mohammadnehal.5z@kyoto-u.ac.jp

ABSTRACT: This paper introduces a platform for image recommendation that can be used in informal learning of foreign words. The platform is based on a distributional semantics model (DSM) that is designed to recommend Feature-based Context-specific Appropriate Images (FCAIs) for representing a word. This technology is for a context-aware ubiquitous learning system that captures ubiquitous learning logs from various learning scenarios. This paper briefly discusses the data capturing tool, methods of employing learning analytics for ubiquitous learning logs analysis, natural language processing techniques applied for word-bank creation, and image embedding methods employed for feature analysis, development of an algorithm that determines the most appropriate FCAI images, and related scientific issues.

Keywords: Image recommendation, lifelogs analytics, ubiquitous learning, word learning

1 INTRODUCTION

Informal vocabulary learning tools such as duolingo, Rosetta stones, VoLT, Rakuten's lingvist etc. on both web and mobile platforms are gaining much popularity among motivated language learners, particularly those who want to memorize foreign words. One of the technological advancements lack in most of the systems is the recommendation of appropriate images in the right time and right learning context. Unarguably, it is not an easy task because a huge amount of educational big data such as a learner's ethnographic information, study location, time, context, and image information etc. processing is required to determine the most appropriate image to represent a word. The objective of this study is to develop a platform that is capable of recommending Feature-based Context-specific Appropriate Images (FCAIs) (Hasnine et al., 2018) for informal learning of foreign words.

2 THE PLATFORM: ARCHITECTURE AND IMPLIMENTATION PATHWAY

The platform is based on a Distributional Semantics Model (Hasnine, 2018; Hasnine et al., 2018) that at first quantifies and categorizes the semantic similarities between various educational data. This analysis allows the model to map the relationship between a word and its visual image features, learning context, geographical location, demographic information, time of learning etc. After that, a word's image representation with a reflection of a learner's cultural-association and learning context is analyzed. SCROLL dataset (Ogata et al., 2018) that contains over 1700 foreign language learners' lifelong learning experiences (such as the geolocation information, vocabulary knowledge, quiz, learning context, contextual image information etc.) is analyzed using lifelog analytics. For the analysis, three kinds of educational data are sent to a Learning Record Store (LRS) as xAPI (Experience

API) statements, are as follows: **Profile data**, the profile metrics consists of word, learner's demographic, culture-specific information, time, place, past knowledge level, and image information; **Word-bank**, is created that contains words labeled as noun, adjectives, sentences, phrases etc. by using Mecab and TreeTagger, two NLP-based tools for Parts of Speech (POS) analysis for English and Japan languages, respectively; **Images**, the Inception v3, VGG16, VGG19, and DeepLoc deep architectures are employed for extracting various deep learning features from images. The image sources are SCROLL system, AIVAS image datasets, and Google image search engine. Finally, using AIVAS-IRA algorithms (Hasnine et al., 2016; Hasnine, Ishikawa, Hirai, Miyakoda, & Kaneko, 2017), the most appropriate image(s) to represent a word's most appropriate representative image under a specific learning location and context. Fig.1. displays the architecture of the platform.

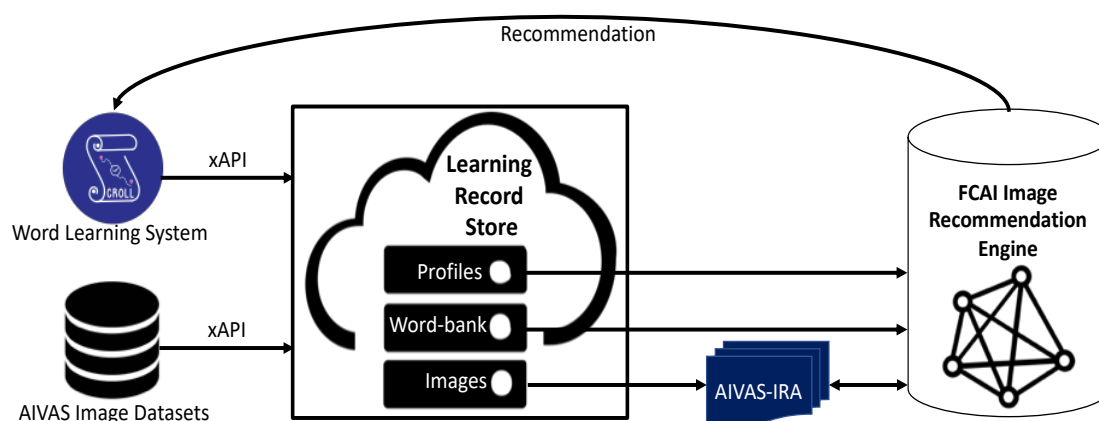


Figure 1. The Architecture of the Platform

ACKNOWLEDGEMENT

This work is supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304 and Start-up Grant-in-Aid Number 18H05745.

REFERENCES

- Hasnine, M. N. (2018). A Distributional Semantics Model for Image Recommendation using Learning Analytics. In Early Career Workshop Proceedings of the 26th International Conference on Computer in Education (pp. 10–12). Manila, Philippines.
- Hasnine, M. N., Hirai, Y., Ishikawa, M., Miyakoda, H., Kaneko, K., & Pemberton, L. (2016). An Image Recommender System that Suggests Appropriate Images in Creation of Self-Learning Items for Abstract Nouns. *International Journal of Management and Applied Science*, 2(5), 38–44.
- Hasnine, M. N., Ishikawa, M., Hirai, Y., Miyakoda, H., & Kaneko, K. (2017). An Algorithm to Evaluate Appropriateness of Still Images for Learning Concrete Nouns of a New Foreign Language. *IEICE Transactions on Information and Systems*, E100.D(9), 2156–2164.
- Hasnine, M. N., Mouri, K., Flanagan, B., Akcapinar, G., Uosaki, N., & Ogata, H. (2018). Image Recommendation for Informal Vocabulary Learning in a Context-aware Learning Environment. In *Proceedings of the 26th International Conference on Computer in Education* (pp. 669–674). Manila, Philippines.
- Ogata, H., Uosaki, N., Mouri, K., Hasnine, M. N., Abou-Khalil, V., & Flanagan, B. (2018). SCROLL Dataset in the Context of Ubiquitous Language Learning. In *Workshop Proceedings of the 26th International Conference on Computer in Education* (pp. 418–423). Manila, Philippines.

Tigris: An Online Workflow Tool for Sharing Educational Data and Analytic Methods

John Stamper, Paulo Carvalho, Steven Moore and Kenneth Koedinger

Carnegie Mellon University

{jstamper, pcarvalh, stevenmo, kk1u}@andrew.cmu.edu

ABSTRACT: This demo will showcase Tigris – an online workflow tool developed as part of the **LearnSphere** project. LearnSphere is a community data infrastructure to support learning improvement online, and brings together a number of data repositories including DataShop (Stamper et al., 2010) and DiscourseDB (Rosé & Ferschke, 2016). Instruction is a data-rich activity — from exams to students’ participation logs. These data can be leveraged to understand instruction and iteratively improve it. However, access to the right tools and how to use them are critical obstacles to this unrealized potential. A Tigris workflow is a component-based process model that can be used to analyze, manipulate and visualize educational data. Using a community based tool repository, educators can quickly build new models, create derivative works, or improve existing tools. Tigris offers a standard set of analysis components which allow researchers to quickly start gathering information about their data and user-contributed workflows and tools to perform other methods of analysis. Tigris enables new opportunities for learning education researchers, course developers, and instructors to better evaluate causal claims, leading to improved teaching and learning. This data-driven course redesign is possible both through better analytics of relational data and through online platform support of controlled experimentation.

Demonstration movie: <https://www.youtube.com/watch?v=p2ecoKBd0q4&t=22s>

Keywords: Learning metrics; data storage and sharing; data-informed learning theories; modeling; data-informed efforts; scalability.

REFERENCES

- Rosé, C. P. & Ferschke, O. (2016). Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses, International Journal of AI in Education, 25th Anniversary Edition, volume 2.
- Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010, June). PSLC DataShop: A data analysis service for the learning science community. In International Conference on Intelligent Tutoring Systems (pp. 455-455). Springer, Berlin, Heidelberg.

Deciphering Dr. Discovery: Data Analytics for Interpreting Museum Visitor Demographics and Engagement with Exhibit Content

Luis E. Pérez Cortés
Arizona State University
luis.perezcortes@asu.edu

Brian C. Nelson
Arizona State University
brian.nelson@asu.edu

Catherine Bowman
Arizona State University
c.bowman@asu.edu

Judd Bowman
Arizona State University
judd.bowman@asu.edu

Brooke Owen
Arizona State University
brooke.owen@asu.edu

Jeff Danas
Arizona State University
jeffrey.danas@asu.edu

Edgar Escalante
Arizona State University
edgar.escalan@gmail.com

Kyle Rogers
Arizona State University
kyle.j.rogers@asu.edu

Abigail Weibel
Arizona State University
abigail.weibel@asu.edu

Jesse Ha
Arizona State University
jesseha@asu.edu

ABSTRACT: The Ask Dr. Discovery study addresses the need for ongoing, large-scale museum evaluation while investigating visitor engagement with museum content. To realize these aims, we developed a mobile app with two parts: 1) a front-end virtual scientist called Dr. Discovery (Dr. D) used by museum visitors that doubles as an unobtrusive data-gatherer and 2) a back-end analytics portal mined by museum staff, evaluators, and researchers. In this poster presentation, we describe the use of data analytics and visualizations gathered from Dr. D to explore visitor characteristics and interpret their engagement with—and movement through—science exhibits at partner museums. We also analyze museum visitor demographics and discuss the implications of their mismatch with the general state population.

Keywords: Museum evaluation, data analytics, informal science education, data-driven decision-making.

1 INTRODUCTION

According to the Pew Research Center (2018), 77% of all American adults own a smartphone. Museums (via apps such as the Smithsonian’s *Infinity of Nations*) as well as researchers (e.g., Bickmore, Vardoulakis, & Schulman, 2013) have quickly seized opportunities provided by the ubiquity of mobile devices. Our project, called “Ask Dr. Discovery” (Dr. D) is a National Science Foundation-funded study aimed at addressing the need for affordable, ongoing, large-scale museum evaluation while investigating innovative ways to encourage museum visitors to engage deeply with museum content. Conducting ongoing museum evaluation is imperative because, while the physical exhibits of museums is often prohibitively costly to change, comprehensive exhibit experiences also depend on flexible elements that can change over time (e.g., docents, multimedia, public events, temporary signage, and webpages). Timely and accessible insight into the minds and behavior of visitors becomes invaluable to improve these elements. Thus, to realize the Dr. D project aims, we designed and developed a mobile app with two parts: 1) a front-end Q&A interface through which visitors can ask questions and receive vetted answers about museum content and 2) a back-end analytics portal that visualizes recorded visitor interactions with the app to be mined by museum staff, evaluators, and researchers. We developed the Dr. D app to function as a platform for research, museum evaluation, STEM informal education, and data-driven decision-making by museum personnel. We describe the data analytics implemented with Dr. D to explore participant-visitor question asking patterns and present data-mined evidence for tracking their movement through respective museum exhibits. Additionally, we seek to understand the visitors of our partner museums by comparing demographic data gathered by Dr. D and exploring how these data compare to the general population as revealed by the state census.

2 DATA SOURCES AND RESULTS

Dr. D generates data via log files containing records of visitor interactions and inquiries (and the path and evolution of those inquiries). Applying techniques for analysis of “big data” from other fields to the Dr. D data can shed light on visitor interests, understandings, and misconceptions. For example, by representing log files as vectors of word frequency in questions, we can employ latent semantic analysis (Deerwester, Dumais, & Harshman, 1990) and closely related principal and

independent component analyses to classify sets of questions into sophisticated broader topics. Dr. D had two conditions: Ask Mode and Game Mode. Ask Mode situates visitor question-asking in a non-game-based version of the app while Game Mode situates visitor question-asking within simple, casual game mechanics.

This study examines data collected from June 2016 to January 2017. Overall, 1,693 questions were asked by participant visitors across both of Dr. D's Game and Ask modes. Overall, the average number of questions per group was 5.1 questions, with Game mode eliciting significantly more questions ($p < .001$) on average than Ask mode (6.9 vs 3.8 questions respectively). We grouped all visitor questions into topics by identifying keywords. These topics are associated with different dedicated sections of each exhibit. We plotted how often participant visitors asked questions related to these topics and correlated it to participants' normalized time-in-exhibit to determine the average path visitors tended to take through the exhibit. Additionally, we compared the demographic data reported by visitor-participants and compared it to the data on the state census. This revealed several key demographics that were either over-represented or under-represented at both partner sites. Whites, for example, represent roughly 55% of the state population, but account for over 60% of all museum visitor groups. Similarly, Latino/Hispanic represent over 30% of the state population, but account for only roughly 15% of all museum visitor groups. While it may not be surprising that museum visitors are not representative of the general state population, it is nonetheless important for museums to empirically gather this information. However, it is difficult for museums to unobtrusively gather demographic information with traditional museum data collection methods. As such, Dr. D might represent one novel way to unobtrusively gather this data. Our results point to three salient implications: 1) the viability of data analytics to better understand museum visitor demographics and their in-museum behavior; 2) the affordances of multimodal methods for collecting and representing evaluative data on visitors; and 3) the viability of an evidence-based approach for museum personnel to better serve their visitor population.

REFERENCES

- Bickmore, T. W., Vardoulakis, L. M. P., & Schulman, D. (2013). Tinker: a relational agent museum guide. *Autonomous agents and multi-agent systems*, 27(2), 254-276.
- Deerwester, S., Dumais, S., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
- Pew Research Center (2018). Mobile Fact Sheet. <http://www.pewinternet.org/fact-sheet/mobile/>. Accessed 7/16/2018/

Exploring Persistence and Regularity Behavioral Analytics in Online Self-Assessments

Cheng-Yu Chung

Arizona State University

Cheng.Yu.Chung@asu.edu

Sharon Hsiao

Arizona State University

Sharon.Hsiao@asu.edu

ABSTRACT: This poster presents exploratory analytics of students' behaviors in an MCQ-based self-assessment platform. We focus on studying persistence and regularity which describe students' voluntary studying behaviors. The persistent behaviors were captured by a Poisson mixture model that revealed four distinct persistence patterns. Self-assessment regularity was measured by an entropy-based score, which discerned regular and irregular study patterns. The experiment results showed that students who were highly persistent and regularly worked on self-assessments achieved better exam performances.

Keywords: self-assessment, regularity, persistence, Poisson mixture model, behavioral analytics

1 INTRODUCTION AND BACKGROUND

Self-assessment tools are designed for students to practice or review the course material along with the regular class content, e.g., lectures, assignments, labs, etc. Since the use of such tools is not required and usually does not count toward the class grade, we believe students' voluntary behaviors in such platform can help us understand their learning conditions and the attitudes toward the class content that cannot be discovered in the analytics of formal assessments. There has been much research showing the benefits of using self-assessments (Bull & Kay, 2007; Hsiao, Sosnovsky, & Brusilovsky, 2010).

In this study we specifically focus on students' studying persistence and regularity in a self-assessment tool. Note we refer *studying persistence* and *studying regularity* to two distinct behavioral analytics in our context, even though these two terms might be used interchangeably in a general context. We define studying persistence as student's continuity of proactive behavior without following a mandatory schedule but his/her own intention to study over a period of time, which fits the context of self-assessment because it is not required or part of the formal class schedule. The study regularity is referred to the degree of repeating a study activity in a certain moment (e.g., in weekends) over a period of time (e.g., four weeks). In other words, study persistence only considers how much a student studies over a period, say it is four weeks. Study regularity considers the amount of everyday study within each week and measures how repetitive such details occur over the four weeks. For instance, a student might use a self-assessment persistently in four weeks, but he/she might not study regularly in every day. In the experiment we

used the Poisson probability model (Park et al., 2018) to model the study persistence and the entropy-based instrument (Boroujeni, Sharma, Kidziński, Lucignano, & Dillenbourg, 2016) to model the study regularity in the log data collected from an MCQ-based self-assessment platform deployed in an undergraduate programming class.

We formulate our research questions as follows: 1) How do students study in an online self-assessment platform, regarding studying persistence and regularity? 2) What are the distinct persistence patterns we can identify by a probabilistic model? 3) How do studying persistence and regularity in the self-assessment relate to the formal exam score?

2 EXPERIMENT RESULTS

We conducted the preliminary experiment in our home-grown MCQ-based self-assessment platform, QuizIT, used in CS undergraduate classes (Alzaid, Trivedi, & Hsiao, 2017). First, we tried to model students' activities in the first four weeks as clickstreams and then used the Poisson mixture model to automatically identify possible persistence patterns. We had found four different behavioral patterns: Active, Inactive, Semi-active, and Cramming. By grouping the students into high-persistence (Active; $\mu = 0.687$, $\sigma = 0.275$) and low-persistence (Inactive, Semi-active, and Cramming; $\mu = 0.532$, $\sigma = 0.280$), the Welch's t-test showed the former group had significantly higher Exam 1 score than the latter one ($p=0.045$ and Cohen's $d=0.551$). To model the regularity, the PWD scores were calculated and in our regression model it was positively correlated to Exam 1 score. To conclude, we have found studying persistence and regularity on the self-assessment platform were positively correlated to the student's performance in the formal assessment. We believe these measurements could also build a bridge to the potential self-regulated learning analytics (Kizilcec, Pérez-Sanagustín, & Maldonado, 2017). The analytics could also help educators assess student's learning strategies quantitatively and provide possible ways to improve their meta-cognitive skills. To make our model close to the existing SRL research, we plan to cross-validate this exploratory work by conventional qualitative measurements of student behavior (e.g., survey) from SRL literature in the future.

REFERENCE

- Alzaid, M., Trivedi, D., & Hsiao, I. H. (2017). The effects of bite-size distributed practices for programming novices. In *Proceedings - Frontiers in Education Conference, FIE*. IEEE.
- Boroujeni, M. S., Sharma, K., Kidziński, Ł., Lucignano, L., & Dillenbourg, P. (2016). How to quantify student's regularity? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI() Open learner modelling framework. *International Journal of Artificial Intelligence in Education*
- Hsiao, I.-H., Sosnovsky, S., & Brusilovsky, P. (2010). Guiding students to the right questions: adaptive navigation support in an E-Learning system for Java programming. *Journal of Computer Assisted Learning*, 26(4), 270–283.
- Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2016.10.001>
- Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., & Warschauer, M. (2018). Understanding Student Procrastination via Mixture Models. In *Educational Data Mining*.

Augmenting Authentic Data Science Environments for Learning Analytics

Anant Mittal

School of Information, University of Michigan
anmittal@umich.edu

Christopher Brooks

School of Information, University of Michigan
brooksch@umich.edu

ABSTRACT: Unlike general learning management systems which have used fine-grained trace behaviours to understand learning processes, data science environments use discipline-specific tools such as Project Jupyter (Perez and Granger, 2015). Augmentation of these tools is necessary in order to surface learner activities in ways which might be used for adaptation (Ferguson, 2012). This is analogous to augmenting problem-solving environments for mathematics (Melis and Siekmann, 2004), where domain-specific tools are necessary for understanding learning activity. In this work, we specifically tackle the augmentation of Project Jupyter. We explain the architecture of the environment along with the types of events we are able to collect and frame research questions we aim to answer with this work.

Keywords: Data science education, Jupyter Notebook, Data Mining, Learning Analytics

1 OVERVIEW OF JUPYTER LOGGING

Project Jupyter, the *de facto* standard learning environment for python data science education, allows developers to extend its functionality through extensions (Perez and Granger, 2015; Kluyver et al., 2016). In order to capture fine-grained activity of learners, we have created an event-based schema of meaningful actions within the notebooks, and then created JavaScript-based extensions which record student activities such as cell insertions, cell executions, and cell deletions¹. These extensions log data in JavaScript Object Notation (JSON) and send them to a webservice back-end. The backend APIs use AWS Kinesis Data Streams with Lambda and S3 to provide a serverless endpoint for learning analytics data collection, allowing us to scale to large numbers of learners with minimal infrastructure costs. The extensions are being deployed in Massive Open Online Courses (MOOCs) which use Jupyter as the source for both course assignments and lecture materials.

We capture five kinds of events: when a notebook has been opened, when a notebook has been scrolled within, when a notebook has been saved, when a notebook cell has been executed, and when cell execution has been finished. For each event we capture high level common data, including the course context where the notebook is deployed (e.g. the assignment or weekly lecture context),

¹ A cell in data science education is similar to a stanza in a poem or paragraph in an essay, and tends to encapsulate a single idea or investigation of the student.

an identifier for the student, a timestamp, and metadata of the notebook. We add to this event-specific metadata, such as which cell in a notebook is being manipulated.

2 LEARNING ANALYTICS IN DATA SCIENCE EDUCATION

We aim to mine the granular student activity data we collect, and we intend to focus on tackling multiple overarching concerns in MOOC environments such as student evaluation and a lack of immediate feedback (Hew & Cheung, 2014). Our environment can help instructors and researchers in understanding student's learning behavior and learning outcomes and help them with more active feedback. Specific investigations this infrastructure will help us understand include:

- What are the common student misconceptions in assignments? For instance, with execution cell events we can identify if a student is struggling on a specific question and provide individual feedback, thus reducing student frustration while scaffolding learning with individual help.
- Are students following along with instructional video? Notebooks for all of the videos are available to students, but at the moment it is unclear how they use these notebooks along with video lectures. Through analysis of cell execution timestamps and the clickstream information (e.g. video heartbeat functions), we should be able to determine if students are following along and practicing as they observe the lectures.
- Do students feel more engaged when given immediate feedback? Through program analysis techniques (e.g. source code analysis), we can identify places where we might provide feedback to the students after their cells have been executed, allowing for just-in-time interventions of learning.

As the online education space continues to grow rapidly, institutions need to see learning analytics and educational data mining as a tool to achieve better learning results. For courses (traditional and online) which use Jupyter for assignments, our extensions to the tool can help instructors proactively monitor student performance, identify students at the risk of dropping out, and implement strategies to improve student engagement.

REFERENCES

- Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304-317.
- Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational research review*, 12, 45-58.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87-90).
- Melis, E., & Siekmann, J. (2004, June). Activemath: An intelligent tutoring system for mathematics. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 91-101). Springer, Berlin, Heidelberg.
- Perez, F., & Granger, B. E. (2015). Project Jupyter: Computational narratives as the engine of collaborative data science. Retrieved September, 11, 207.

Critical Thinking Training and Formative Measurement Using Student Questions

Turner Bohlen¹, Carolyn Bickers¹, Linda Elkins-Tanton^{1,2}, James Tanton¹, Calvin Dunwoody¹, Megan Allen¹
Beagle Learning¹, Arizona State University²
turner@beaglelearning.com, carolyn@beaglelearning.com, lelkins@asu.edu, james@beaglelearning.com,
calvin@beaglelearning.com, megan@beaglelearning.com

ABSTRACT: The skills for the new economy, including self-teaching, critical thinking, team work, and problem solving are becoming increasingly important for success in the modern world. New learning analytics are needed to help instructors train these skills in large scale or online classes, and to measure the outcomes of students in these non-content domains. Student questions have been found to relate closely with a number of these skills (Biddulph & Osborne, 1982). We are developing a software system that uses natural language processing to automatically analyze and categorize student questions, giving instructors actionable summaries of student content understanding and critical thinking.

Keywords: critical thinking, 21st-century skills, collaboration, peer feedback, questions, nlp, natural language processing, formative assessment

1 THE VALUE OF QUESTIONS IN LEARNING

Students need a new set of skills to succeed in the new economy, including critical thinking and collaboration. Lecture-based teaching methods are no longer sufficient, but large or online classes, often fall back on these methods due to a lack of time or tools. As a result, only select students in small, in-person classes are receiving the skills training they most need. The ability for students to ask meaningful questions has been recognized as related to key skills including critical thinking (Biddulph & Osborne, 1982). The quality of the questions asked by students also reveals how much they know and how well they learn (e.g., White & Gunstone, 2014).

We have developed a new software system for student question asking with the goals of 1) supporting students in the generation of meaningful questions, 2) supporting instructors in integrating question-asking into their classrooms, and 3) quantitatively measuring improvement in question quality and critical thinking over time. Our software gathers student questions in focused question-asking cycles centered on specific pieces of content, analyzes those questions to measure critical thinking, and categorizes questions based on the topic discussed. The resulting actionable summary allows question-asking activities to be included in classes with hundreds or even thousands of students, and adds a formative measure of critical thinking to help track student progress.

At LAK19, We will share a simple software platform that provides this analysis to instructors. You can see a video demo at <https://youtu.be/N3NHhAF2Ev8>. In addition, we will share an initial design concepts for a future system that extends our analysis to peer feedback and collaboration skills.

REFERENCES

- Biddulph, F., & Osborne, R. (1982). Some issues relating to children's questions and explanations. University of Waikato.
- White, R. T., & Gunstone, R. F. (1992). *Probing understanding*. London. The Falmer Press.

The Role of Learning Analytics in Redefining Nursing Skills for Artificial Intelligence and Robotization in the Healthcare

Gábor Kismihók

Leibniz Information Centre for Science and Technology
Gabor.Kismihok@tib.eu

Martina Hasseler

University of Heidelberg
martina.hasseler@uni-heidelberg.de

ABSTRACT: In our rapidly changing healthcare system, digitalization, e-health and robotization are gaining influence. Due to the existing nurse shortage in Europe, a demand for healthcare, and therewith nurses, will continue to grow, whilst the supply of available nurses is projected to drop. This fact, together with the growing robotization will create a disruption in the provision of health and nursing care. Furthermore, research shows that the confident usage of ICT is still limited within healthcare professions and thus in nursing. Therefore the NursingAI project aims at analyzing and forecasting skills and competencies needed by healthcare professionals for working together with robots and Artificial Intelligence (AI) driven technology. Furthermore, utilizing state of the art Learning Analytics (LA) tools and methods, curricula for local and European nursing education programmes will be updated. This update includes the testing and the evaluation of relevant skill assessment and training prototypes.

Keywords: curriculum development, nursing education, healthcare, robotization

1 NURSING EDUCATION AND THE ROBOTIZATION OF HEALTHCARE

The NursingAI project works towards an assessment and training framework for skills related to AI, robotization, and e-health in the nursing sectors of Europe. This effort is critical, since 1), nurses should be able to understand and work with novel digital technology in order to improve the general quality of care (Clipper, 2018); 2) the current offer of assessment and training methods on AI, robotization, and e-health skills in nursing in Europe is very limited, 3) in order to have significant amount of digitally native nurses in place in 5-10 years' time in Europe, investments and changes in the nursing curricula need to be initiated now.

Up until now, little research has been done on which skills and competencies for digital healthcare are needed or how nurses respond to these new applications (Maalouf, 2018). Healthcare professionals who use Information and Communication Technologies (ICT) complain about the lack of skills and tailored trainings for their needs. Usually nurses learn ICT related skills on the job within their working duties. For that, they see ICT as a nuisance and very time consuming instead of recognizing its benefits. To ensure a seamless integration of AI and robots in various care systems and to foster the effective use of ICT by nurses, it is important to gain insights in the competencies and skills required (McGonigle, 2014). Therefore, our NursingAI project reviews and analyzes existing assessment and teaching methods focusing on AI and robotization. On the basis of this analysis, the

project points out critical skills and sets of skills, what are essential for nurses when working and interacting with AI. Once this is done, a prototype skill assessment and intervention method will be developed and tested. As a result, we expect that both nurses, patients (and the society) will benefit from the outcome, in terms of safety, and reliability of healthcare system.

2 PROJECT OUTCOMES

As it is visible on Figure 1, this work in progress project aims at the following objectives:

1. **NursingAI Assessment and Training Framework.** Through a structured review of academic and practical literature, existing assessment and training methods for skills necessary to understand and work with robots and AI is analyzed. This analysis is transformed into a concrete assessment and training framework for European nursing VET.
2. **NursingAI Assessment Prototype.** Developing prototype assessment(s) for at least one of the critical, abovementioned skills.
3. **NursingAI Training Prototype.** Developing a prototype intervention, involving training of nurses and nursing students in relation to AI and robotization described by the framework.
4. **NursingAI Learning Analytics.** Assessment and training will be monitored in order to provide feedback to learners. Analysis will also suggest curriculum changes for nursing VET.

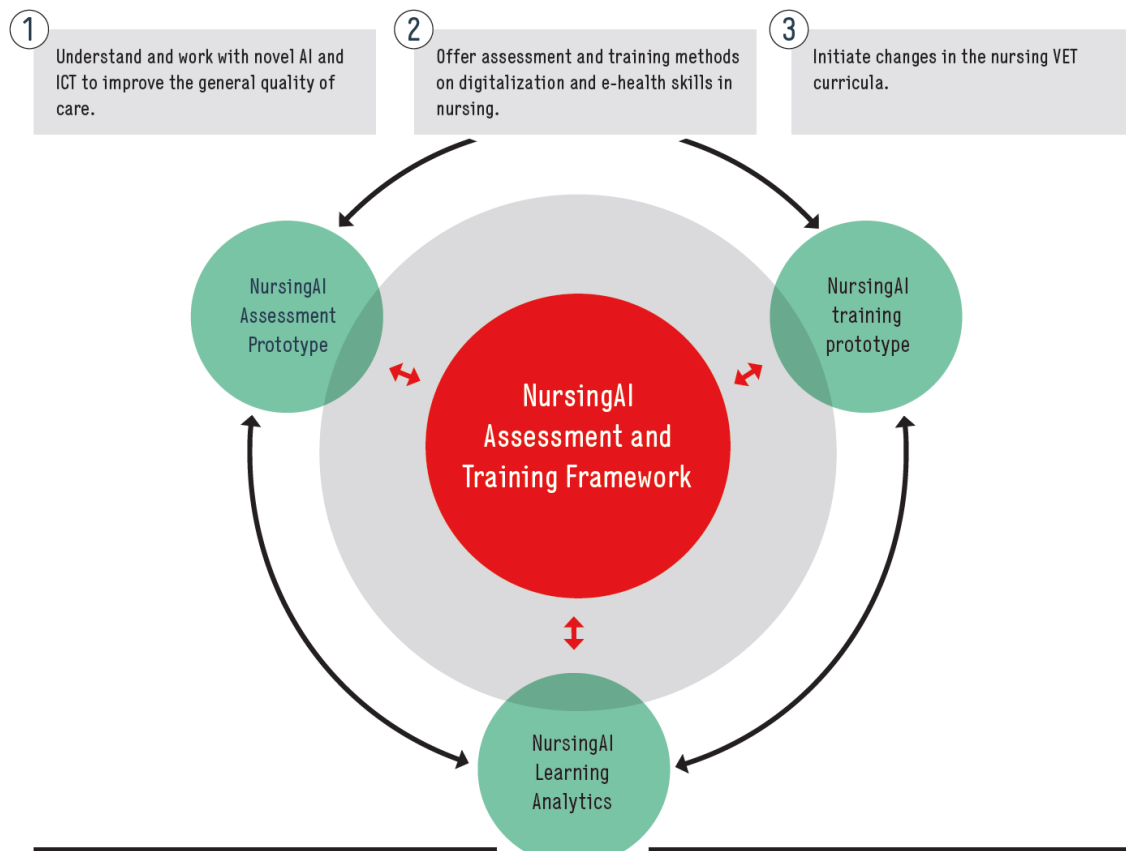


Figure 1. Assessment and Training Framework

REFERENCES

- Clipper, B., Batcheller, J., Thomaz, A. L., & Rozga, A. (2018). Artificial Intelligence and Robotics: A Nurse Leader's Primer. *Nurse Leader*, 16(6), 379–384. <https://doi.org/10.1016/j.mnl.2018.07.015>
- Maalouf, N., Sidaoui, A., Elhajj, I. H., & Asmar, D. (2018). Robotics in Nursing: A Scoping Review. *Journal of Nursing Scholarship: An Official Publication of Sigma Theta Tau International Honor Society of Nursing*, 50(6), 590–600. <https://doi.org/10.1111/jnu.12424>
- McGonigle, D., Hunter, K., Sipes, C., & Hebda, T. (2014). Why nurses need to understand nursing informatics. *AORN Journal*, 100(3), 324–327. <https://doi.org/10.1016/j.aorn.2014.06.012>

Accessible Learning Analytics

Mohammed Ibrahim¹

Daniel McSweeney

Geraldine Gray

Technological University Dublin, Ireland

¹B00108128@student.itb.ie

ABSTRACT: Although there is wide agreement on the benefits of Learning Analytics (LA), many institutes still struggle to operationalize their LA Strategy across campus for a number of reasons. Many stakeholders in the learning process generate data, but are unsure of how to derive actionable intelligence from their data. To maximise the benefit from analysis, data needs to be easily converted to well formatted, visualised output that is accessible and meaningful. This poster outlines progress made towards an Accessible Learning Analytics environment based on Moodle data readily available to faculty, with the main goal of enabling faculty apply LA, moving LA toward more participatory and inclusive approaches.

Keywords: Participatory design; do-it-yourself; do-it-together; accessible learning analytics

1 MOTIVATION FOR ACESIBLE LEARNING ANLAYTICS

Two questions recently posed to the learning analytics community summarise the motivation for this project: “What Are We Measuring?” (Conijn et al., 2017); and “How do we enable teacher Do-It-Yourself (DIY) Learning Analytics?” (Jones et al. 2017). Both questions reflect a practice of LA modeling that has been primarily data-driven and somewhat disconnected from theoretical arguments and reasoning. They direct attention first to existing gaps, then to considering solutions in a context of Accessible Learning Analytics. There is a requirement to have faculty more engaged with LA, thus connecting analysis of data with the learning environment that generated it. In this project, we examine how to bring learning analytics to the wider audience of practitioners with the potential impact of more insightful analysis of learning data to support evidenced based practice.

2 WHY DO WE NEED YET ANOTHER NEW TOOL?

The focus of this project is higher education institutes using Moodle. Moodle is one of the most used Learning Management System (100,153 registered sites and 17,486,372 courses¹); reporting capabilities available to faculty include activity logs and aggregates by student or activity. Moodle’s recent “Students at risk of dropping out” prediction model (Olivé et al., 2018) (available from V3.4, Nov 2017) is reported to manage a prediction model life cycle, but has a number of limitations. Analysis is available at site level only and cannot be tailored for specific curriculum or courses. Processing power and memory requirements restricts how and when models run, explaining why Moodle Analytics is disabled by default. In addition, their supervised learning models are black-box and give a model accuracy metric only, inhibiting users’ ability to interpret and build on model results. Moodle’s analytics framework can be extended to support new prediction models as a

¹ <https://moodle.net/stats>. Last accessed on: January 22, 2019

Moodle plugin. However, the advanced programming knowledge and understanding of Moodle architecture means enhancements to the core offering will take time, and there is an additional lag for educational institutes to adopt such upgrades. Currently available Moodle Plugins range from ad-hoc reporting to visualisation and prediction. The decision to install a plugin, even if free, is governed by many considerations such as performance, data privacy, security and hosting restrictions. Adding to this, each plugin comes with a different aim and does not offer a complete solution. In our opinion, supporting LA on a Moodle installation via multiple plugins does not meet a goal of accessible and applicable LA, while Moodle's core of basics reporting with some advanced analytics is limited and costly. Thus, a solution that can move analysis of Moodle data from available to accessible, and from accessible to contextualised is desirable.

3 ACCESSIBLE LEARNING ANALYTICS (ALA) FOR MOODLE

While many LA studies have analysed Moodle data, there is a lack of common practices and standards which then cause inconsistency between research conclusions (Conijn et al., 2017), and difficulties in adoption and comparisons. Our proposal² is based on Jupyter Notebook powered by a number of common Python open-source scientific libraries to process, analyze, and visualize student activity data in an interactive web-based environment bundling code, documentation, and results. To ensure ease of use, the widely-used Pandas library is utilized instead of pure python code or complex SQL queries; hence, those with basic knowledge of programming can make code changes to customize output. We provide complete examples and functions doing the tough tasks from transforming log data and driving engagement metrics through to prediction models. The code can run on any exported Moodle log and grade book, or faculty's own spreadsheet of grades, and with only course start and end dates required as input. The next step is to interview faculty from a range of disciplines who contributed data to the project to evaluate the usability of the proposed tool and so inform our iterative Dashboard design process. Our tool aims to be a valuable contribution in enabling teacher DIY analytics, and should also be helpful to a Learning Analytics Dashboard (LAD) design team helping them ensure realistic and accurate capturing of user requirement in a participatory Do-It-Together (DIT) fashion. It can also facilitate more meaningful, and reproducible research studies to progress our understanding of what we are measuring and modeling.

REFERENCES

- Conijn, R., Snijders, C., Kleingeld, A. and Matzat, U. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), pp.17-29.
- Jones, D., Jones, H., Beer, C., & Lawson, C. (2017). Implications and questions for institutional learning analytics implementation arising from teacher DIY learning analytics. Paper presented at the ALASI 2017: Australian Learning Analytics Summer Institute.
- Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2018). A supervised learning framework for learning management systems. In *Proceedings of the First International Conference on Data Science, E-learning and Information Systems* (p. 18). ACM.

² <https://github.com/csmibrahim/Accessible-Learning-Analytics>.

Towards a Process to Integrate Learning Analytics and Evidence-Centered Design for Game-based Assessment

Yoon Jeon Kim, José A. Ruipérez-Valiente, Philip Tan, Louisa Rosenheck, and Eric Klopfer

Massachusetts Institute of Technology
{yjk7, jruipere, philip, louisa, klopfer}@mit.edu

ABSTRACT: To fully leverage data-driven approaches for measuring learning in complex and interactive game environments, the field needs to develop methods to coherently integrate learning analytics (LA) throughout the design, development, and evaluation processes to overcome the downfalls of a purely data approach. In this paper, we introduce a process that weaves three distinctive disciplines together—assessment science, game design, and learning analytics—for the purpose of creating digital games for educational assessment.

Keywords: Evidence-Centered Design, Learning Analytics, Game Learning

1 INTRODUCTION

Digital games are gaining popularity in educational settings in addition to playing an essential role in young people's everyday lives. To fully understand what young people learn from playing these games and how they do so, the educational research community has developed ways to unobtrusively measure students' learning using log data generated from their engaging and authentic experiences in the game itself (Shute, 2011). This approach, called stealth assessment, promises that robust inferences can be made without interrupting the flow of gameplay, while at the same time reducing learners' anxieties about assessment (Kim & Shute, 2015). Evidence-centered design (ECD) is a framework commonly used by assessment designers to establish coherence across all aspects of stealth assessment development. Because much of the ECD effort is focused on formalizing assessment models (e.g. competency models), this can make the process less iterative. On the other hand, pure data-driven modeling approaches often found in fields like LA can be subject to bias, when they forget the human nature of the field. By integrating LA with ECD, the application of the assessment design process can avoid these pitfalls by intentionally incorporating expert-informed design decisions. We have been applying this process in our game-based assessment project called Shadowspect (see Figure 1 and a [video online](#)), which aims to measure common core geometry standards (e.g. visualize relationships between 2D and 3D objects) and relevant reasoning skills (e.g. spatial reasoning).

2 OVERVIEW OF THE PROCESS

The process describes the steps for game designers and developers to ensure that the assessment needs are well-balanced with the goal of a playful and engaging gameplay experience. Figure 1 on the right illustrates the three iterative phases of this framework—design, development, and evaluation—and shows how the LA, game design, and assessment disciplines inform one another to build coherence across all aspects of game-based assessment. Next, we provide a short description of each phase:

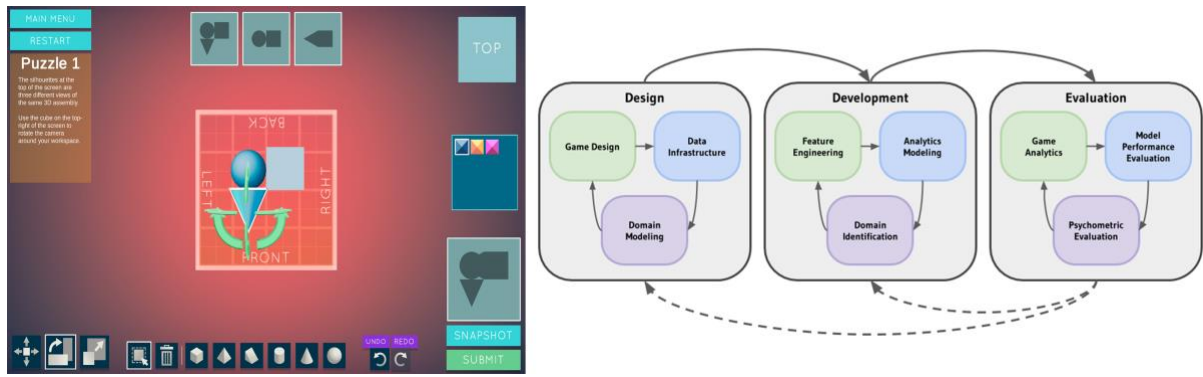


Figure 1. Screen capture of the Shadowspect game and overview of the process.

1. **Design:** The first step for design is to accomplish domain modelling by reviewing existing literature to further define what skills, knowledge, and attributes constitute the competencies that we want to measure. After the game designer has a reasonable understanding of the target competency (i.e. the competency model) and what evidence of that competency might look like (i.e. the evidence model), then the designer is able to explore the design space to come up with game mechanics that are compatible with the assessment mechanic. Additionally, this phase can include design of a data infrastructure that can accommodate the game events, algorithmic and assessment machinery, and reporting tools for the instructor. This can be a challenge due to the open-ended nature of game environments.
2. **Development:** The first step in developing assessment machinery is to work on feature engineering to create variables related to the target competencies. This is a step that entangles a mix of domain and analytics expertise, thus it is good to be carried out by a multidisciplinary team. Then, we use those features as the input of our analytics modeling. Where traditional ECD applies fixed rules based on human experts by tightly controlling the game elements for each target competency, this step can be improved with LA to discover learners' attributes or behaviors that are related to the competencies by generating automatic scoring rules or applying machine learning models. Then, we start mapping which evidence of the game and algorithmic outputs are linked to the domain that we aim to evaluate.
3. **Evaluation:** The last phase is evaluation of the analytic model in terms of both construct validity as well as performance metrics, and then embedding it within the assessment machinery. As part of this process, we also want to invest time in evaluating our game through analytics, so that we can identify game elements that could directly affect the psychometrics of the assessment and further iterate ways to improve player experience (fun) while using that data to identify "random patterns" or "off-task behaviors" that are the product of game design flaws. Finally, once evaluated algorithms and analytics models are incorporated as part of the assessment machinery, we then need to evaluate the psychometric qualities of the overall game-based assessment regarding generalizability, reliability, validity, and fairness (AERA, 1999).

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*
- Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340-356.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2), 503-524.

Elements of Success: Supporting At-risk Student Resilience through Learning Analytics

Jae-Eun Russell

University of Iowa
jae-russell@uiowa.edu

Anna M. Smith

University of Iowa
anna-smith-1@uiowa.edu

ABSTRACT: Learning analytics is a growing field in higher education which aims to increase student success as well as support institutions' retention and graduation efforts. While many of the early learning analytics efforts were intended to help educators and academic advisors identify and intervene with students who are likely to struggle, more recent applications have been designed for students' direct access. In particular, student-facing dashboards, which visually display data derived from educational technologies, have become prevalent. Although some studies have reported overall positive associations between student use of dashboards and learning outcomes, little is known about the effects of learning analytics dashboards on at-risk students, particularly when a predicted course grade is displayed. This study examined the relationship between at-risk students' use of a learning analytics platform and the risk of withdrawal and course achievement. Results indicated that viewing performance feedback including an undesirable predicted final grade did not increase the likelihood that an at-risk student would withdraw from the class, and furthermore, may have encouraged students to take actions to improve their standing in the course.

Keywords: Learning Analytics, Predicted grade, At-risk students, Self-regulated learning

1 INTRODUCTION

Although some research studies have reported positive associations between the use of learning analytics dashboards and student success (e.g., Van Horne et al., 2018), little is known about the effects of learning analytics dashboards on at-risk students' learning outcomes, particularly when a predicted course grade is displayed. In fact, there is concern that exposure to negative performance feedback such as a low predicted course grade could discourage at-risk students (Teasley, 2017). The goal of this study was to examine the relationship between at-risk students' use of a learning analytics platform, Elements of Success, and their subsequent risk of withdrawal and course achievement.

2 ELEMENTS OF SUCCESS

Elements of Success (EoS) is a learning analytics platform that provides visual performance feedback in each grade category, weekly progress updates, and a predicted final grade in the course after each major benchmark (e.g., exams). Through EoS, students also receive elaborated feedback based on their performance from their instructors.

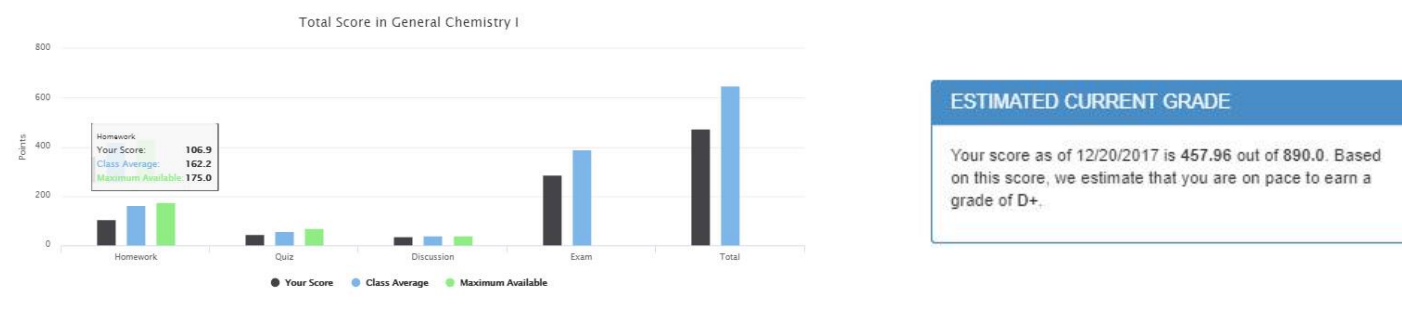


Figure 1: Student View of Elements of Success

3 DATA AND METHODS

This study examined data from 948 students who received a low predicted final grade after the first exam across two semesters of an introductory chemistry course. Among this group, 677 (71.4%) of students used EoS. We identified two levels of risk based on their predicted grade after the first exam: *high* and *moderate*. High-risk students were estimated to receive a predicted final grade of a D+ or lower, and moderate-risk students were estimated to receive a predicted grade of a C or C-. Our primary interest was comparing withdrawal rates and final grades between at-risk students who did and who did not use EoS. To analyze withdrawal rates, survival analysis with a time-varying covariate for EoS use was used, and logistic regression was employed to model final grades. To control for possible confounders, we included in our model variables likely to be associated with both EoS use and course outcomes including students' prior GPA, early course performance scores (i.e., homework and exam 1), frequency of clicks on EoS and the gradebook, demographics, self-reported study skills, GPA expectations, and resiliency (Grit scale).

4 RESULTS AND CONCLUSIONS

The association between EoS use and course outcomes was dependent on student risk level. For the high-risk group, EoS use was not related to risk of withdrawal, but was positively associated with course outcomes, after controlling for model covariates. In contrast, moderate-risk students who used EoS were significantly less likely to withdraw compared to non-users. However, these students ultimately did not achieve better course outcomes compared to non-users. Overall, these results suggest that, despite concerns, use of EoS may not be detrimental to the perseverance of at-risk students. Despite at-risk students entering the course with lower resiliency (Grit) and prior learning outcomes compared to students not at risk, receiving performance data including an undesirable predicted final grade was not associated with an increased risk of withdrawal. Furthermore, this information may have prompted some students to take actions to improve their standing in the course.

REFERENCES

- Teasley, Stephanie. (2017). Student facing dashboards: One size fits all?. *Technology, Knowledge and Learning*, 22(3), 377-384. <https://doi.org/10.1007/s10758-017-9314-3>
- Van Horne, S., Curran, M., Smith, A., VanBuren, J., Zahrieh, D., Larsen, R., & Miller, R. (2018). Facilitating student success in introductory chemistry with feedback in an online platform. *Technology, Knowledge and Learning*, 23, 21-40. <https://doi.org/10.1007/s10758-017-9341-0>

Promoting college readiness in math with ALEKS: How restudy and learning behaviors relate to enrollment, achievement, and retention

Matthew L. Bernacki

University of North Carolina – Chapel Hill
mlb@unc.edu

Kat Campise

RT Solutions
datasciencewonk@gmail.com

Megan Romero

megan.romero@unlv.edu

William Speer

University of Nevada, Las Vegas

william.speer@unlv.edu

Diane Chase

diane.chase@unlv.edu

ABSTRACT: This **Poster** examines the implications of engagement with the ALEKS intelligent tutor for admitted, aspiring undergraduates (N=296) who initially scored “not ready for college math coursework.” Analyses confirm 10 hours of additional study in ALEKS had large effects on subsequent placement assessment scores. Problem-solving successes and seeking of explanations during study in ALEKS predicted achievement in later undergraduate courses.

Keywords: Intelligent tutors, mathematics, college readiness, STEM retention

1 THEORETICAL FRAMEWORK AND STUDY AIMS

Assessment and Learning in Knowledge Spaces (ALEKS) is a widely used adaptive math problem solving software designed to provide math instruction to a wide range of learners [1]. In line with prior research on the effects of intelligent tutors on math learning [2], a recent meta-analysis found ALEKS conferred modest benefits to classroom learners ($g = .08$) with greater effects in post-secondary settings ($g = .16$ [3]). Universities use ALEKS placement assessments to determine students’ readiness for college level mathematics. The placement, preparation, and learning (PPL) interface assesses proficiency needed to enroll in credit-bearing university math courses and, for students not yet proficient, provides adapted problem-solving lessons to improve their proficiency prior to future placement attempts. Little is known about how not-yet-proficient learners use ALEKS to prepare for second attempts, and whether they benefit from doing so. If aspiring students can indeed achieve proficiency and eligibility to enroll in credit bearing math through brief restudy, universities can adopt ALEKS PPL to speed credit acquisition, and shorten time to degree. The study focuses on events of 296 aspiring college students who earned admission to university, but whose initial placement score confirmed a lack of proficiency needed for credit bearing math. Enrolled students primarily graduated from low-resource high schools. Most are the first in their family to attend college students. Most also identify with ethnic groups not well-represented in higher education.

2 METHODS

We examined trace data from ALEKS sessions after an initial failed placement assessment attempt and before the second assessment attempt. Students (N = 296) graduated high school, failed their initial placement assessment, spent 10 additional hours solving problems in ALEKS prior to a second assessment attempt. We examined (1) improvement in ALEKS assessment scores after 10 hours of

additional study, (2) how traced problem solving events in ALEKS relate to learning outcomes AND (3) whether positive outcomes are associated with elective, self-regulated learning behaviors not prompted by ALEKS: seeking out problem solutions using Explain features and Review mode.

Table 1: Definitions of Variables for the Analysis

Variable	Definition
C vs. W	Correct / Wrong response to the item.
E vs. L	Student looks at an explanation page called by the student / prompted by ALEKS
Review	Student election to access a review mode to review any topics previously mastered.
%Correct	Correct responses divided by total correct + wrong responses to items

3 RESULTS

We compared placement scores on attempts 1 and 2 using a dependent samples t-test and found a significant and large effect of restudy on assessment scores, $t(295) = 31.224$, $p < .001$, $d = 2.29$. Scores rose 29.7 points (SD = 16.3) from 33.1% mastery (SD = 8.6) to 62.8% (SD = 17.3) and of the 296 who completed 10 hours of study and a second assessment, a significantly greater proportion achieved proficiency (i.e. > 46% mastery) and gained eligibility to enroll in credit bearing math than did not, $\chi^2(296) = 109.46$, $p < .001$. Having confirmed that scores increased after engagement with ALEKS, we next examined how learning events during that engagement (i.e., Table 1) associated with six learning outcomes: ALEKS placement attempt 2 score, improvement across attempts, Eligibility to enroll in Credit-bearing math (binary), Fall Math course GPA, Semester GPA, and Spring re-enrollment at the university. Bivariate correlations revealed no association between behaviors and second placement score nor credit-bearing math eligibility. Those who earned higher placement scores answered more items correctly during study ($r[296] = .119$, $p = .04$). Those who engaged with more ALEKS-prompted explanations achieve greater gains ($r[296] = .125$, $p = .03$); the relation between gains and learner prompted views of explanations was weaker and not significant, $r(296) = .065$. The act of seeking explanations was, however, significantly associated with performance in math courses once enrolled, $r(211) = .213$, $p = .002$, and may thus confer benefits at university. Additional results indicate more complex relations wherein performance during learning in ALEKS predicts initial Math GPA differentially by course type (credit-bearing vs. developmental).

4 ACKNOWLEDGEMENTS

This work was undertaken as part of ACAO Digital Fellows Project, which was supported by the Bill & Melinda Gates Foundation.

REFERENCES

- [1] ALEKS. (2017). Overview of ALEKS. Retrieved from https://www.aleks.com/about_aleks/overview
- [2] Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106 (2), 331–347.
- [3] Ying Fang, Zhihong Ren, Xiangen Hu & Arthur C. Graesser (2018): A meta-analysis of the effectiveness of ALEKS on learning, *Educational Psychology*.

(Un)Readiness for College Algebra: Using Learning Analytics to Design Interventions for Student Success

Goutam Sarker

University of Texas at Arlington
gsarker@uta.edu

T. Lisa Berry

University of Texas at Arlington
lberry@uta.edu

Shanna Banda

University of Texas at Arlington
sbanda@uta.edu

George Siemens

University of Texas at Arlington
gsiemens@uta.edu

ABSTRACT: This poster will present the findings from an on-going research study examining course data to investigate the best conditions for student success in order to implement strategies and interventions that support student learning in a College Algebra course at a large, publicly funded university in the U.S. Many college students continue to fail or withdraw from College Algebra courses even though they achieved the prerequisite standards to enroll. The results from this study indicate that a student's state-mandated college-readiness score (TSI), college entrance exam score (ACT), gender, and admission type are not statistically significant predictors of final course grade. A student's score on the readiness exam developed by the mathematics department was a strong predictor for a passing final grade. Future work from this project will make use of productive persistence interventions to target mathematical concepts identified to predict course failure.

Keywords: College Algebra, Learning Analytics, Learning Success, Mathematics, Postsecondary Student Success, Readiness, Remediation

1 INTRODUCTION

In 2013, the National Assessment of Education Progress (NAEP) reported that 61% of 12th grade students were not prepared for college-level mathematics (McClarty, Matttern, & Gaertner, 2018). During the last decade, U.S. policy makers and educators have designated an increasing number of resources to address the costs associated with students who graduate from high school without the requisite skills to succeed in college. College Algebra serves as a gateway course at many institutions and prevents students from matriculating and continuing to graduation (Li, et. al., 2010). In addition to impeding graduation, students need quantitative skills to acquire high-labor-market-value credentials, as these skills are often prerequisites to many professional advancement opportunities (Radford & Horn, 2012).

The purpose of this study is to make use of predictive learning analytics to identify specific course concepts that predict failure and to design productive persistence driven (Edwards & Beattie, 2016) interventions to increase the passing rate. Early identification of barriers not only improves student

outcomes but also decreases education costs to the student, remediation costs to the university, and the costs of developmental education efforts by state and local governments. Since there are many individual traits, characteristics, and interventions that contribute to student success, the research goal is to make them work more reliably for diverse learners in the hands of a diverse group of instructors in diverse contexts. The mathematical concepts that students need in order to be successful in College Algebra are universal and the goal of this study is develop interventions that can be delivered at scale - by any instructor in any location with any number of students.

2 METHOD AND DATA ANALYSIS

For this study, the researchers examined course data from the spring 2018 offerings of College Algebra at a large, public university ($n= 267$ students). The Mathematics Department developed a Readiness test to assess each student's weakness areas in College Algebra. All enrolled students were required to take the 30-question Readiness pre-test at the beginning of the semester, and course instructors assigned a series of homework assignments based on the results. After completion of the homework assignments, the student completed a Readiness post-test, which is the only assignment counted in the final grade calculation. Additionally, the researchers collected student information (SAT score, ACT score, TSI score, Gender, Admission type) from the university's Student Information System (SIS).

3 RESULTS AND DISCUSSION

The overall passing rate for students from all course offerings was 48.9%. Data suggested that the remediation work did have a positive impact on the student's grade; the Readiness post-test score was much more predictive of course success than the Readiness pretest (p value = $7.432e-13$) with a negligible (marginally small) effect size (Cohen's d 0.17). The results of the Backwards Stepwise Logistical Regression revealed that TSI score, ACT score, Gender, Admission type, orientation quiz, and five of the homework scores were not statistically significant predictors for student success. Data analysis revealed that the Readiness exam was the best predictor with a log odd of 0.09 (95% confidence Interval 0.08-0.12). Additionally, the regression results indicated that a subset of 12 of the 30 mathematical concepts in the Readiness exam predicted course failure. The researchers will design interventions targeting these 12 concepts in the next phase of the project. The interventions will follow the principles of productive persistence and be delivered in real time using student feedback software (On Task).

REFERENCES

- Edwards, A.R. & Beattie, R.L. (2016). Promoting student learning and productive persistence in developmental mathematics: Research frameworks informing the Carnegie Pathways, *NADE Digest* 9(1), 30-39
- Li, K., Uvah, J., Amin, R., & Okafor, A. (2010). A study of college readiness in college algebra. *Journal of Mathematical Sciences & Mathematics Education*, 5(1), 52-66.
https://www.researchgate.net/publication/265535647_A_Study_of_College_Readiness_for_College_Algebra
- McClarty, K.L., Mattttern, K.D., & Gaertner, M.N. (Eds.). (2018). *Preparing students for college and careers: Theory, measurement, and education practice* (1st ed.). New York, NY: Taylor & Francis.
- Radford, A.W. & Horn, L. (2012). *An overview of classes taken and credits earned by beginning postsecondary students* (Web Tables, NCES 2013-151rev). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
<http://hdl.voced.edu.au/10707/240874>

CoderBot: AI Chatbot to Support Adaptive Feedback for Programming Courses

David Azcona¹, Enric Moreu¹, I-Han Hsiao², Alan F. Smeaton¹

Insight Centre for Data Analytics, Dublin City University¹

School of Computing, Informatics & Decision Systems Engineering, Arizona State University²

{David.Azcona, Enric.Moreu}@insight-centre.org,

Sharon.Hsiao@asu.edu, Alan.Smeaton@dcu.ie

ABSTRACT:

Conventionally, learning analytics are used to notify students regarding their predicted performance and further resources using email or via a university's Learning Management System. To support students to engage in learning and become more pro-active about their learning, we designed CoderBot. CoderBot, is an Artificial Intelligent Chatbot service deployed on WhatsApp¹ as a coding assistant to support learning of computer programming. CoderBot has been deployed in our University's **Python Programming I** course. Students are able to interact with the assistant and find out the following:

(a) Personalized messages about predicted performance. A Predictive Machine Learning classification model is built by aggregating multiple sources of student data (academics, programming work, and logged interactions with offline and online resources), handcrafting features and extracting patterns of success on the course leveraging Artificial Intelligence techniques. The model is trained with two years of ground-truth data and cross-validated. Predictions are generated weekly for incoming student data. Using the classification probabilities, we divide students into deciles and designed a message for each group.

(b) Recommended material. Students are suggested material such as slides and exercises they might want to check out based on their progression and effort on the course.

(c) Short code snippets. Students can avail of code snippets that showcase functionality such as slicing lists, reading from files or printing arguments. 100+ snippets have been hosted on GitHub's gists, as that website is already optimized for easy code reading on smartphones.

In addition, students can ask for further help from the Lecturer or the University's support services, consult the terms of the project and opt-out at any time. Phone numbers are deleted at the end of the semester. Efforts are now being made to include Natural Language Understanding so students can ask questions in natural language.

Keywords: Computer Science Education, Learning Analytics, Feedback, Predictive Model.

Technologies: Python, Pandas, Numpy, Scipy, Scikit-learn, Whatsapp Wrapper API, Flask, MongoDB, Docker, Selenium, GitHub, GitHub's gists, Google's Phone Validator.

Code: <https://github.com/dazcona/code-assistant>

Video: <https://www.youtube.com/watch?v=9HSLwvVzN8E>

¹ WhatsApp Messenger is a freeware and cross-platform messaging and Voice over IP service owned by Facebook

Exploring Writing Analytics and Postsecondary Success Indicators

Jill Burstein

Educational Testing Service

jburstein@ets.org

Daniel McCaffrey

Educational Testing Service

dmccaffrey@ets.org

Beata Beigman Klebanov

Educational Testing Service

bbeigmanklebanov@ets.org

Guangming Ling

Educational Testing Service

gling@ets.org

Steven Holtzman

Educational Testing Service

sholtzman@ets.org

ABSTRACT: Writing is a challenge and a potential obstacle for students in U.S. 4-year postsecondary institutions lacking prerequisite writing skills. This study aims to address the research question: Is there a relationship between specific features (analytics) in coursework writing and broader success predictors? Knowledge about this relationship could contribute to more immediate personalized learning support for students. To investigate, we collected authentic coursework writing from students enrolled at one of six 4-year colleges. We then extracted natural language processing (NLP) writing features (analytics) from the writing samples and examined relationships between the analytics and college grade point average (GPA). Consistent with Burstein et al. (2017), findings suggest that NLP writing analytics may contribute to college GPA prediction. Our findings imply that real-time NLP writing analytics from authentic coursework writing could be used to efficiently track success and flag potential obstacles during students' college careers.

KEYWORDS: natural language processing, writing analytics, higher education

1 INTRODUCTION

Writing is a challenge and postsecondary students who lack prerequisite writing skills may not persist in U.S. 4-year postsecondary institutions (NCES, 2012). This study aims to address the research question: Is there a relationship between specific features (analytics) in coursework writing and broader success predictors? Knowledge about this relationship could contribute to more immediate personalized learning support for students. Previous work has found statistically-significant relationships between reading comprehension and writing features in postsecondary contexts (Allen et al, 2014). Studies related to reflective writing reveal relationships between reflective writing features, learning, and college success outcomes (Gibson et al., 2017; Beigman Klebanov et al., 2017). Consistent with Burstein et al. (2017), preliminary findings presented here suggest that NLP writing analytics generated from authentic coursework writing assignments are predictors of college GPA. The broader implication is that analytics may be applied to authentic college student and may serve to efficiently contribute to technology for tracking success and obstacles throughout college.

2 METHODS

Participants. Authentic coursework writing was collected from 693 students enrolled in first-year courses who participated across the 2017-18 academic year at 6 4-year postsecondary sites. Writing samples represented 7 academic disciplines across Social Sciences, Humanities and STEM.

Data. Nine-hundred and thirty-two assignments were collected. This analysis represents a *slice* of a larger study. We examine writing submissions from a subset of students (N=369) completing multiple required study tasks.

Table 1. College GPA Writing Analytics Predictors (N=369)

Variable	Standardized Coefficient	p-value	Overall R ²	Increase in R ²
personal reflection	-0.17	0.00	0.27	0.02
vocabulary richness	0.20	0.00	0.28	0.03
vocabulary sophistication	0.18	0.00	0.28	0.03
discourse structure	0.17	0.01	0.26	0.01

Analysis. Thirty-six NLP features were automatically extracted from each writing assignment. Features represented writing construct (e.g. *argumentation, coherence, discourse, grammar, and vocabulary*). Using the NLP feature values, we ran a separate hierarchical linear mixed model analysis that contained: 1) one NLP feature, plus 2) *length* (*square root of number of words in the text*), plus 3) *school site*. Each NLP feature, *length* and *site* were the independent (or predictor) variables, and college GPA was the dependent variable. We control for length to ensure that features are not length proxies, and school to control for *site effects* in GPA.

Results and Discussion. **The Overall R² baseline (length + site-only model) for college GPA is 0.25.** Table 1 shows a subset of models where NLP features were stronger statistically-significant predictors of college GPA, and exceeded the baseline R² (Inc. R²). These features are related to vocabulary such as, personal reflection, and richer and more sophisticated vocabulary use (e.g., use of less common words) and discourse structure. Additional features (not shown) with p-values < 0.05 included *grammar and mechanic errors, coherence, and use of contractions*. Findings imply that real-time NLP writing analytics for authentic coursework writing from college students could be leveraged during students' college careers to track success and flag obstacles.

ACKNOWLEDGMENTS

Research presented in this paper was funded by the Institute of Education Science, U.S. Department of Education, Award Number R305A160115. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of the IES. Many thanks to our partner sites, Bowie State University, Bloomburg University of Pennsylvania, California State University, Fresno, Jacksonville State University, Slippery Rock University, and University of North Carolina, Wilmington. Thanks to Michael Flor, Binod Gywali, Ben Leong, and Maxwell Schwartz for engineering support. Many thanks to our research assistants, Patrick Houghton, Hillary Molloy and Zydrune Mladineo, for managing a complex data collection.

REFERENCES

- Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., & McNamara, D. S. (2014). Reading comprehension components and their relation to writing. *L'Année psychologique*, 114(4), 663-691.
- Burstein, J., McCaffrey, D., Klebanov, B. B., & Ling, G. (2017). Exploring Relationships Between Writing & Broader Outcomes With Automated Writing Evaluation. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 101-108).
- Beigman Klebanov, B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective Writing About the Utility Value of Science as a Tool for Increasing STEM Motivation and Retention—Can AI Help Scale Up?. *International Journal of Artificial Intelligence in Education*, 27(4), 791-818.
- Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback.
- NCES. National Center for Education Statistics (2012). The nation's report card: Writing 2011 (NCES 2012–470). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf>

Growing an Institutional Data Lake into a Community Good

Kevin Hartman

National University of Singapore

khartman@nus.edu.sg

ABSTRACT: During the past two years, the National University of Singapore has worked to centralize institutional, academic and learning analytics data into a single repository. The original intent of the repository was to provide researchers with accessible measures of learning and a source of important covariates. This poster illustrates how, over time, the amount of data in the repository and its use cases have multiplied. As new representations of the data have been released to larger stakeholder groups and the data's entry points have been made more accessible, the ALSET Data Lake has become a critical piece of the university's data management policy, a context for data science course activities, and an anchor for inquiry-based learning in addition to its intended use as a research tool.

Keywords: academic analytics, innovation, participatory design

1 INTRODUCTION

During the past decade, many educational institutions have embarked on individual journeys to create pipelines for transforming student data into actionable insights. The success stories include projects that consolidated institutional and academic analytics into data warehouses (Campbell, DeBlois, & Oblinger, 2007) and mined data logs to create early warning dashboard systems for students and instructors (Duval, 2011).

At the National University of Singapore (NUS), the conversations informing the development of its data and analytics pipeline revolved more around creating a community good with the potential to benefit all students, instructors, and researchers, than creating safety nets for certain types of students or insights for a particular type of instructor. With this goal in mind, NUS created an institute—ALSET—to coordinate the aggregation and use of student data and to facilitate learning sciences research. The ALSET Data Lake (ADL) serves as a reservoir of student data collected from the different touch points students have with the university. Its contents include de-identified admissions information, wi-fi activity, learning activity data, and responses from annual surveys. During the past two years, ALSET has developed an ecosystem around the ADL by first crafting a set of guiding principles and policies for its use and then introducing activities to promote its use. The accompanying poster highlights the development of the ADL and the activities it supports.

2 GUIDING PRINCIPLES AND DEFINING POLICIES

Analytics projects relating to the ADL are reviewed by a panel of university faculty and staff for their adherence to 10 principles. Many of the principles on the list, like maintaining privacy and ensuring uses of learning analytics are non-evaluative in nature, were collected from existing learning analytics evaluation and policy efforts (Slade & Prinsloo, 2013). The most important principle on the list is that ADL projects must not only be focused on learning, they must also benefit learners

(Gašević, Dawson & Siemens, 2015). To balance this principle with maintaining student privacy and the security of the data, an early decision was made to restrict direct access to the ADL to a small set of researchers. The byproduct of this decision was the unintentional marginalization of most of the university community from exploring uses of the ADL.

In past year, the institute worked to make the ADL policies more inclusive by creating a second path for initiating ADL project proposals. These proposals do not need to be focused on answering research questions. They can be about using the existing data to benefit students, proposing new sources of data, or transforming how the data is visualized. The new path accepts proposals from researchers, instructors, and even university students.

3 SPURRING GROWTH AND INNOVATION

In its bid to include the entire university in proposing ADL projects, ALSET also expanded the resources available to potential project teams. A version of the ADL's data catalog, a description of every field from every table in the ADL, is now provided to any individual affiliated with the university who declares an interest in starting a project. Additionally, ALSET developed a synthetic student dataset (SSD) that preserves the real relationships between the fields and tables in the ADL without exposing any real student data. Project proposers can use the SSD to explore the contents of the ADL, validate the syntax of their queries, and develop their applications without ever needing access to the actual data. The SSD has been successfully used in undergraduate courses, ADL onboarding sessions, and is the foundation for an online training program. By increasing the number of people who understand what is in the ADL and how the data fit together, ADL innovations now come from courses teaching data science and software design as well as large community events. As a side benefit, the ALSET's researchers now receive unsolicited emails about new sources of data that may lend themselves for inclusion into the ADL.

4 CONCLUSION

While university administrators manage the ADL and the policies that surround it, and only a small number of researchers will ever have direct access to it, NUS is working to make good on its promise of making the ADL a community good by providing multiple paths for engagement. At this point, everyone on campus is a potential source for generating new data for the ADL or a contributor of new ideas for ADL projects.

REFERENCES

- Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4), 40.
- Duval, E. (2011, February). Attention please!: learning analytics for visualization and recommendation. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 9-17). ACM.
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64-71.
- Slade, S. & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1509–1528.

Multimodal Tutor Builder Kit

Jan Schneider

DIPF | Leibniz Institute for Research and Information in Education
Schneider.Jan@dipf.de

Daniele Di Mitri

Open University of the Netherlands
Daniele.Dimitri@ou.nl

Hendrik Drachsler

DIPF | Leibniz Institute for Research and Information in Education
drachsler@dipf.de

ABSTRACT: Traditional Learning Analytics (LA) focuses on the collection and analysis of the interaction between learners and learning platforms, such as Learning Management Systems (LMS). Human learning, however, is not constrained to the direct interactions with these systems. There is a vast number of scenarios where the learning process can happen. With the use of sensors, it is possible to capture the learning process unobtrusively for learning scenarios that go beyond the direct interaction between learners and LMS. Sensor data is noisy and has low semantic value. Therefore, a multimodal approach towards sensor data is usually needed in order to make sense out of it. In recent years several research prototypes have emerged showing the potential of multimodal sensor data to support learning for a diverse range of learning scenarios. However, the development of these type of prototypes is still very time consuming and therefore expensive. In this demo publication, we present the Multimodal Tutor Builder, a set of tools that allows users to connect generic sensor applications in order to build their customized Multimodal Learning solutions. The Multimodal Tutor builder consists of two different applications: The *Multimodal Learning Hub* (*LearningHub*) and the *Visual Inspection Tool* (VIT).

The *LearningHub* allows users to collect and integrate sensor data from customized sets of sensor applications, in order to create recordings (Multimodal Learning Experiences) of Meaningful Learning Tasks. The MLH also is able to broadcast on real-time rule-based feedback to customized sets of feedback applications.

The Multimodal Learning Experiences consist of a .zip file. The .zip file includes a .JSON for each sensor application and a video file of the recording. The .JSON files contains all the frames recorded by the sensor application.

With the VIT it is possible to analyze the Multimodal Learning Experiences generated by the *LearningHub*. It allows users to annotate segments of the recorded Multimodal Learning Experiences by looking at the recorded video, then Machine Learning Techniques that can be used to automatically make sense out of the annotated Multimodal Data.

The promo video of is available on: <https://www.youtube.com/watch?v=cJOqcUsS8Oo>

Keywords: Multimodal Learning Analytics, Artificial Intelligence, Sensor-based learning support.

Preparing Successful Facilitation: Designing A Teacher Dashboard to Support PBL Classroom Orchestration in A Game-based Learning Environment

Yuxin Chen, Asmalina Saleh, Cindy Hmelo-Silver, Krista Glazewski

Indiana University

Yc58@iu.edu, asmsaleh@iu.edu, chmelosi@iu.edu, glaze@iu.edu

James Lester

North Carolina State University

lester@ncsu.edu

ABSTRACT: In small group computer-supported collaboration, teachers face challenges as they engage in classroom orchestration (Dillenbourg, & Jermann, 2010). These challenges are further compounded when using problem-based learning (PBL) approach to design a game-based learning environment. In this complex learning environment, students learn across different forms of learning activities: individual data collection, collective inquiry, and discussion. Teacher dashboards enable teachers to get access to students learning activity and allow them to provide real-time feedback and appropriate scaffolds. By investigating students' learning actions around a structured PBL whiteboard in an educational game, we identified challenges in collaboration and how to support students' discussion effectively. In this paper, we propose a teacher dashboard design in hopes of informing teacher-oriented learning analytics to advance our understanding of PBL facilitation for group collaboration.

Keywords: classroom orchestration, problem-based learning, teacher dashboard, game-based Learning, learning design

1 INTRODUCTION

Empowered with technologies, teachers have new opportunities but also face challenges with increased demands on how to monitor a class at group level and classroom level. Dillenbourg and Jermann (2010) defined the process of managing students learning as classroom orchestration, which requires teachers to provide, maintain and modify facilitation on the fly. Problem-based learning (PBL) design principles provide facilitating strategies to support group collaboration but has less to say about managing a PBL classroom (Hmelo-Silver, Kapur, & Hamstra, 2018). Similarly, game-based learning environments can provide excellent contexts for PBL (Rowe, Shores, Mott, & Lester, 2011) but depend on pedagogical approaches such as PBL to support productive collaboration and learning. A critical tool to support that teachers are a teacher dashboard. To design teacher dashboards, we utilized Clow's (2012) learning analytics cycle that delineated four steps starting with involving learners, then capturing relevant data, generate metrics, and drive interventions. In this paper, we focused on the second and the third steps to discuss data through the pilot testing and propose a design framework of a teacher dashboard matched with students learning process. We aim to investigate two questions: 1) What are the critical indicators during students collaborative problem-solving process and 2) How might we design a teacher dashboard that present collaborative problem-solving indicators to support successful classroom orchestration?

2 LEARNING DESIGN OF CRYSTAL ISLAND

In this study, 6th graders engage in a story-driven game to investigate why fish in a local hatchery in the Philippines is sick. Four students are assigned to an in-game group to solve the problem. Each student is also assigned a unique narrative by talking to different stakeholders. After meeting stakeholders, students in the game group use the PBL whiteboard to select the evidence from their personal notebooks as either in support or to argue against the different five hypotheses presented in the whiteboard (Figure 1). When there is agreement among members, the note turns green and red when there is disagreement. Orange is the default state of the note and indicates that the information is unevaluated. Students can also remove a counterfactual hypothesis once they reached a group agreement and justified their rationales.

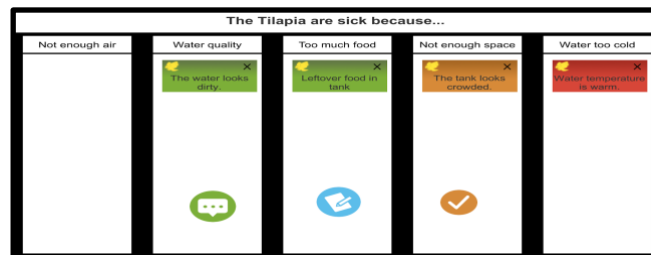


Figure 1. The PBL whiteboard: chat (green icon), notebook (blue icon) and checklist (orange icon)

3 A PROPOSED DESIGN METRICS OF A TEACHER DASHBOARD

Data highlights that students engaged in the PBL inquiry process as they use the whiteboard. Specifically, students shared their ideas, evaluate their peers' ideas and negotiated what pieces of information can be used as evidence to support or reject a hypothesis. The ability for group members and facilitators to see color differences about the salience of a piece of evidence were especially productive in discussions about the viability of hypotheses. To design the teacher dashboard, we have sought to include an overview of student consensus as it relates to each hypothesis. Based on our findings, we believe that this feature can help teachers examine the quality of students' justification and provide sufficient support when there are disagreements, misunderstanding or misconceptions. Below, we proposed several metrics that underlie our teacher dashboard (table 1). We hope our study could provide some insights of designing teacher dashboards to facilitate collaborative problem-solving and to practice their instructional skills in classroom orchestration.

Table 1: A proposed design metrics of a teacher dashboard for Crystal Island: Eco Journey

Learning design	Teacher Dashboard Design	Facilitation and Scaffolding
Deep Content learning: Narrative engagement	Task completion: 1) Individuals' data collection 2) Groups' decision making	<ul style="list-style-type: none"> Overview of learning Alerts for emergent and critical situations Formative assessment on collaboration Enable teachers to provide contingent and effective scaffolds to facilitate group collaboration and problem solving
Collaborative Problem Solving: 1) Structured whiteboard 2) Hypothesis board 3) Real-time chat	Anomaly detection: 1) Inactive and/or students who are lagging behind 2) Anti-social behaviors	
System Support: Virtual agents' prompt	Collaboration process: Substantive and forms of participation	

REFERENCES

- Clow, D. (2012). The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 134-138). ACM.
- Dillenbourg, P., & Jermann, P. (2010). Technology for classroom orchestration. In *New science of learning* (pp. 525-552). Springer, New York, NY.
- Hmelo-Silver, C. E., Kapur, M., & Hamstra, M. (2018). Learning through problem solving. In *International handbook of the learning sciences* (pp. 210-220). Routledge.
- Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21(1-2), 115-133.

Bttn: A Simple Data Collection App for Learning Analytics

Charles Lang

Teachers College, Columbia University
charles.lang@tc.columbia.edu

Xiaoting Kuang

Teachers College, Columbia University
kuang@exchange.tc.columbia.edu

Sai Raj Reddy

Microsoft Research India
t-sared@microsoft.com

ABSTRACT: Demo. In this submission we describe Bttn, a simple mobile data collection app for use in learning analytics education and research data collection. The app consists of a single button that can be assigned with any meaning that individuals or groups would like to collect data about. For example, emotional state (happiness), physical state (tired) observations of the world (temperature). Users can then input measures of the construct they have assigned to the button on a circular visual-analogue scale, the length of time the button is held down the larger the magnitude of the recording. Reminders can be assigned to appear randomly or on a schedule to request users to input data. The primary goal of Bttn is to allow easy collection of data as part of learning analytics education efforts with students collecting data about their own learning that they deem relevant. Bttn may also find use in research settings where preset configurations can be disseminated through the app to research participants. Data is stored as recordings on a ten point scale and is held for ten days on Bttn servers before deletion in which time users can download it for their own use.

Keywords: Data collection, multi-modal analytics, student-driven data collection

1 VIDEO URL

<http://bit.ly/bttnapp>

2 BACKGROUND

We use a visual analogue scale as it has been reported that ratio relationships between phenomena are preserved better by users than with numerical scales (Wewers & Lowe, 1990). We may test numerical scales in later work.

REFERENCES

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, 13(4), 227–236.
<https://doi.org/10.1002/nur.4770130405>

Toward More Meaningful Analytics: Refining Social Presence Within the Community of Inquiry Model

Valerie Barbaro

University of Minnesota – Twin Cities
barb0094@umn.edu

ABSTRACT: The Community of Inquiry model (Garrison, Anderson, & Archer, 2000) is a familiar fixture in online education research, providing a useful template for designing and assessing online learning communities. As the model makes its way into the learning analytics field, developing a more fine-tuned instrument becomes imperative. This poster presents a revised version of the Community of Inquiry model's element of social presence to be used in conjunction with developing learning analytics. With more exact measures of students' social interaction within a course, more precise learning analytics tools can be created. This, in turn, will allow researchers to more accurately correlate social presence to the other presences and to learning outcomes as well as enable online educators to make pedagogical choices that better promote inclusion and success.

Keywords: Community of Inquiry, social presence, online discussion forum, learning analytics

1 INTRODUCTION

As the popularity of online courses continues to grow, so too do the popularity of models for measuring their effectiveness. One framework that has gained significant traction in the 19 years since its introduction is Garrison, Anderson, and Archer's Community of Inquiry model (2000). The model is illustrated by a Venn diagram with three domains: teaching presence, social presence, and cognitive presence. This poster focuses on social presence from the Community of Inquiry model. Social presence, which is comprised of affective, interactive, and cohesive subdivisions, is the ability for individuals to present themselves in a way that allows them to be seen as real people in a virtual environment (Garrison, Anderson, & Archer, 2000).

Considerable attention has been paid to the interaction between the presences (Garrison, Cleveland-Innes, & Fung, 2010) as well as their effect on learning outcomes (Lee 2014; Morueta, Lopez, Gomez, & Harris, 2016), and, more recently, they are seen in learning analytics research (Kovanoić, Gašević, Joksimović, Hatala, & Adesope, 2015; Kovanoic, Joksimović, Gašević, & Hatala, 2014). Despite the model's omnipresence in online education research in the past two decades, though, there has been minimal discussion regarding developing the presences. With a growing body of research touting the importance of social aspects to learning (Garrison, Anderson, & Archer, 2000; Ke, 2010; Lee, 2014; Oztok et al., 2015; Rovai, 2002), more fully understanding levels of social presence becomes increasingly important, especially for generating precise learning analytics.

2 REFINING SOCIAL PRESENCE

This poster proposes low, medium, and high levels to each of the three elements of social presence, which would enable researchers to more precisely measure the presence. The model is not without

such internal measures in the presences. One presence, cognitive presence, includes four states, which suggest stages of critical thinking. This internal breakdown of cognitive presence has enabled more accurate analytics and more accurate assessment of the effectiveness of interventions, such as the use of a nontraditional discussion forum. Social presence, on the other hand, contains no such progression. The three elements of affective, interactive, and cohesive designate types of social presence, but they do not suggest a hierarchy, with one element demonstrating more advanced social presence than another. While this researcher agrees that none of the three elements is “higher” than another, the lack of social presence levels becomes problematic when learning analytics seek to model the presence or to correlate it to the other presences or to learning outcomes—it cannot be applied as neatly as cognitive presence with its four progressive states.

Consider, for example, this reply post: “Thanks for your post, _____. Great job!” This common response typifies posts online educators see all too frequently in discussion forums. Technically, this acknowledgement of another’s work would register as a gesture of interactive social presence. But has a “real” interaction taken place in this simple acknowledgement? As it now stands, the Community of Inquiry model could be used to code such a response as interactive social presence, which would skew analytics results. With nothing to separate such superficial posts from ones that illustrate more authentic social presence, correlating social presence becomes less meaningful.

This poster proposes a breakdown of the elements of social presence, as illustrated in Table 1. Developed through iterative coding cycles that analyzed over 750 student posts, this refined scale can be used in conjunction with tool development to create more precise learning analytics. Results can then be applied pedagogically to promote inclusion and success.

Table 1: Social Presence Refined.

Element	Level	Characteristics
Affective	Low	Low stakes self-exposure with few personal details; non-personal opinion unrelated to course discussion
	Medium	Self-disclosure with some personal elements; use of humor; expressing emotions
	High	Expressing empathy; showing vulnerability, such as sharing a highly personal story
Interactive	Low	Short response that does not elicit further interaction or demonstrate comprehension; agrees/disagrees/compliments but does not elaborate on why
	Medium	Agrees/disagrees/compliments with some explanation; adds own opinion; asks a question; quotes another’s post
	High	Agrees/disagrees with more in-depth explanation; adds own opinion with support; seeks to engage discussion further with questions, elaboration, challenges, etc.
Cohesive	Low	Use of salutations; use of “we” or “us,” but in a general sense
	Medium	Use of “we” or “us” to suggest team mentality; addressing by name
	High	Addressing others’ ideas by specifically referring to them by name in a response that is not directed to them; showing camaraderie

KEY REFERENCE

Garrison, D. R., Anderson, T., & Archer, W. (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, 2(2–3), 87–105. [http://dx.doi.org/10.1016/S1096-7516\(00\)00016-6](http://dx.doi.org/10.1016/S1096-7516(00)00016-6)

Log-based Learning Analytics in Vector Space

Di Sun

Syracuse University

dsun02@syr.edu

Pengfei Xu

Beijing Normal University

xupf@bnu.edu.cn

Junlei Du

Beijing Normal University

junlei007.love@163.com

Qinhua Zheng

Beijing Normal University

zhengqinhua@bnu.edu.cn

Jingjing Zhang

Beijing Normal University

jingjing.zhang@bnu.edu.cn

ABSTRACT: Raw web logs are widely available in web-based learning environments, and are sometimes the only format that curators could provide to analysts. This paper studied the possibilities of performing learning analytics based on raw web logs directly by mapping URLs into a 2D vector space. Experimental results shows that different student groups are easily distinguishable and interpretable in the proposed vector space. It would be interesting to explore better spaces and explore their applications in tasks like student clustering and performance predication in the future.

Keywords: Learning analytics, Educational data mining, Log file

1 INTRODUCTION

Web-based online learning environments often store students' activity in raw web logs. These web logs are generally generated by hosted web servers, each record in web log refers to a HTTP request made by user's browser. Although schemes like MOOCdb (Veeramachaneni et al., 2014) provides much more benefits to analysts, raw web logs are more general and are sometimes the only format that curators could provide to analysts. To perform learning analysis on web logs or convert web logs into schemes like MOOCdb, one needs to map URLs in web logs to students' study behaviors, which is not a trivial task to the best of authors' knowledge. In this paper, we studied the possibilities of performing learning analytics based on raw web logs directly, without inferring students' specific study behaviors from URLs or classifying URLs into groups (Baykan, Henzinger, Marian, & Weber, 2009). Specifically, URLs are mapped into a simple two-dimensional space based

on web log data itself, and the feasibility of performing learning analysis on this two-dimensional space is studied. This simple space is designed heuristically and is by no means an optimized space, however it serves as an exemplar to study the possibility of log-based learning analytics in vector space.

2 PROPOSED METHOD

The proposed method assumes the web log contains at least three fields, which are UID, URL and TIME. UID is an unique identifier points to specific user in the learning environment, URL is an unique resource locator to locate resource in the learning environment, and TIME denotes the timestamp when the specified resource is accessed by the specified user. For each UID, the records are ordered by TIME ascendingly to form a personal list of study records, which will be referred as listuid for simplicity. Each listuid is scanned to calculated the relative position and the time spent for each record. For instance, if listuid contains 100 records, then the first record will have relative position of 0.01 and the 50th record will have relative position 0.5. Time spent for each record is calculated by time difference between this record and the next record, followed by a truncation to deal with special cases. For each listuid, the relative position and time spent values for each record are then grouped and averaged for each URL, which could be denoted as vector $v_{uid,url}$. $v_{uid,url}$ are then averaged over all UIDs to get the final vector for each URL. The final vector is expected to represent the amount of information and the relative position in course for each URL. Figure 1(e) shows a scatter plot of URLs from Chapter 1 to Chapter 13 using proposed method, it could be seen that the course structure is well revealed using proposed method.

3 EXPERIMENTAL RESULTS

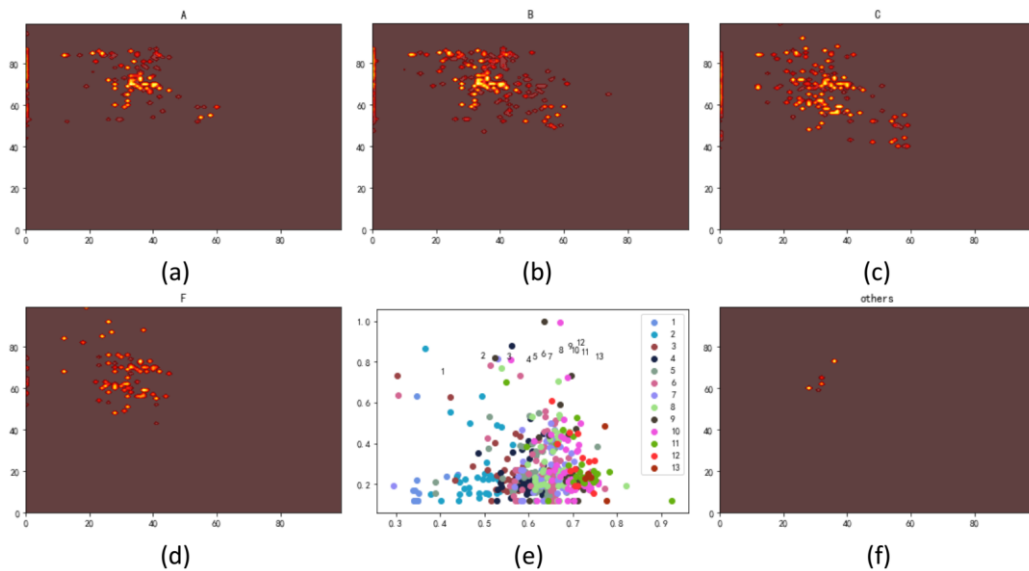


Figure 1: Heat maps for different student groups (a-d, f) and scatter plot of exemplar URLs (e).

In experiments, log data are converted into a 2D space using proposed methods. To illustrate feasibility of perform learning analytics on such vector spaces, heat maps for different student groups are generated in proposed space, as in Figure 1. Figure 1 (a), (b) and (c) corresponds to student groups with grade A, B and F, respectively. These three groups of students clearly have

different learning pattern, with Group A is the most focused on high informative resources, follows by Group B. Figure 1 (d) and (f) corresponds to the student group who registered but did not participate in final exam and guest student group, respectively. These two groups clearly have shorter spans in X-axis, comparing with the three groups above.

4 DISCUSSION AND FUTURE WORK

This paper proposed a method to convert raw web logs into a two-dimensional vector space. In experiments, different student groups are easily distinguishable and interpretable in the proposed vector space. The proposed method and space is only an exemplar to study the possibility of log-based learning analytics in vector space. It would be interesting to explore better spaces and explore their applications in tasks like student clustering and performance predication in the future.

REFERENCES

- Baykan, E., Henzinger, M. R., Marian, L., & Weber, I. (2009). *Purely URL-based topic classification*. Paper presented at the International World Wide Web Conference Poster Track.
- Veeramachaneni, K., Halawa, S., Derroncourt, F., O'Reilly, U. M., Taylor, C., & Do, C. (2014). MOOCdb: Developing Standards and Systems to Support MOOC Data Science. *Computer Science*.

Using Natural Language Processing to Assess Explanation Quality in Retrieval Practice Tasks

Kathryn S. McCarthy¹ & Scott R. Hinze²

¹ Georgia State University, ² Middle Georgia State University
kmccarthy12@gsu.edu; scott.hinze@mga.edu

ABSTRACT: This study explored the potential for automated assessment of students' explanations during retrieval practice. Regression analyses indicate that the linguistic features analyzed by the natural language processing tools Coh-Metrix and CRAT predicted 66% of the variance in the quality of students' retrievals. These findings indicate that both the content and connections in student retrievals are relevant to the quality of the explanation. Limitations and future work will be discussed.

Keywords: Natural language processing; reading comprehension

1 INTRODUCTION

Work in *retrieval practice* indicates that practice tests are more effective for long-term learning than restudying. Further, prompting students to *explain* what they have just read as a practice test leads to additional retention and comprehension. One such study demonstrated that students who wrote higher quality explanations during retrieval scored significantly better on a comprehension test seven days later as compared to students who merely recalled as much as they could (Hinze, Wiley, & Pellegrino, 2013). Despite the fact that open-ended practice tests are more effective than multiple-choice or fill-in-the-blank tests (Hinze & Wiley, 2011), open-ended practice tests are rarely used in classrooms due to the arduousness of providing individualized evaluation and feedback.

Thus, the current study explored if natural language processing (NLP) could be used to automate the assessment of open-ended practice tests (explanatory retrievals). Two tools were selected. The Constructed Response Assessment Tool (CRAT; Crossley et al., 2015), which calculates linguistic and semantic similarities between a source text and a constructed response was selected because it was predicted that good explanations would reflect more of the important content from the source text than poor explanations. Coh-Metrix (McNamara et al., 2014), which evaluates lexical, semantic, and cohesive features of text was selected because discourse comprehension theories assume that a more cohesive explanation is reflective of a more coherent and durable mental model (i.e., deeper comprehension).

2 METHOD

The corpus consisted of 186 retrievals collected from a study in which undergraduates ($n = 62$) read three science texts and then engaged in retrieval of information in each text from memory. Half of the participants were asked to *recall* and the other half were asked to *explain*, providing some variability in the quality of retrieval attempts. Two researchers scored the quality of the retrievals holistically from 1-5, consistent with how instructors typically evaluate open-ended responses ($\gamma_s = .80-.89$; Hinze et al., 2013 Exp. 3).

3 RESULTS

Retrievals were submitted to Coh-Metrix. Linguistic indices with non-normal distributions and those with high multicollinearity ($r > .80$) were removed. Indices that were highly correlated with quality score were retained and submitted to a stepwise regression to determine which were most predictive of the quality score. This yielded two significant linguistic indices: 1) *narrativity* (inversely related) and 2) *givenness*, a measure indicative of cohesion. The same procedure was conducted for measures in CRAT. These analyses revealed two predictors: 1) *lexical sophistication* and, 2) *semantic overlap between the source text and the retrieval*.

Finally, a hierarchical regression was conducted to determine if these linguistic indices predicted human ratings of retrieval quality. The final model, $R = .813$, $R^2 = .66$, accounted for 66% of the variance in the retrieval quality score.

Table 1: Regression analysis predicting human ratings of retrieval quality

Entry	Variables Added	R^2	ΔR^2
Entry 1	Number of Words, Text	.50	.50
Entry 2	Coh-Metrix: Narrativity, LSA Givenness CRAT Indices: Lexical complexity	.55	.06
Entry 3	(AoA), LSA Content Overlap	.66	.11

4 DISCUSSION

This exploratory study demonstrated that a combination of natural language processing tools (Coh-Metrix, CRAT) could be used to reliably predict human ratings of explanation quality in an open-ended retrieval practice. Entering indices of cohesion and content overlap significantly improved model fit, providing support for the notion that the benefits of explanatory retrieval are due not only to an increase in what is remembered, but the way that information is organized in memory.

Automating the evaluation of open-ended practice tests can make tasks like explanatory retrieval practice more amendable to classroom implementation as well as to intelligent tutoring. Given that the quality of these retrievals predicts later test performance, the ability to quickly assess what students know during practice can also serve as a form of formative feedback that instructors can use to provide remediation prior to summative tests. This study serves as an initial proof-of-concept and more will be done to improve scoring accuracy. Future work will also be conducted to replicate and generalize these findings using larger corpora on different topics. We also plan to develop and test feedback messages to help students attend to key words as well as the relations between those key words.

REFERENCES

- Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*, (pp.336-340). Raleigh, NC: International Educational Data Mining Society.
- Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, 19(3), 290-304.
- Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, 69(2), 151-164.
- McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Data analysis and visualization for supporting academic writing and its instruction – the example of Thesis Writer (TW)

Christian Rapp

Zurich University of Applied Sciences

rapp@zhaw.ch

Jakob Ott

Zurich University of Applied Sciences

ottj@zhaw.ch

Peter Kauf

PROGNOSIX AG, Zurich, Switzerland

Peter.Kauf@prognosix.ch

Academic writing and its instruction are increasingly being supported electronically (Allen, Jacovina, & McNamara, 2015). Many such systems work as Software-as-a-Service (SaaS), allowing for fine granular analysis of system usage, text production, and revision via logfiles. [TW](#) is a bilingual (German, English) system supporting the genre of research report writing (IMRaD) (Rapp & Kauf, 2018).

We demonstrate two ways of employing log file analysis for research and learning purposes: (1) Studying writing processes with keyloggers is an established field. Screen recording allows for user system interaction research. TW unobtrusively combines these two aspects in the natural user setting by employing logfile analysis (Dumais, Jeffries, Russell, Tang, & Teevan, 2014). A replay function was implemented with a time slider. Within a web browser, it replays the user's primary system function usage simultaneously with their text production. It is therefore possible to research how text production changes following usage of tutorial or linguistic support functions. (2) Aggregated user data analysis & visualization: An API from TW's database to the R statistics package aggregates and visualizes logfile data, which is displayable in TW. Additional to research purposes, data can be displayed to learners to support their learning processes (Vieira, Parsons, & Byrd, 2018).

Academic writing is difficult to both learn and supervise. Many systems offer tracking and analysis of user-system interaction. One major issue is which data should be analyzed, how, and for whom. We presented two directions – process data and aggregated data. Additional to general research interest, both support more practical goals, by helping to understand the impact of pedagogical interventions for users, and by displaying data relevant to users to improve their learning process. However, ethical and privacy issues have to duly be taken into account. Link to demo video: <https://tube.switch.ch/videos/09bec47a>

Keywords: writing analytics, academic writing, intelligent tutoring systems

REFERENCES

- Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 316-329). New York: The Guildford Press.
- Rapp, C., & Kauf, P. (2018). Scaling Academic Writing Instruction: Evaluation of a Scaffolding Tool (Thesis Writer). *International Journal of Artificial Intelligence in Education*. 28(4), pp 590-615. <https://doi.org/10.1007/s40593-017-0162-z>
- Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding user behavior through log data and analysis. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 349-372). Springer, New York, NY.
- Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122, 119-135.

Determining Learning Pathway Choices Utilizing Process Mining Analysis on Clickstream Data in a Traditional College Course

Authors: Matt Crosslin¹, Nikola Milikic², Igor Jovic², Justin T. Dellinger¹ & Kim Breuer¹

¹University of Texas Arlington, ²University of Belgrade

[matt, jdelling, breuer]@uta.edu, nikola.milikic@gmail.com, jovic.i@me.com

ABSTRACT: This poster presentation will detail preliminary research into the pathways that learners utilize to move through a course when given two modalities to choose from: one that is instructor-led and one that is student-directed. Process Mining Analysis was utilized to examine and cluster clickstream data from an online college-level History course designed with dual modality choices. This poster examines some of the results from different approaches to clustering the available data. The results of this analysis could potentially lead to the creation of predictive artificial intelligence models that can assist learners as they navigate modality choices.

Keywords: Learning pathways, process mining, self-regulated learning

1 SELF-MAPPED LEARNING PATHWAYS

The Self-Mapped Learning Pathways instructional design methodology is a course design process with the goal of allowing learners to develop a personalized pathway throughout a course that has options for instructor-led and student-directed modalities. Learners can change and mix modalities at any point through the duration of the course. To date, this option has mostly been utilized in massive open online courses (Crosslin, 2018). This study seeks to understand how learners navigate these options when they are a part of a traditional 15-week college course. Process Mining analysis was initially utilized to quantitatively document the clickstream artifact evidence of the pathways that learners mapped through the mixture of structured and less-structured options. Additionally, learners completed a survey about their choices that will be analyzed alongside textual data from course forums and assignments to create a qualitative companion to the data analysis results.

1.1 Process Mining Analysis of Clickstream Data

Process mining consists of a set of techniques for analyzing data coming from event logs. Process Mining Analysis was initially chosen because recent research has found it can be helpful in identifying and detecting process patterns in self-regulated learning events (Bannert, Reimann, & Sonnenberg, 2013), patterns in learning behavior (Jovanović, Gašević, Dawson, Pardo, & Mirriahi, 2017), and learning strategies (Saint, Gašević, & Pardo, 2018).

1.1.1 Preliminary Data Analytics Results

Log data from course software was collected and organized by actions (events) that were triggered by a user, as well as metadata about the generated event. The total number of users was 104, while the total number of sessions generated from the dataset was 4784. User sessions were generated based on the 'login' and 'logout' events, but also after 30 minutes of inactivity occurred. Finally,

sequences were generated for each actor. Figure 1 shows a full process model containing all interaction sequences of actor sessions.

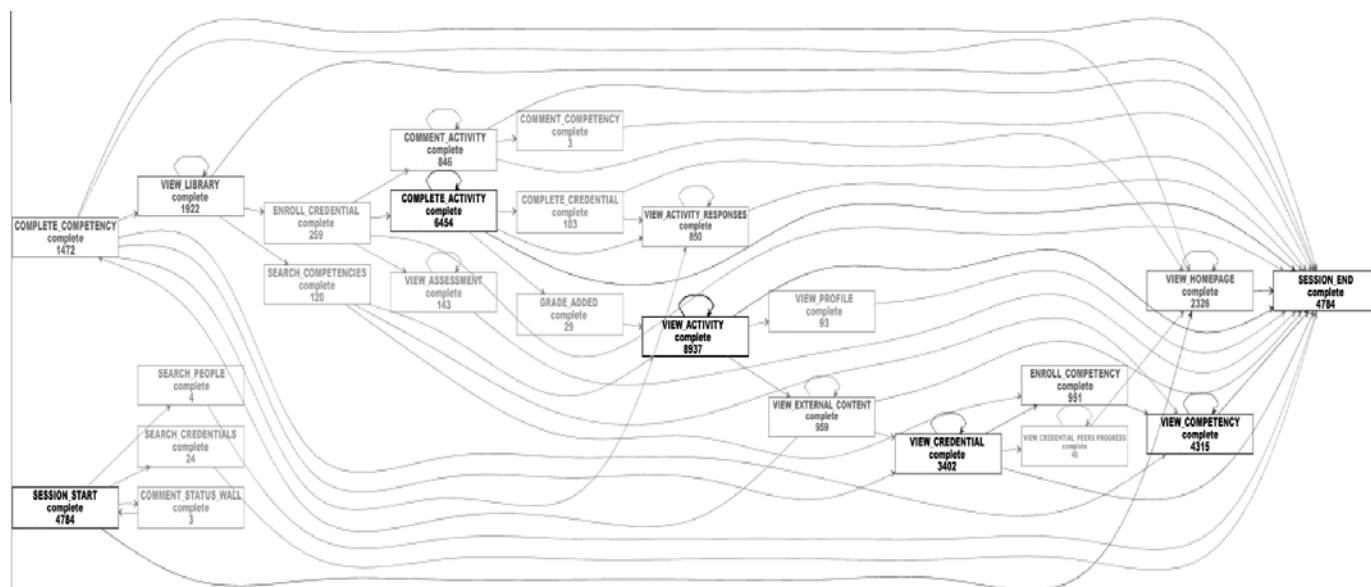


Figure 1: Heuristic net based on 30 minute session data

2 FUTURE WORK

Cluster analysis is being applied to the data to determine what insights and patterns can be gleaned from learner choices. The approach includes applying agglomerative hierarchical clustering algorithm, based on Ward's method, where the similarity measure between learning sequences is based on the optimal matching distance metric (similar to the approach presented in Jovanović et al. (2017)). The algorithm will produce clusters of similar sequences representing different patterns in student behaviors. Once patterns can be established, work can begin on predictive models to help guide learners through modality choices (for example, identifying when learners need instructor help and offering options for joining the instructor-focused modality). Conversely, if no patterns are detected, predictive models can focus on how to guide learners through the full range of choices based on other factors (such as responses to interactive artificial intelligence agents).

REFERENCES

- Bannert, M., Reimann, P., & Sonnenberg, C. (2013). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, 9(2), 161–185. doi:10.1007/s11409-013-9107-6
- Crosslin, M. (2018). Exploring self-regulated learning choices in a customisable learning pathway MOOC. *Australasian Journal of Educational Technology*, 34(1), 131-144.
- Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, 33, 74-85.
- Saint, J., Gašević, D., & Pardo, A. (2018). Detecting learning strategies through process mining. *Lecture Notes in Computer Science*, 385–398. doi:10.1007/978-3-319-98572-5_29

Page-wise Difficulty Level Estimation using e-Book Operation Logs

**Tetsuya Shiino, Atsushi Shimada, Tsubasa Minematsu, Kohei Hatano, Yuta Taniguchi,
Shin'ichi Konomi, Rin-ichiro Taniguchi**
Kyushu University, Japan
shiino@limu.ait.kyushu-u.ac.jp

ABSTRACT: We propose a new approach to analyze the page-wise difficulty of lecture materials. The dataset used in this study was collected from an e-Book system. The e-Book operation logs contain page movement operations as well as learning operations such as bookmark on a page, highlight on keywords. We analyzed a total of 110,894 e-Book operation logs. 6 kinds of page-wise features are calculated from the e-Book operation logs, and 921 pages were evaluated via 10-fold cross validation. Eventually, our method could provide better performance than the chance rate. In this paper, we give our analytics strategy and report primal results.

Keywords: difficulty level estimation, lecture material, page-wise analysis, e-Book logs

1 INTRODUCTION

Analytics of lecture material is important to improve lectures. In particular, if teachers know where students feel difficult in the lecture materials, they can do flexible lecture based on that point. A questionnaire-based approach is one of the realistic approaches to investigate the difficulty of learning materials, but this approach forces students to evaluate the materials in each course. For the purpose of automatic estimation, gaze information and mouse actions are utilized (Nakamura 2008). However, this approach is not suitable for the evaluation involving a large number of students. In this paper, we propose a new approach to analyze the page-wise difficulty of lecture materials. We utilize event stream logs collected from an e-Book system (Ogata 2015). Our approach provides prediction of difficulty for each page of lecture materials to use machine learning.

2 METHOD

First, the dataset used in this study was collected from an e-Book system during 90 min lectures of information science in Kyushu University. The target students of the lectures were beginners of information science. When an e-Book system is used, its timestamp, user id, material id, page number, and operation name are automatically recorded as an operation event. There are many types of operations; for example, OPEN indicates that a student has opened the e-Book file and NEXT indicates that the student has clicked the next button to move to the subsequent page. The number of lectures was 25, and two of five lecture materials were used in each lecture. A total of 110,894 e-Book operation logs were collected from 456 students. Then, we asked 15 students who got these lectures to evaluate the difficulty level of e-Books giving 5-level scores (1 for easy, 5 for difficult) for each page of lecture materials. we classified the pages into difficult or not as correct labels using the aggregated evaluation results.

Finally, we designed 6 features. we assumed that each page was explained by a teacher at the most browsed time. We defined that time as t (see Figure 1). Browsing before/after t shows that students were preparing/reviewing. Thus, we calculated following 3 features: the number of students browsing the page at t , the number of students browsing the page before/after t . In addition to these features, we calculated the number of page visits by the operation of NEXT/PREVIOUS, and the total browsing time. Therefore, each page has 6 dimensional feature vector and its corresponding label (difficult or not). We applied the Random Forest Classifier (Tin 1995) to acquire the classification model.

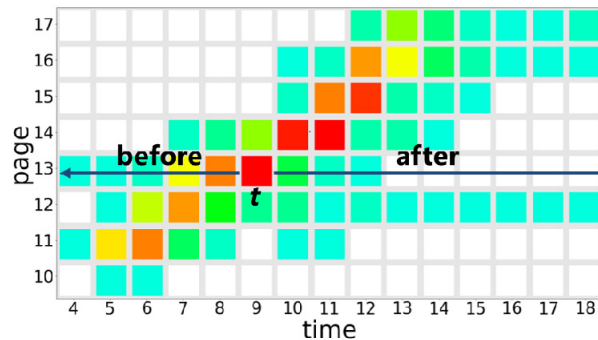


Figure 1: The number of page viewers per minute. The red color indicates that the large number of students are browsing the page.

3 RESULTS

We conducted the 10-fold cross validation using 921 samples including 185 samples in “difficult” label and 736 samples in “not difficult” label. The precision and recall ratio of “difficult” label were 0.430 and 0.454, respectively. Note that the chance rate to indicate the difficult pages is about 20% due to the imbalanced samples. Our method could provide better performance than the chance rate. Therefore, the features extracted from event stream logs contributes to the estimation. However, the result is not sufficient as an automatic difficulty estimation. Thus, we have to improve the feature extraction strategy and extract more effective features from event stream logs to make better models. Furthermore, since difficulty level is different for every person, it is not very easy which pages to regard as difficult as a correct label. We are going to design a strategy to effectively collect difficulty evaluation results from many students via the e-Book system.

REFERENCES

- Nakamura, K., Kakusho, K., Murakami, M., Minoh, M. (2008). Estimating Learners' Subjective Impressions of the Difficulty of Course Materials in e-Learning Environments. In: Distance Learning and the Internet Conference 2008, pp. 199–206.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., Yamada, M. (2015). E-Book-based learning analytics in university education. Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015) pp.401-406.
- Tin Kam Ho (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, 1995, pp. 278-282 vol.1.

Analytics of Time Management Strategies in a Flipped Classroom

Nora'ayu Ahmad Uzir

School of Informatics, University of
Edinburgh, United Kingdom
Faculty of Information
Management, Universiti Teknologi
MARARA, Malaysia
n.uzir@ed.ac.uk

Dragan Gašević

Faculty of Education, Monash
University, Australia
School of Informatics, University of
Edinburgh, United Kingdom
dragan.gasevic@ed.ac.uk

Wannisa Matcha

School of Informatics, University of
Edinburgh, United Kingdom
w.matcha@ed.ac.uk

Jelena Jovanovic

Faculty of Organizational Sciences,
University of Belgrade, Serbia
jeljov@fon.rs

Abelardo Pardo

Division of Information
Technology, Engineering and the
Environment, University of South
Australia, Australia
abelardo.pardo@unisa.edu.au

ABSTRACT: This study aims to explore time management strategies followed by students in a flipped classroom through the analysis of trace data. The study was conducted on the dataset collected in three consecutive offerings of an undergraduate computer engineering course (N=1,134). Trace data about activities were initially coded for the timeliness of activity completion. Such data were then analyzed using agglomerative hierarchical clustering based on the Ward's algorithm, first order Markov chains, and inferential statistics to detect time management tactics and strategies from students' learning activities. The results indicate that meaningful and theoretically relevant time management patterns can be detected from trace data as manifestations of students' tactics and strategies. In addition, this study also showed that time management tactics had significant associations with academic performance.

Keywords: Learning Analytics, Time Management, Flipped Learning, Self-Regulated Learning

1 BACKGROUND

Learning analytics allow for comprehensive data capture, however, connecting this data with higher level constructs such as learning strategies still remains a challenge. This study is an initiative to explore the capacity of data analytics methods to uncover patterns and trends in students' time management practices based on the trace data captured by a learning management system. It makes use of trace data to reveal individual differences in time management tactics and strategies and how these relate to the students' learning achievements (Broadbent & Poon, 2015), especially in flipped classroom setting. Time management was analyzed by looking at times when the students completed pre-class activities, as evidenced in the trace data and validated against the course schedule provided by the course instructor. Each week students were required to study one topic. Modes of study were assigned to each learning action based on its timing with respect to the week's topic: i) preparing - if students were on the topic that they were supposed to study for the given week, ii) ahead - if they were advance of the schedule, iii) revisiting - if students had visited the required activities for the behind-the-schedule topic at some earlier point in time, and iv) catching-up - if students had never before accessed activities related to the behind-the-schedule topic. By examining the students' modes of study, we expected to obtain insights that could inform the provision of feedback. In line with this objective, we defined our research questions as follows: i) Can we detect theoretically meaningful tactics and strategies of student time management from trace data about students' interactions with online preparatory learning activities in a flipped classroom? ii) What is the association between the students' time management strategies in the online component of a flipped course and their achievement? In particular, this study focuses on online learning activities that were designed to prepare students for face-to-face sessions. Trace data were collected

from three consecutive student cohorts enrolled, in years 2014, 2015, and 2016, in a computer engineering undergrad course ($N_{2014} = 290$, $N_{2015} = 368$, and, $N_{2016} = 476$) that followed a flipped classroom design. Meanwhile, the second data source was derived from midterm and final exam scores. These data were used to examine time management practices of high performing and low performing students, and if / how the two differed. In terms of analysis, first, each learning session was encoded as a sequence of learning modes indicative of student time management tactics. This was done using the TraMineR R library (Gabadinho, Ritschard, Mueller, & Studer, 2011). Second, agglomerative hierarchical clustering based on Ward's method was used to identify: i) time management tactics by grouping similar learning mode sequences and ii) time management strategies by grouping students with similar time management tactics. For both cluster analyses (i.e. tactics and strategies), the optimal number of clusters was determined by inspecting dendrograms. Finally, First Order Markov Model (FOMM) was generated for each time management tactic to further explain the tactics identified through clustering. FOMM allows for modeling the changing of states based on the probability theory and the assumption that the next state depends only on the current state. The pMineR R package was used to compute and visualize process models (B et al., 2017).

The clustering of the learning mode sequences produced four clusters that could be considered as manifestations of the time management tactics adopted by the students, namely: i) Tactic 1 (*Mixed and Short*) typically started their learning in the *preparing mode*; that is, by engaging with the activities required for the week's face-to-face session, or by revisiting the learning activities they have previously done as a part of preparation tasks, ii) Tactic 2 (*Revisiting*) shows high probability of revisiting events performed by the students for the entire course, iii) Tactic 3 (*Short Preparing*) is distinguished by high probability of preparing events throughout the course iv) Tactic 4 (*Long Preparing*) is strongly focused on preparation events throughout the duration of the course, but unlike Tactic 3, preparation events tended to form long learning sessions. Subsequently, three time management strategies were identified by grouping students with similar time management tactics, as follow: i) *Comprehensive and Active* strategy group mostly used the *Mixed and Short* tactic. They also demonstrated how to use effectively spaced practice (Tactic 2 – *Revisiting*) and combined that with tactics focused on preparation only (Tactics 3-4), ii) *Selective and Active* strategy group showed a low use of Tactic 1 (*Mixed and Short*) and almost no use of Tactic 4 (*Long Preparing*), iii) *Limited Activity* strategy group included students who focused mainly on Tactic 1 (*Mixed and Short*) for the entire duration of the course but not as intensively as in the previous two groups, while, all other tactics were very rarely used. Our analysis also suggests that students with higher academic performance were characterized by consistent efforts and diverse time management tactics throughout the entire course (*Comprehensive and Active*) compared to mid-performing (*Selective and Active*) and poorly performing students (*Limited Activity*).

In conclusion, the methodology proposed in this work allows for identifying patterns in students' time management behavior on the basis of study session data. Our findings indicated that time management patterns can be detected from student learning session as manifestation of students' time management tactics. Such observable patterns in learning behavior led to the detection of several strategy-based student groups. In addition, consistent with previous research, we found that more active and directive time management strategies promoted effective self-regulation and positive association with academic performance.

REFERENCES

- B, R. G., Lenkiewicz, J., Vallati, M., Rojas, E., Damiani, A., Sacchi, L., Valentini, V. (2017). Artificial Intelligence in Medicine, 10259, 351–355. <http://doi.org/10.1007/978-3-319-59758-4>
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education*, 27, 1–13. <http://doi.org/10.1016/j.iheduc.2015.04.007>
- Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.

Examining Science Learning by At-Risk Middle School Students in a Multimedia-Enriched Problem-Based Learning Environment

Sa Liu, Harrisburg University of Science and Technology, saliu@harrisburgu.edu

Min Liu, The University of Texas at Austin, MLiu@austin.utexas.edu

Zilong Pan, The University of Texas at Austin, panzl89@utexas.edu

Wenting Zou, The University of Texas at Austin, ellenzou@utexas.edu

Chenglu Li, The University of Texas at Austin, li.chenglu@utexas.edu

ABSTRACT: Little research on problem-based learning (PBL) exists for disadvantaged middle school students, especially students who are considered at risk of failing academically. To promote inclusion and success for all learners, this poster presentation will share our study on at-risk students using learning analytics. We examined science knowledge of a group of at-risk middle school students as they used a multimedia-enriched PBL environment. The results showed that these students significantly improved their science knowledge after they engaged in PBL learning. While there were no differences in the scores between the genders, the gain scores from pre- to post-tests in science knowledge for the girls were larger. Visualizations were used to present the findings from qualitative data. Such research should provide much needed insights on the effect of PBL for all students.

Keywords: Science, At-Risk Students, Problem-based learning, Visualization

1 INTRODUCTION

To understand and optimize at-risk middle school students' learning in multimedia-enriched environment, we investigated their science learning after they used a 3D immersive PBL environment—Alien Rescue. This environment is designed as a 15-hour curriculum unit in sixth-grade space science. Students assume the role of young scientists and participate in a rescue operation to find suitable relocation sites for six displaced alien species within our solar system. Our research questions were: Are there any differences in these at-risk students' science knowledge after they used a multimedia-enriched PBL science environment? Are there any gender differences?

2 METHOD

2.1 Participants

Participants were thirty-two middle school students (boys=17, girls=15) from three Title I schools with a high percentage of students on free/reduced lunch and minority populations in a US northeast state. These students were enrolled in a free STEM summer program in 2017 funded by a state grant that served at-risk youth and used the environment as summer curriculum for eight days.

2.2 Data Sources

Student science knowledge was assessed before and after they used this environment. A 20-item science knowledge test ($\alpha = .77$) was used to measure student understanding of scientific concepts introduced in the environment. Two open-ended questions were given as the post-questionnaire: 1) What have they learned from Alien Rescue? And 2) Compare their use of Alien Rescue to other science classes/activities and if they had learned science better? To supplement the quantitative data, interviews were conducted with a total of 25 students, randomly selected and interviewed by

the summer program staff. The responses to the open-ended questions and interviews were analyzed following the qualitative data analysis framework (Creswell, 2014). The qualitative data were coded by two researchers and checked by the entire research team to reach until 100% inter-rater reliability. Visualizations were also used to present the findings, specifically TermsBerry visualization to explore high frequency words, Mandala visualization to show the relationships between words and document(s), and StreamGraph visualization to depict the change of the frequency of words within a single document (Sinclair & Rockwell, 2016).

2.3 Results

ANOVA indicated that student science knowledge scores increased significantly from pretest to posttest: $F(1, 30) = 10.26, p < .01, ES = .26$ ($M_{\text{pretest}} = 45.78$; $M_{\text{posttest}} = 52.03$). Although, there were no differences in the scores between boys and girls, the gain scores from pre- to post-tests for girls were bigger ($\text{ScienceKnowledgeGainScore}_{\text{boys}} = 4.7$; $\text{ScienceKnowledgeGainScore}_{\text{girls}} = 8$). Visualizations were used to present the findings (See Figures 1, 2 and 3). As shown by the TermsBerry visualization in Figure 1.a, more students stated they learned science better after using this environment. Mandala visualization in Figure 1.b indicated student reasons for learning better: knowledge, fantasy, new, experience, computer, fun, and interesting—the shorter distance between the word and the document (word in the middle) represents a higher frequency of the word in it. For example, more students cited fun in providing their reasons—“[I prefer] this game. Because, I guess, science class is fun and everything, but this game gives me a new chance to have even more fun.” We further examined the positive words the students used to describe this environment and the frequency of these coded responses by genders. The coded responses by girls were presented first in Figure 2 and 3, followed by boys. That is, the left side of the X-axis in the visualization showed the codes for girls (X-axis from 0 to 3), while the right side indicated the codes for boys (X-axis from 4 to 9). More units on the X-axis represented a larger corpus by boys because more boys participated in the interviews. The Y-axis indicated the relative frequency each code appeared among all the codes. Both genders indicated they liked science better after using this environment (see Figure 2). Both the boys and girls listed these reasons: “fun,” “computer,” “new.” The girls also listed “knowledge” while the boys listed “experience,” as well as “fantasy” when they explained why they liked this environment (see Figure 3).

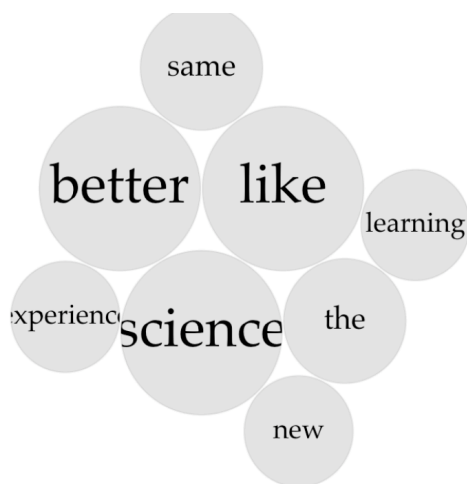


Figure 1.a TermsBerry Visualization

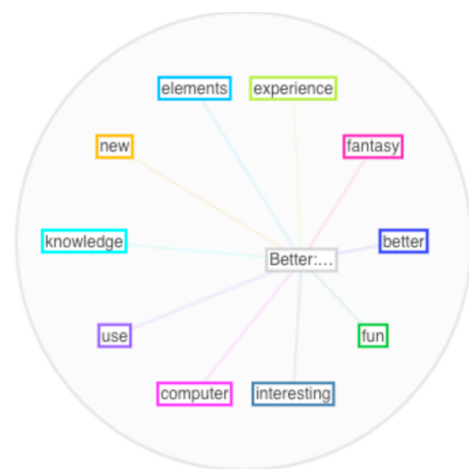


Figure 1.b Mandala Visualizations

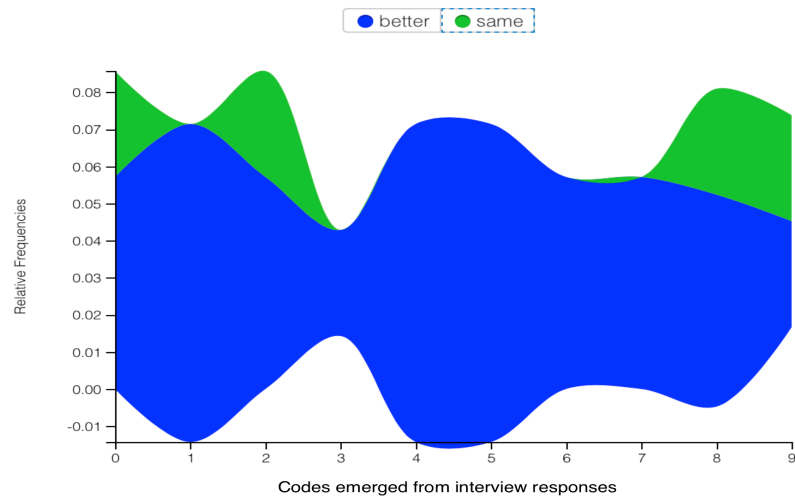


Figure 2. Streamgraph Visualizations on Students Interview Responses

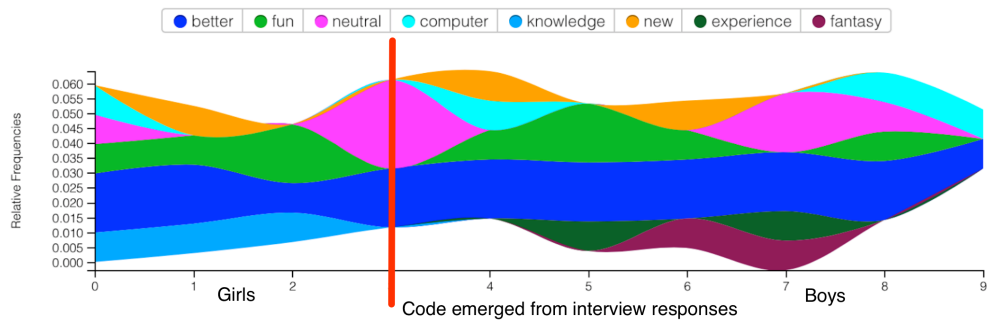


Figure 3. Streamgraph Visualizations on Students Interview Responses based on Category

REFERENCES

- Creswell, J. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches (4th ed.)*. Thousand Oaks, CA: Sage.
- Sinclair, S., & Rockwell, G. (2016). *Voyant Tools*. Retrieved from <http://voyant-tools.org/>

Examining gameplay of high score achieving students: comparison of replaying after a failed gameplay

Jihyun Rho¹, Gahgene Gweon²

Seoul National University, Seoul, Republic of Korea

¹whohehe@snu.ac.kr, ²ggweon@snu.ac.kr

ABSTRACT: In this paper, we used Kitkit School, a tablet-based educational game for learning basic concepts in mathematics, to investigate how children's achievement is related to children's replaying after a failed gameplay. For example, children's decision to replay after a failed gameplay might have an impact on children's achievement since children can practice their revised knowledge through retrials. To examine our research question on children's replay depending on children's achievement level, we assigned children into four types of high or low achievement groups based on the median score of pre-test and post-test. For types of replay, we classified the selection of next game to play after failing a game into four types: playing a new game, replaying a currently failed game, replaying a previously failed game, or replaying a previously passed game. Statistical analysis conducted on 82,385 instances of log-data from 91 children's plays showed that compared to learners with low achievement in the post-test, those with high achievement had a higher rate of selecting a currently failed game rather than a previously played failed/passed game.

Keywords: Educational game, achievement level, replaying after gameplay

1 INTRODUCTION AND RESEARCH QUESTIONS

Replay of an educational game can be explained by the concept of judgment-behavior-feedback cycle, a repeated loop in mastery learning (G. A. Gunter et al., 2008). In particular, a learner's autonomous choice to replay a game could advance their achievement, since a learner can apply feedback gained from the prior round of game play (Long, Y., et al., 2014). However, little is known about how children's achievement in test scores is related to the selection of a next game to replay after a failed gameplay. Therefore, in this paper we examine the relationship between the level of achievement (low and high pre/post test scores) and the four types of playing after a failed gameplay: playing a new game, replaying a currently failed game, and replaying a previously failed/passed game. We hypothesized when learners select the next game to play after a failed gameplay, high-achievement learners will select a currently failed game more often than low-achievement learners.

2 METHOD

The study data was collected from 91 students who used Kitkit School in rural primary schools in Tanzania for 30 minutes daily for three months, from September through December 2017. Kitkit School is a tablet-based educational game that helps children practice basic math in K-2 Knowledge. A total of 82,385 instances of gameplay data was collected. The average age of the children was 9.09 years, and 51.64% were female. We also conducted pre-test and post-test evaluations at the beginning and end of data collection period. For each test, students could score up to a total of 72 points. The contents of all tests were based on the Early Grade Mathematics Assessment (EGMA) of

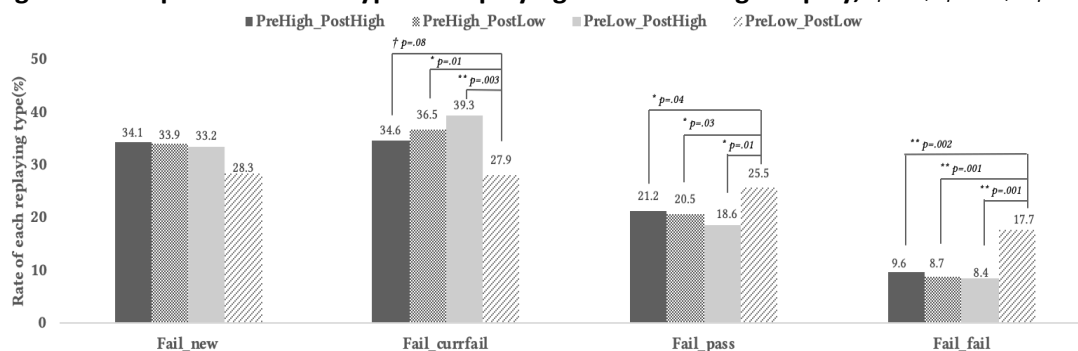
the 2014 National Baseline Assessment developed by the United States Agency for International Development (USAID).

We examined children's achievement groups and types of replaying after a failed gameplay. The children were assigned to four types of achievement groups; preHigh_postLow (n=14), preHigh_postHigh (n=33), preLow_postHigh (n=13), and preLow_postLow (n=31). Students were assigned to each group based on the median score of their pre-test (m=29.0) and post-test (m=44.0) scores. For instance, if a learner scored 25 on pre-test and 58 on post-test, she was assigned to the preLow_postHigh group. We also identified four types of playing after a failed gameplay: playing a new game (Fail_new), replaying a currently failed game (Fail_currfail), replaying previously passed games (Fail_pass), and replaying previously failed games (Fail_fail). Note that in Fail_currfail, a learner immediately replays the same failed game right after a failed gameplay. In comparison, in Fail_fail, a child replays a failed game which is not the game that was played in the immediate prior round, but is a game that was played in previous rounds. For Fail_pass, a child replays a passed game that was played in previous rounds.

3 RESULTS AND CONCLUSION

In this study, we investigated learners' replaying according to their achievement in the post-test. As shown in Figure 1., compared to the preLow_postLow group, children with high achievement, regardless of their pre-test score, had a significantly higher rate of replaying a currently failed game (Fail_currfail). On the contrary, compared to the preLow_postLow group, children with high-achievement showed a significantly lower rate of replaying previously failed and passed games (Fail_pass/ Fail_fail). Our results suggest that in educational games, learning support that encourages learners to replay the currently failed game after a failed gameplay may help them to practice, and thereby achieve better learning outcomes through educational games. However, considering PreHigh_PostLow who adopt similar strategies but ended up with low achievement, further studies examining the detailed impact of learning support that encourages immediate replay after a failed gameplay are needed in order to test the generalizability of our results.

Figure 1. Comparison of four types of replaying after a failed gameplay, [†]p<0.1, *p<0.05, **p<0.01



REFERENCES

- G. A. Gunter., et al., (2008). Taking educational games seriously. *Educational Technology Research and Development*, 56(5-6), pp.511–537.
- Long, Y., et al., (2014). Gamification of joint student/system control over problem selection in a linear equation tutor. In *International Conference on Intelligent Tutoring Systems*, pp. 378-387.

Predicting Graduation at a Public R1 University

Henry Anderson

University of Texas at Arlington
henry.anderson@uta.edu

Afshan Boodhwani

University of Texas at Arlington
afshan.boodhwani@uta.edu

Ryan Baker

University of Pennsylvania
rybaker@upenn.edu

ABSTRACT: In this **poster**, we build a set of high-performance machine learning models to predict 6-year graduation for university undergraduate students, a critical metric for state and federal reporting and university evaluation, using Linear Support Vector Machines, Decision Trees, Logistic Regression, and Stochastic Gradient Descent binary classifiers. We use a data set of over 14,000 students from six Fall cohorts, containing 104 features, drawn from pre-existing university data. This minimizes sparsity and data collection time, while improving coverage of the student body and student activities. Our models achieve high performance, and identify GPA and completed credit hours as the most important predictors.

Keywords: graduation, predictive modelling, first time in college, machine learning

1 INTRODUCTION

For many universities, graduation is an important measure of institutional effectiveness, particularly in an era when some institutions are criticized for very low graduation rates. Researchers have thus sought to understand and predict students' graduation, frequently using machine learning and data mining techniques (Raju and Schumacker, 2015; Kuh et al., 2008; Karamouzis and Vrettos, 2008), and achieving high predictive accuracies. This poster reports on early, but promising, results of our own such efforts to predict 6-year graduation for first time in college (FTIC) undergraduate students in a public, four-year university.

2 DATA

We used a data set taken from a publicly-funded, four-year state research university in the southern United States, which serves a diverse population, and is a federally designated Hispanic-Serving Institution. The data set includes 14,706 FTIC students admitted in the Fall terms of 2006-2012 (inclusive). Only data from a student's first academic year were included, since prior research has shown that this early period of a student's college career is the most critical for retention and graduation outcomes (Tinto, 2006; Arnold and Pistilli, 2012).

We only used data that the university collects as part of its routine reporting efforts, which allows us to make use of a large number of features for each student, while minimizing the data’s overall sparsity. Compared to features only available for a smaller number of students—e.g. surveys, interviews—this makes our resulting predictions more reliable given our choice of modeling algorithms.

3 METHODOLOGY

We extracted 104 features related to students’ first academic year, in order to capture a broad view of students’ experiences and activities. Through prior reporting work (e.g., to state and federal agencies), a number of variables had been identified by the university that provided our data as likely predictors of student graduation. We use these variables, along with several closely related measures, as features in our models. These features fall into four major categories: *academic performance* (e.g. GPA, credit hours completed), *financial information* (e.g. scholarships, unmet need), *pre-admission information* (e.g. SAT/ACT scores, high school rank), and *extra-curricular activities* (e.g. involvement in Greek Life or Athletics).

The target for classification was defined as the binary variable: *did the student graduate from this university within 6 years of first enrolling?* Using this definition, 46% (6,787) of the students in the data set were assigned a label of “true” (graduated). This does not distinguish different types of failure to graduate—students who drop out, transfer to another institution, or graduate in more than six years are all assigned a “false” value for classification. While these do represent very different student outcomes, each still represents a student who is not being fully served by their university, and whom we wish to identify early in their academic career.

We trained a set of four binary classifiers on the data set to predict FTIC students’ 6-year graduation, using the scikit-learn 0.20.0 (Pedregosa et al., 2011) implementations: Linear-kernel Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Stochastic Gradient Descent classifier (SGD). All predictor variables were scaled to zero mean and unit variance when training and evaluating the SVM, Logistic Regression, and SGD models. We held out 20% of the data for testing, and performed 5-fold cross-validation on the remaining 80% to tune model parameters. Models with the highest AUC-ROC score during cross-validation were evaluated on the held-out data.

4 RESULTS AND DISCUSSION

Table 1: AUC-ROC and F1 scores for each model, evaluated on the held-out testing set.

Model	AUC	F1
Decision Tree	0.786	0.785
Linear SVM	0.801	0.795
Logistic Regression	0.814	0.810
SGD Classifier	0.824	0.822

The scores on the testing set are reported in Table 1. As that table shows, each of the four classifiers performed approximately equally well on the held-out data.

The models' feature weights are not directly comparable, making it difficult to identify the most important predictors overall. To account for this, we compute an approximate "overall importance" metric. For each model, we sort features by the absolute value of their assigned weight, then calculate each feature's average rank across the four models. Total credit hours completed, cumulative GPA at the end of the first academic year, out-of-major GPA, and the percent completed credit hours (the student's completed credit hours as a percent of the credit hours they enrolled for) were consistently the highest-ranked features (both in the overall ranking and within each model), which is in keeping with much of the prior work on graduation prediction that finds GPA and credit hours to be the most important predictors.

5 FUTURE WORK

Our current results are encouraging, though they only represent an early analysis of the data. The predictions and feature rankings need to be tested experimentally, to investigate whether they are useful for guiding student interventions and changes in university policy. The models may also be suppressing the effects of lower-ranked features, which may be more directly useful for informing interventions or instructional practices; this merits further investigation, e.g. by re-fitting the models using only a subset of the available features. Given the possible applications in interventions and policy decisions, these models should also be thoroughly tested for algorithmic bias, e.g. lower performance for specific student races or ethnicities. We see this as the most pressing, avenue of future work.

REFERENCES

- Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM. <https://doi.org/10.1145/2330601.2330666>.
- Karamouzis, S. T., & Vrettos, A. (2008). An artificial neural network for predicting student graduation outcomes. In *Proceedings of the World Congress on Engineering and Computer Science* (pp. 991-994).
- Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The journal of higher education*, 79(5), 540-563.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), 2825-2830.
- Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, 16(4), 563-591.
- Tinto, V. (2006). Research and practice of student retention: What next?. *Journal of College Student Retention: Research, Theory & Practice*, 8(1), 1-19.

Person-Oriented Approach to Profiling self-regulation in STEM learning

Juan Zheng^{1,5}; Wanli Xing^{2,5}; Gaoxia Zhu^{3,5}; Guanhua Chen^{4,5}; Henglv Zhao^{2,5}; Xudong Huang^{4,5}
McGill University¹; Texas Tech University²; University of Toronto³; Concord Consortium⁴;
Learning Genome Collaborative⁵

juan.zheng@mail.mcgill.ca; wanli.xing@ttu.edu; gaoxia.zhu@mail.utoronto.ca;
gchen@concord.org; Henglv.zhao@ttu.edu; xhuang@concord.org

ABSTRACT: Students can display different types of self-regulation: they can be “cognitively oriented,” “behaviorally oriented,” or “minimally self-regulated.” Instead of evaluating the self-regulation profiles of individuals, previous studies have generally used variable-oriented approaches. This study used principal component analysis and cluster analysis to classify learners’ self-regulation profiles and to determine the relationship between self-regulation and student performance. The results revealed that the behaviorally oriented learners performed better than did the minimally self-regulated learners, though the cognitively oriented learners performed the best. The results also offer new insights into SRL with emerging learning analytics. Learning analytics used in person-oriented approach have the potential to enable data-driven assessments that could be used to provide adaptive feedback to learners.

Keywords: self-regulation, STEM learning, person-oriented profiling, Energy3D

1 INTRODUCTION

“Self-regulation” refers to one’s ability to actively monitor and control one’s learning using a variety of cognitive and behavioral strategies (Zimmerman, 2000). Although a wealth of research has demonstrated the impact of self-regulation on learning performance (Sitzmann & Ely, 2011), these studies have generally employed variable-oriented statistical approaches. Though Ning and Downing (2015) used a person-oriented approach to identify four self-regulation profiles—“competent,” “cognitively oriented,” “behaviorally oriented,” and “minimally self-regulated”—they relied exclusively on self-reports. In addition, few studies have used trace data to determine students’ self-regulation profiles. This is especially true for STEM learning, in which computer-based learning environments are used to support SRL. To fill this gap in the research, this study sought to answer the following research questions: (1) Do students display different self-regulation profiles when engaging in computer-based STEM learning? (2) How the various profiles of self-regulated learners differ in terms of performance?

2 METHOD

The participants were 108 9th-grade students from a suburban high school in the northeastern United States. The participants spent 50–80 minutes each day during a science course that lasted nine consecutive days performing three design tasks on Energy3D, a simulated environment in which students can complete home-design projects that produce renewable energy (Xie, Schimpf, Chao, Nourian, & Massicotte, 2018). Energy3D provides students not only 3D modeling tools that allow them to design realistic buildings, but also plenty of tools for quantitative analysis that they can use to evaluate their buildings’ energy performance. Energy3D also captures students’ actions on a timeline,

and these can be used to determine their self-regulation profiles. In this study, each student experienced three self-regulation processes: “orientation” (i.e. adding walls, windows, and solar panels to increase their awareness of the learning environment), “monitoring” (i.e. changing the energy setup to achieve better learning outcomes), and “self-reflection” (i.e. taking notes to elaborate and evaluate their progress). Students’ performance was measured according to the net energy required by the house they built—the lower the better. In determining the students’ self-regulation profiles, we first used a principle component analysis to reduce the high dimensionality of 93 types of actions. A k-means cluster analysis was then conducted to determine the students’ self-regulation profiles. Finally, an ANOVA was performed to examine the performance difference among different profiles of students.

3 RESULTS AND CONCLUSION

In response to research question 1, Table 1 displays the results for the three SRL profiles. Only the events that occurred most frequently and were most relevant to self-regulation are presented. The students in Cluster 1 frequently engaged in monitoring and self-reflection, so they were described as “cognitively oriented” (Ning & Downing, 2015). The students in Cluster 2 engaged in orientation most frequently, so they were described as “behaviorally oriented.” The students in Cluster 3 engaged the least of all of the clusters in all three self-regulation processes, so they were described as “minimally self-regulated.” In response to research question 2, the ANOVA analysis revealed that the cognitively oriented learners performed the best ($M = -1077.5$, $SD = 1836.0$), the behaviorally oriented students performed intermediately ($M = -999.27$, $SD = 2616.93$), and the minimally self-regulated learners performed the worst ($M = 2985.56$, $SD = 9928.35$). These findings lend empirical support to the self-regulation framework, and they reveal the importance of self-regulation to student performance in STEM learning.

Table 1: Cluster Result.

Self-regulation processes	Sample of SRL events	Cluster 1 (N=31)	Cluster 2 (N=26)	Cluster 3 (N=51)
Orientation	add solar panel, add wall, add window	91.81	111.19	55.59
Monitoring	change inside temperature, change color of solar heat map, edit solar panel, efficiency change for selection	80.45	69.12	26.18
Self-reflection	note-taking	494.55	198.35	83.59

4 REFERENCES

- Ning, H. K., & Downing, K. (2015). A latent profile analysis of university students' self-regulated learning strategies. *Studies in Higher Education*, 40(7), 1328-1346.
- Sitzmann, T., & Ely, K. (2011). A Meta-Analysis of Self-Regulated Learning in Work-Related Training and Educational Attainment: What We Know and Where We Need to Go. *Psychological Bulletin*, 137(3), 421-442. doi:10.1037/a0022777
- Xie, C., Schimpf, C., Chao, J., Nourian, S., & Massicotte, J. (2018). Learning and teaching engineering design through modeling and simulation on a CAD platform. *Computer Applications in Engineering Education*, 26(4), 824-840.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. *Handbook of Self-Regulation*, 13-39. doi:10.1016/B978-012109890-2/50031-7

Understanding the factors contributing to persistence among undergraduate engineering students in online courses

Samantha R. Brunhaver, Jennifer M. Bekki, Eunsil Lee, & Javeed Kittur

Arizona State University

Samantha.Brunhaver@asu.edu

ABSTRACT: This poster will report on the research design and methodology planned for a recently funded National Science Foundation-sponsored project focused on advancing knowledge about the factors that influence the decisions of undergraduate engineering student to complete (rather than drop out of) online courses. Through the application of both social science and learner analytics-based research methods, the research will explore how students' perceptions about the characteristics of their online undergraduate engineering courses and engagement with their course learning management system (LMS) influence their persistence. To support these studies, we draw on the undergraduate engineering student population at a large, public university in the southwestern United States that has been an early adopter of comprehensive online undergraduate engineering education. The findings from this work will be both important and timely, as the field of engineering education shows signs of embracing the online presence critical to increasing access and participation in engineering.

Keywords: Persistence, Online learning, Learner analytics, Structural equation modeling

1 PROJECT OVERVIEW

Ensuring widespread access to education is both a national imperative and a call to action for the engineering education community (National Research Council, 2007). Online education is simultaneously disrupting and transforming the educational landscape, demonstrating potential to address the issues of access (Allen et al., 2016). In contrast to many other fields, until quite recently, engineering education has been slow to adopt or research the online pathway. However, there are now some indications the field is in transition to a greater online presence, with ABET now accrediting several online undergraduate degree programs (ABET, Inc., 2018) and an increasing number of other undergraduate engineering programs offering online courses as well. The work proposed here will take advantage of the early adoption of online engineering education at a large, public university in the southwestern United States to study and report critical information to the online and engineering education communities on factors that influence its efficacy.

2 RESEARCH DESIGN AND METHODOLOGY

This project has the overarching goal of advancing understanding of the factors influencing course-level persistence among the population of online undergraduate engineering students. The choice of course-level persistence as a measure of educational efficacy is in line with much of the literature related to online learning (e.g., Xu & Jaggars, 2013) and has clear links to more traditional persistence-related measures such as degree completion. A Model of Online Course-level Persistence in Engineering (MOCPE), which is grounded in online and undergraduate engineering

student persistence and which combines findings and ideas from theories of student motivation (Keller, 1987; Wigfield & Eccles, 2000), will be developed and empirically evaluated.

The project will be comprised of three studies. The Diary Study will use a within-person diary method (Bolger & Laurenceau, 2013) to investigate how online undergraduate engineering students' perceptions of their course affect: (i) their beliefs about their chances of success in the course, (ii) their perceptions of the value of the course, (iii) their level of engagement with the online course learning management system (LMS), and (iv) their decision to complete the course. Students will be recruited from 7.5 week-long online engineering courses and surveyed bi-weekly until they complete or drop out of the course. The LMS Interaction Study will then apply the learner analytics-based technique of associative classification (Sun et al., 2006) to historical data from online undergraduate engineering courses in order to generate "rules of engagement" that describe LMS-interaction behaviors associated with course-level persistence. These rules will be combined with measures of students' perceptions and beliefs to develop a complete model of course-level persistence in the Persistence Modeling Study. This model will be tested using longitudinal structural equation modeling with data from a sample of current online engineering students to determine whether the complete model predicts student persistence better than LMS data or student attribute data alone.

3 IMPLICATIONS FOR ONLINE AND ENGINEERING EDUCATION

While this study is not without limitations (such as the possibility to influence course completion and drop-out rates), knowledge will be generated about whether and how online course characteristics related to the LMS, instructor practices, and peer support influence students' persistence decisions. Additionally, the development and evaluation of the MOCPE will yield evidence to support a proposed theoretical framework upon which future research and educational practice can build.

REFERENCES

- ABET, Inc. (2018). *Online programs accredited by ABET*. Retrieved from <https://www.abet.org/accreditation/find-programs/>.
- Allen, E.I., Seaman, J., Poulin, R., & Strout, T.T. (2016). *Online Report Card: Tracking Online Education in the United States*. Babson Survey Research Group & Quahog Research Group, LLC. Retrieved from <https://www.onlinelearningsurvey.com/reports/online-report-card.pdf>
- Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68-81.
- Keller, J.M. (1987). Development and use of the ARCS model of motivational design. *Journal of Instructional Development*, 10(3), 2-10.
- National Research Council. (2007). *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. Washington, DC: The National Academies Press.
- Sun, Y., Wong, A., Wang, A. (2006). An overview of associative classifiers. In Proceedings of the 2006 International Conference on Data Mining, 138-143.
- Xu, D. & Jaggars, S.S. (2013). The impact of online learning on students' course outcomes: Evidence from a large community and technical college system. *Economics of Education Review*, 37, 46-57.

CanoPy: Using Python Scripts to Promote Teacher-Driven Learning Analytics

Charles Lang

Teachers College Columbia University
charles.lang@tc.columbia.edu

Detra Price-Dennis

Teachers College Columbia University
dmp2192@tc.columbia.edu

ABSTRACT: Poster. CanoPy is an open source Python module that aims to make learning analytics tools accessible to classroom teachers in a user-friendly but highly flexible way. The project consists of converting teacher-formulated problems of practice into Python scripts, using a vocabulary that is generated by teachers and therefore more intuitive to teachers. The hope is that this process generates both relevant and requested tools and a gateway for teachers to learn and access those tools.

Keywords: Teacher professional development, data literacy, coding, accessible analytics

1 INTRODUCTION

Although the advent of online and mobile computing in the classroom has meant a tremendous growth in the amount and variety of data collected about students (Merceron, Blikstein, & Siemens, 2015), we are at the very early stages of utilizing this data for educational purposes in K-12 classrooms (Rodríguez-Triana, Martínez-Monés, & Villagrà-Sobrino, 2016). Integrating data into teaching-practice for school-age students is a complex task, requiring thoughtful consideration of context-specific pedagogical goals, technical and practical limitations, stakeholder concerns and ethics. These needs require learning analytics tools to be highly adaptable and flexible, but at the same time they must also be highly usable to avoid over-burdening teachers with professional development. This usability-flexibility trade-off (Rocha, Correia, Adeli, Reis, & Costanzo, 2017) has led us to the development of CanoPy, a scripting language in Python that aims to be flexible enough to adapt to complex, heterogenous contextual analytic needs of individual classrooms but intuitive enough to allow a low barrier to utilization.

1.1 Analytic Strategy

The strategy we are investigating involves developing a Python module of scripts from teacher-generated pseudocode. Teachers attend a short learning analytics workshop in which they define a problem of practice, decompose that problem into parts and then devise pseudocode to generate analytics that could help them solve the problem. This pseudocode is then converted into working Python code by our development team and provided back to the teachers to determine if it makes sense to them and to see if they can use it in new ways. We hope this strategy can a.) ensure that the scripts reflect the heterogeneity of authentic problems of practice from the teachers perspective

and b.) create an intuitive coding vocabulary that allows teachers to creatively utilize analytics in their work in an extensible way.

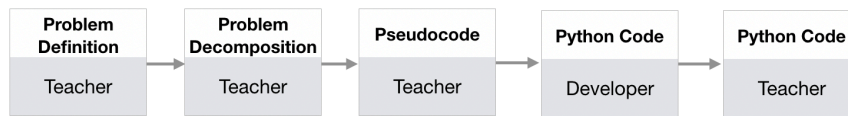


Figure 1: Process for developing teacher-relevant Python scripts

1.2 Results

We have currently worked with 30, K-12 STEM teachers from across a large metropolitan city in the Northeast of the USA. So far, teachers have been largely interested in analytics that concern process (attendance records, homework completion, disruptive behavior) and less that directly looks at measuring learning. We hope that these process questions can be leveraged into questions that more directly impact learning in the future. Below is an example of the type of problem that was posed by a teacher, how they decomposed the problem, their pseudocode and the definition of the Python command that was coded:

Table 1: Example of teacher generated problem and script.

Problem	Decomposition	Pseudocode	Python Script
Student disruptive behavior changes with season, want to predict changes	Count disruptive behavior, count season, predict future change	Predict(sum(DB) by month)	Season = plot(DB, month) predict(Season, student)

1.3 Conclusion

Through the process of code development we hope that teachers will be able to gain the skills necessary to implement learning analytics solutions within their classrooms but also these solutions will be able to be shared through a common intuitive vocabulary stored as a Python module. This would seem to negotiate between usability and flexibility by allowing a highly extensible format that could take into account the heterogeneity of classrooms but one that can be altered with less effort than a GUI or dashboard. It also provides a natural tool to introduce teachers to learning analytics concepts and a step towards “programming classes” rather than “planning classes”.

1.4 References

- Drachsler, H., & Greller, W. (2016). Privacy and analytics: It’s a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 89–98). New York, NY, USA: ACM. <https://doi.org/10.1145/2883851.2883893>
- Merceron, A., Blikstein, P., & Siemens, G. (2015). Learning Analytics: From Big Data to Meaningful Data. *Journal of Learning Analytics*, 2(3), 4–8. <https://doi.org/10.18608/jla.2015.23.2>
- Rocha, Á., Correia, A. M., Adeli, H., Reis, L. P., & Costanzo, S. (2017). *Recent Advances in Information Systems and Technologies*. Springer.
- Rodríguez-Triana, M. J., Martínez-Monés, A., & Villagrà-Sobrino, S. (2016). Learning Analytics in Small-Scale Teacher-Led Innovations: Ethical and Data Privacy Issues. *Journal of Learning Analytics*, 3(1), 43–65.

Advances in Writing Analytics: Mapping the state of the field

Antonette Shibani

Connected Intelligence Centre, University of Technology Sydney, Australia
antonette.shibani@uts.edu.au

Ming Liu

Connected Intelligence Centre, University of Technology Sydney, Australia
School of Computer and Information Science, Southwest University, China
ming.liu@uts.edu.au

Christian Rapp

ZHAW, School of Management and Law, Center for Innovative Teaching and Learning,
Winterthur, Switzerland
rapp@zhaw.ch

Simon Knight

Faculty of Transdisciplinary Innovation, University of Technology Sydney, Australia
Simon.Knight@uts.edu.au

ABSTRACT: Writing analytics as a field is growing in terms of the tools and technologies developed to support student writing, methods to collect and analyze writing data, and the embedding of tools in pedagogical contexts to make them relevant for learning. This workshop will facilitate discussion on recent writing analytics research by researchers, writing tool developers, theorists and practitioners to map the current state of the field, identify issues and develop future directions for advances in writing analytics.

Keywords: writing analytics, learning analytics, collaborative writing, writing theories, writing analytics advances

1 BACKGROUND

As technological capabilities progress in the field of understanding natural language, there is increasing interest in their application to study and improve writing. *Writing analytics* has emerged as a sub-domain of learning analytics to support the analysis of written products and processes in educational contexts (Buckingham Shum et al., 2016). The time-consuming and labor-intensive process of assessing writing makes it hard for educators to provide formative feedback on students' writing, which could be supported by writing analytics. An application of writing analytics that has gained traction is the use of tools that provide automated feedback and writing instruction to improve students' writing skills (Allen, Jacovina, & McNamara, 2015; Liu, Li, Xu, & Liu, 2017; Woods, Adamson, Miel, & Mayfield, 2017). Such tools developed across different educational levels engage students directly to aid in the improvement of their writing skills. Another objective of writing analytics tools

and techniques is to understand the writing *products* and *processes* deeper to contribute to the theory and research on writing, which can then lead to its application in writing contexts (McNamara, Graesser, McCarthy, & Cai, 2014). In addition to studying user behavior and interaction through log data, this can inform design choices in writing tool development. These applications build on the main notion of developing a synergy between writing analytics technology and pedagogical practice, so that the educational context is meaningfully embedded in the use of these technologies. Three previous workshops run on this topic have focused on critical perspectives and community building around writing analytics in LAK (Buckingham Shum et al., 2016), developing a writing analytics literacy and practitioner capacity (Knight, Allen, Gibson, McNamara, & Buckingham Shum, 2017) and a hands-on-training for developing this literacy by understanding technical affordances and aligning them to pedagogical feedback (Shibani, Abel, Gibson, & Knight, 2018).

2 WORKSHOP FOCUS

The proposed fourth workshop in the series will build on the previous writing analytics workshops to develop writing analytics literacy and map the field for the future. The focus will be on critically assessing the current state of work being done in the field, and how it could be directed towards the future by considering key issues. The key thread of integrating writing analytics with pedagogy will be emphasized, by connecting theory, pedagogy and assessment to close the feedback loop (Knight, Shum, & Littleton, 2014; Shibani, Knight, Buckingham Shum, & Ryan, 2017). The pedagogic relevance and the question of why writing analytics is employed and what it can add to the existing system will be brought into discussion by practitioners. In this way, we maintain a productive dialogue among different stakeholders like educators, researchers and developers for effective implementation of learning analytics in the classroom (Thompson et al., 2018; Shibani, Knight, & Buckingham Shum, 2019).

The landscape of tools that offer support for writing is constantly changing with new tools getting introduced and the existing ones getting updated, to incorporate the technical advances and the data made available over time (Liu, Calvo, Pardo, & Martin, 2015; McDonald, Moskal, Gunn, & Donald, 2018; Rapp & Ott, 2017; Woods et al., 2017). The ways in which we study writing, and respective systems that support its instruction and practice, have also considerably changed with technological affordances like keystroke-level analysis which allow for a more fine-grained level of analysis, and multiple sources of data which allow for triangulation and validation while studying writing processes. It is important to share knowledge from related work on writing, for instance process-mining and temporal analysis, that can contribute to writing analytics research. This will expand the knowledge base of the community and find relevant opportunities to meaningfully collect, analyze, visualize and use data to derive insights that are relevant for the learning contexts. Hence, the workshop will encourage presentations on various tools and techniques to understand and improve writing.

With growth in the field of Writing Analytics, the multidisciplinary of the field, and the different ways in which researchers engage with its development, it is important to align the goals of the field within the community. Community building generates a shared understanding and common goals to work towards the future of the field. While considering the potential pathways for the field to progress, we

will also include discussions on the pushbacks and critical perspectives that can affect how the field moves forward. This includes legal and ethical considerations on the use of students' data, development of learning theories to support writing analytics technology, and evaluation methods to assess these advances for their real impact to meaningfully contribute to writing.

Thus, the fourth workshop is intended to:

1. Build on the existing dialogue around developing writing analytics literacy and pedagogic integration by connecting different stakeholders like practitioners and researchers.
2. Expand the knowledge of the field by discussing about novel approaches and tools being developed by different researchers that contribute to writing analytics research.
3. Move the field forward by building a community for writing analytics research and thinking about pushbacks and potential future steps.

3 SUBMISSIONS AND WORKSHOP FORMAT

Workshop activities and schedule

The full-day workshop will include a number of presentations and demonstrations from researchers to share their work within the writing analytics community (depending on the interest generated). It will include round-table and open discussions throughout the day to steer the direction of writing analytics work and possible pathways for future advances in the field. The provisional program is given below:

Introductions (30 minutes): Introductions of workshop organizers and participants, and a quick background to the field of writing analytics.

Presentations (10-15 minutes each): Presentations and demonstrations from accepted papers and invited researchers on their writing analytics tool or technology, the data collected by the tool, analysis of writing data and how it contributes to writing theory, and the direction of future work.

Discussion Blocks (5-10 minutes each): Discussion blocks will follow each presentation to ask critical questions on what can be done and analyzed from the tool/data, how and why.

Round-table discussion (1 hour): Key topics for discussion from the presentations will be selected for round-table discussion. Participants can move around tables to discuss more in detail on the topic they are interested in. Potential topics include collaborative writing analytics, analytical and reflective writing analytics, writing feedback visualization and writing theories.

Open discussion (30 minutes): Open discussion facilitated among all participants on the advances in writing analytics and its potential future, co-creation of shared notes and resources.

Writing analytics community engagement (30 minutes): Building the community of writing analytics researchers by connecting existing and new researchers in the field. Formation of a formal writing analytics committee if participants are interested.

Concluding remarks and future directions (15 minutes): Brief summary and closing remarks on the workshop with future steps.

Program Committee

Co-chairs of the workshop will invite researchers and companies active in the field of writing analytics to present their work in the form of tool demonstrations or presentations. They will also review submissions for presentations by extending an open call for participation.

Participation, Required Equipment and Dissemination

Participation will be 'mixed' – in addition to participants who are invited to present their work, any interested delegate may register to attend. An invitation will be extended to participants of previous workshops, writing researchers who are not (yet) involved with the technology side, and international researchers active in the field to share their work and different perspectives on Writing Analytics. An open call for participation will be put out to encourage others to present their research and become more actively involved in the LAK writing analytics community. A website setup for the workshop will archive the event and disseminate the notes to participants. Papers accepted for presentation will be published in the companion proceedings and linked to the website.

The workshop will be of interest to a wide range of LAK delegates including: students and researchers engaged in writing research and the use of writing tools; educators in schools, universities and businesses; data analysts; and companies active or potentially active in the field. The workshop does not require any special equipment (WiFi, data projector and power strips aside). Flexible seating is preferred for breakout discussion groups. Participants will be encouraged to bring their own devices to contribute to shared notes. Workshop organizers will make use of listservs (SoLAR, Learning Analytics Google group, EDM-announce, ISLS, SIG-LS, ICCE) and their own personal networks to advertise the workshop.

REFERENCES

- Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Computer-based writing instruction. *Handbook of writing research*, 316-329.
- Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). *Critical perspectives on writing analytics*. Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.
- Knight, S., Allen, L., Gibson, A., McNamara, D., & Buckingham Shum, S. (2017). *Writing analytics literacy: bridging from research to practice*. Paper presented at the Proceedings of the Seventh International Learning Analytics & Knowledge Conference.
- Knight, S., Shum, S. B., & Littleton, K. (2014). Epistemology, assessment, pedagogy: where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2), 23-47.
- Liu, M., Calvo, R. A., Pardo, A., & Martin, A. (2015). Measuring and visualizing students' behavioral engagement in writing activities. *IEEE Transactions on Learning Technologies*, 8(2), 215-224.
- Liu, M., Li, Y., Xu, W., & Liu, L. (2017). Automated Essay Feedback Generation and Its Impact on Revision. *IEEE Transactions on Learning Technologies*, 10(4), 502-513.
- McDonald, J., Moskal, A. C. M., Gunn, C., & Donald, C. (2018). Text analytic tools to illuminate student learning. *Learning Analytics in the Classroom: Translating Learning Analytics for Teachers*.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*: Cambridge University Press.

- Rapp, C., & Ott, J. (2017). Learning Analytics in Academic Writing Instruction—Opportunities Provided by Thesis Writer (TW). *Bildungsräume* 2017.
- Shibani, A., Abel, S., Gibson, A., & Knight, S. (2018). *Turning the TAP on Writing Analytics*. Paper presented at the Learning Analytics and Knowledge, Sydney.
- Shibani, A., Knight, S., Buckingham Shum, S., & Ryan, P. (2017). *Design and Implementation of a Pedagogic Intervention Using Writing Analytics*. Paper presented at the 25th International Conference on Computers in Education, New Zealand.
- Shibani, A., Knight, S., Buckingham Shum, S. (2019). *Contextualizable Learning Analytics Design: A Generic Model, and Writing Analytics Evaluations*. In Proceedings of the International Conference on Learning Analytics and Knowledge (LAK'19). ACM, New York, NY, USA. <https://doi.org/10.1145/1234567890>.
- Thompson, K., Alhadad, S. S., Buckingham Shum, S., Howard, S., Knight, S., Martinez-Maldonado, R., & Pardo, A. (2018). Connecting expert knowledge in the design of classroom learning experiences *From data and analytics to the classroom: Translating learning analytics for teachers*.
- Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). *Formative Essay Feedback Using Predictive Scoring Models*. Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

WAT: Writing Assessment Tool

Danielle S. McNamara¹, Laura K. Allen², & Scott Crossley³

¹Arizona State University, ²Mississippi State University, ³Georgia State University
¹dsmcnamara1@gmail.com, ²laura.allen22@gmail.com, ³sacrossley@gmail.com

This presentation describes our progress on a project to develop the Writing Assessment Tool (WAT): an on-line platform to provide students, teachers, and researchers access to automated writing analytics. WAT will comprise three access points, each tailored to the needs of these three types of end-users. From a single-entry point: *Students* will receive summative and formative feedback via automated writing evaluation (AWE) on three types of essays: persuasive (independent) essays, summaries, and source-based (integrative) essays. *Teachers* will have access to a teacher interface allowing them to administer essay assignments, which they can choose to be scored using AWE or grade themselves using scaffolded rubrics. *Researchers* will have access to a web-based tool, a downloadable tool, and editable software, which will allow them to conduct computational analyses of writing. WAT will be packaged and disseminated such that researchers and software developers can easily integrate components of WAT into existing tools to provide natural language processing (NLP) extensions in educational software.

Our aim is to provide students, teachers, and researchers with writing analytics that will directly contribute to their knowledge of writing. For researchers, this knowledge may be theoretical or computational; for teachers, this knowledge may be pedagogical and relate to developing a better understanding of linguistic and semantic features of higher quality writing and pedagogical approaches to improve writing; finally, for students, this knowledge may be metacognitive, such that they develop a better understanding of how features of language affect their audience and essay scores. Our overall aim is to provide a writing analytics tool that will enhance students' ability to produce high-quality texts across multiple genres. Thus, we aim to develop a tool with broad impact on current practices in writing research and instruction across multiple dimensions.

One of our objectives with WAT is to provide students and teachers with writing tasks that provide automated feedback. Previous projects have informed our natural language processing (NLP) algorithms to drive feedback for persuasive essays and summaries. As such, our main focus currently is to collect additional corpora of source-based essays, analyze those essays to identify important linguistic and semantic features, and develop NLP algorithms. We will discuss work with our collaborators in which we are conducting NLP analyses of source-based essays collected in previous projects as well as on-going projects.

We also invite our colleagues to join the Distributed Literacy Coalition (DLC; distributedliteracy.org), which aims to integrate laboratories distributed across the world focused on understanding and improving literacy. Distributed literacy refers to the multiple, intertwined aspects of literacy including reading and writing, as well as science, health, math, and social media literacies. DLC members work together on the common objective to improve literacy worldwide, recognizing the vital societal importance of literacy and the need for multidisciplinary and multicultural approaches to solve literacy problems.

Understanding the ‘Black-Box’ of Automated Analysis of Communicative Goals and Rhetorical Strategies in Academic Discourse

Elena Cotos

Iowa State University
ecotos@iastate.edu

Despite the appeal of automated writing evaluation (AWE) tools, many writing scholars and teachers have disagreed with the way such tools represent writing as a construct. This talk will address two important objections – that AWE heavily subordinates rhetorical aspects of writing, and that the models used to automatically analyze student texts are not interpretable for the stakeholders vested in the teaching and learning of writing. The purpose is to promote a discussion of how to advance research methods in order to optimize and make more transparent writing analytics for automated rhetorical feedback. AWE models will likely never be capable of truly understanding texts; however, important rhetorical traits of writing *can* be automatically detected (Cotos & Pendar, 2016). To date, AWE performance has been evaluated in purely quantitative ways that are not meaningful to the writing community. Therefore, it is important to complement quantitative measures with approaches stemming from a humanistic inquiry that would dissect the actual computational model output in order to shed light on the reasons why the ‘black box’ may yield unsatisfactory results.

Drawing on an ongoing project, which involves a systematic analysis of a collection of erroneous feedback produced by a genre-based AWE tool (Cotos, 2016), I will describe a hybrid – computer--driven/human-informed – approach with an exponential interpretive strand. The approach entails a linguistic investigation of the communicative goals analyzed both by AWE and the human. New heuristic taxonomies were developed to compare AWE detection and human interpretation of rhetorical intent, examine differences, and construe the nature of AWE errors. The resulting qualitative insights describe error patterns and reveal the role of linguistic features in automated detection of communicative goals. These insights help describe and interpret the reasons why error patterns in automated rhetorical analysis occur and how they may hinder computational representation of the writing construct. The findings can inform future interdisciplinary research aimed at developing augmented approaches for improving the quality of automated rhetorical feedback on student writing. In terms of immediate practical implications, the outcomes of this work can be translated to teaching and learning materials addressing possible feedback errors and providing strategies for how to use the feedback more effectively. More broadly, interpretable writing analytics can potentially power paradigmatic shifts and drive innovation at the level of research methodology, computational operationalization, interdisciplinary collaborations, and writing pedagogy – all interconnected to serve the purpose of students’ writing development.

REFERENCES

- Cotos, E. (2016). Computer-assisted research writing in the disciplines. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 225–242). Routledge: New York and London.
- Cotos, E., & Pendar, N. (2016). Discourse classification into rhetorical functions for AWE feedback. *CALICO Journal*, 33(1), 92-116.

Learning with Text: Toward a Multi-Dimensional Perspective on Text-Based Communication

Laura K. Allen, Lacey C. Zachary
Mississippi State University
{Laura.Allen, lcz15}@msstate.edu

Danielle S. McNamara
Arizona State University
dsmcnama@asu.edu

A commonly held belief among educators, researchers, and students is that high-quality texts are easier to read than low-quality texts, as they contain more engaging narrative and story-like elements. Interestingly, these assumptions have typically failed to be supported by the writing literature. Research suggests that higher quality writing is typically associated with decreased levels of text narrativity and readability. Although narrative elements may sometimes be associated with high-quality writing, the majority of research suggests that higher quality writing is associated with decreased levels of text narrativity, and measures of readability in general.

One potential explanation for this conflicting evidence lies in the situational influence of text elements on writing quality. In other words, it is possible that the frequency of specific linguistic or rhetorical text elements alone is not consistently indicative of essay quality. Rather, these effects may be largely driven by individual differences in students' ability to leverage the benefits of these elements in appropriate contexts. Indeed, recent research points to the contextual variability of linguistic features across different audiences, prompts, and assignments (Allen, Snow, & McNamara, 2016; Crossley, Roscoe, & McNamara, 2014). Crossley and colleagues (2014) for example, found that there were multiple profiles of high-quality writing, which demonstrated different linguistic properties. This evidence points toward the need to examine writing in more situated contexts.

This presentation will further explore the hypothesis that writing proficiency is associated with an individual's flexible use of text properties, rather than simply the consistent use of a particular set of properties. Across three experiments, this study relies on a combination of natural language processing, dynamic methodologies, and behavioral methodologies to examine the role of linguistic flexibility during the writing process. Overall, this study provides important insights into the role of flexibility in writing skill and develop a strong foundation on which to conduct future research and educational interventions.

REFERENCES

- Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, 108(7), 911.
- Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, 31(2), 184–214.

A Novel Writing Analytics Approach to Study Multiple Information Sources Integration by Students

Yoram M Kalman*, Ester Adam, Ina Blau

The Open University of Israel

yoramka@openu.ac.il, adamf@gedu.openu.ac.il, inabl@openu.ac.il

ABSTRACT: This poster describes a work in progress (WIP) research project that will explore the way students merge and summarize multiple information sources. The experiment examines the effect of merging digital versus analog texts on the digital writing process, the quality of the written outcomes, and the level of plagiarism in summaries written by students. The project incorporates a novel writing analytics approach that uses a logger which tracks not only keystrokes and timestamps, but also their impact on the evolving text, allowing an in-depth analysis of writing and editing processes. The study contributes to the writing analytics literature by improving our understanding of multiple information source integration and of plagiarism in student writing, as well as by offering a novel method to track and analyze computer-based writing processes.

Keywords: writing analytics, text integration, logger, plagiarism

1 MERGING AND SUMMARIZING MULTIPLE INFORMATION SOURCES

One of the top skills required by participants in the knowledge economy is that of reading multiple information sources and creating a new document that integrates these information sources in a coherent and effective manner (Barzilai, Zohar, & Mor-Hagani, 2018). A study of this skill intersects with several research themes related to reading and writing, including research on the differences between reading from paper versus from digital sources (e.g. Fortunati & Vincent, 2014; Mangen, Walgermo, & Brønnick, 2013), research on the cognitive and metacognitive processes that are associated with these integration tasks (Barzilai & Zohar, 2012), research on writing processes and their evaluation (e.g. Shibani, Knight, & Shum, 2018), and research on academic integrity in the use of information sources (e.g. Blau & Eshet-Alkalai, 2017).

2 RESEARCH QUESTIONS

The study described in this WIP poster is an experiment that requires participants to merge and summarize three texts into a single coherent digital text. The study explores three research questions:

- a. Are there differences between the processes of creating a summary document from digital sources versus paper-based information sources?
- b. Are there differences between the quality of outcomes - a summary document from digital sources versus paper-based information sources?

- c. Is there a difference in the extent of plagiarism between creating a summary document from digital sources versus paper-based information sources?

3 EDIT-TRACKING KEYSTROKE LOGGER

A unique keystroke logger is currently under development in order to study the writing process of the participants in the study. Like a regular keystroke logger, this logger tracks every keystroke performed by users as they type within an HTML window. Furthermore, with each keystroke (both down-stroke and up-stroke) the logger also records the text that is in the HTML window when the keystroke occurred. These timestamped records are then exported in a json file which contains a highly detailed record of the writing process. This json file is then analyzed using scripts that identify the various writing and editing activities performed by the users.

4 THE EXPERIMENT

In the experiment, sixty participants will be recruited and randomly assigned into two groups. Both groups will be asked to merge and summarize three identical texts, either digital (group A) or paper based (group B). Both groups will perform the merging using the logger described in section 3 above. The three RQs will be explored by analyzing the writing process as well as the resultant text written by the participants. This novel writing analytics approach contributes to our understanding of multiple information source integration, of digital versus paper-based reading and writing, and of student plagiarism. It also presents a novel method for tracking and analyzing computer-based writing processes.

REFERENCES

- Barzilai, S., & Zohar, A. (2012). Epistemic Thinking in Action: Evaluating and Integrating Online Sources. *Cognition and Instruction*, 30(1), 39–85. <https://doi.org/10.1080/07370008.2011.636495>
- Barzilai, S., Zohar, A. R., & Mor-Hagani, S. (2018). Promoting Integration of Multiple Texts: a Review of Instructional Approaches and Practices. *Educational Psychology Review*, 30(3), 973–999. <https://doi.org/10.1007/s10648-018-9436-8>
- Blau, I., & Eshet-Alkalai, Y. (2017). The ethical dissonance in digital and non-digital learning environments: Does technology promotes cheating among middle school students? *Computers in Human Behavior*, 73, 629–637. <https://doi.org/10.1016/j.chb.2017.03.074>
- Fortunati, L., & Vincent, J. (2014). Sociological insights on the comparison of writing/reading on paper with writing/reading digitally. *Telematics and Informatics*, 31(1), 39–51. <https://doi.org/10.1016/j.tele.2013.02.005>
- Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58, 61–68. <https://doi.org/10.1016/j.ijer.2012.12.002>
- Shibani, A., Knight, S., & Shum, S. B. (2018). Understanding Revisions in Student Writing Through Revision Graphs. In *International Conference on Artificial Intelligence in Education* (pp. 332–336). Springer.

Am I planning smart? – Analyzing student goals

Sebastian Wollny

DIPF | Leibniz Institute for Research and Information in Education
wollny@dipf.de

Jan Schneider

DIPF | Leibniz Institute for Research and Information in Education
schneider.jan@dipf.de

Marc Rittberger

DIPF | Leibniz Institute for Research and Information in Education
rittberger@dipf.de

Hendrik Drachsler

DIPF | Leibniz Institute for Research and Information in Education
Goethe University, Frankfurt am Main
Open University of the Netherlands
drachsler@dipf.de

ABSTRACT: Goal setting is an important step in Self-Regulated Learning. Setting goals is not a straight forward task. Some types of goals are more useful than others. The SMART goal setting guideline helps to generate more meaningful goals. In this paper, we present a research roadmap designed to assist learners with the generation of meaningful learning goals. The roadmap consists of a three-stage process: structure goal extraction, continuous text goal extraction, and dialogue-based goal extraction. Findings from each of the stages will support with the implementation of the next one.

Keywords: NLP, Learner Goals, Recommender Systems, Self-Regulated Learning, Chatbot

1 BACKGROUND

Self-regulated Learning (SLR) describes the area of learning strategies, self-assessments, and self-reflection of learners. Learning planning and goal setting is a crucial process of SRL that allows learners to draw conclusions from the learning process through self-reflection (Zimmerman & Moylan, 2009). Goals can be defined in many ways, nevertheless not every formulation is of equal value. By the requirements of the well-known SMART Framework (Doran, 1981), they can be evaluated through a simple set of rules.

With the increasing digitalization of our everyday lives, written texts are gaining more and more importance. For many students, writing text messages has become the preferred method of communication, which they use to communicate with others (Rideout & Robb, 2018). Popular extensions of these classic text messages are chatbots and digital assistants. They open up new

possibilities in the networking of learners and learning support systems by using the same communication channels (Winkler & Söllner, 2018).

We want to help learners with their goal setting by offering a system that can be operated in natural language. Such a dialogue-oriented system should give students the opportunity to compose goals and track their achievements in the context of SRL (Locke & Latham, 1990).

In this article, we present our research roadmap of a system that starts with the evaluation of written learning goals and leads into a dialogue-based learning tool for goal setting. This research roadmap follows the design-oriented approach (Wang & Hannafin, 2005), in which context and theory are examined in an iterative process.

2 THEORETICAL BACKGROUND

The basis of the following goal extraction is the SRL theory. A popular model in SRL is the three phases model of (Zimmerman & Moylan, 2009). It describes an SRL cycle with Forethought Phase, Performance Phase and Self-Reflection Phase (fig. 1).

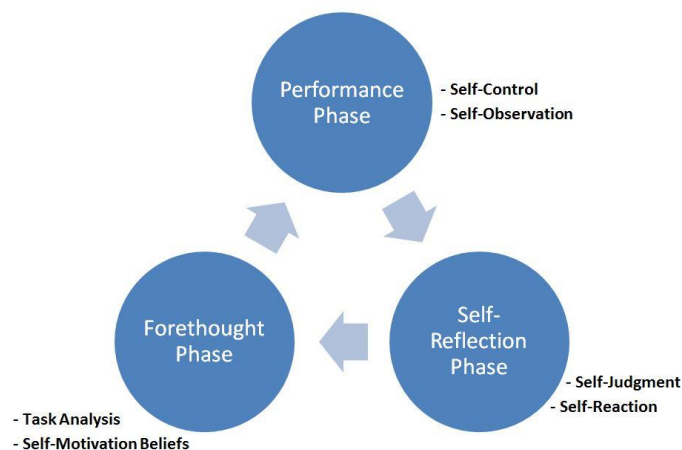


Figure 1: A cyclical phase model of self-regulation (Zimmerman & Moylan, 2009)

Many applications of student facing Learning Analytics can be assigned to the second phase, where learners observe themselves within the learning process. With the introduction of a goal dialogue system, we plan to contribute to the learning planning phase of SRL, which is in many cases overlooked (Jivet, Scheffel, Drachsler, & Specht, 2017).

3 SMART GOAL SETTING

In order to be meaningful, goals should inherit several features as defined by (Doran, 1981). This guideline consists of the acronym “SMART”, which says that goals should be:

- **S**pecific
- **M**easurable

- Assignable
- Realistic
- Time-related

The following example (fig. 2) is intended to illustrate the features of a SMART goal:

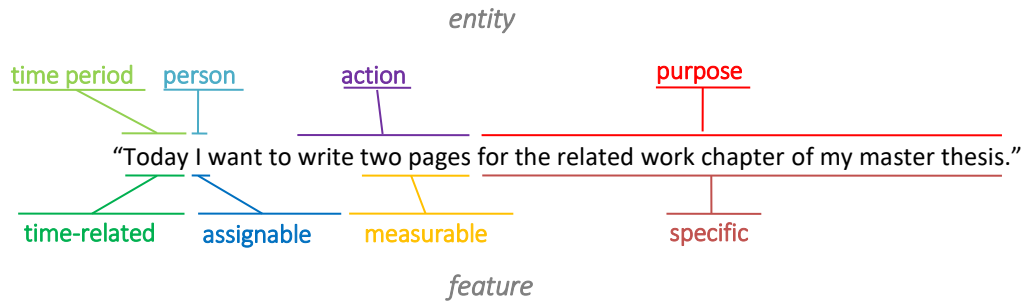


Figure 2: SMART goal example

As this example shows, many features match entities. It turns out, the assessment of the realistic feature is not included in the wording and strongly dependent on author and context. The SMART guideline contains the idea that the progress of goal achievement has to be assessed in the future. Therefore, the measure is strongly related to the defined time period, which represents a deadline. The specificity feature can be described as a connection between the actual action and a superior intention. It can be seen as a hierarchy of goals, in which the achievement of subordinate goals also benefits superior goals (Cropanzano, James, & Citra, 1993).

4 RESEARCH ROADMAP

Our Research roadmap has the purpose to create a system that helps learners to set smart learning goals. It consists of a three-stage process, which leads from a structured to a dialogue-oriented input (fig. 3). These stages are defined as:

- Stage 1 – Structured Goal Extraction
- Stage 2 – Continuous Text Goal Extraction
- Stage 3 – Dialogue-based Goal Extraction

In the transition from one stage to another, learners gain degrees of freedom in the possibility of defining goals. This increases the variability of the used wording and requires more complex extraction rules and procedures.



Figure 3: Research Roadmap for extracting learning goals

Stage 1 - Structured Text Goal Extraction

This stage is the beginning of the roadmap and focuses on the extraction of goals from a predefined wording. It simplifies the definition of learning goals to one sentence, which has to be completed by the learners. As already seen in fig. 2, SMART features are comprised by textual entities. The most variable entities in this context are actions and purposes. With actions, learners describe conditions for achieving a goal, while purposes are used to place goals in a higher context.

Actions should be examined for measures (see chapter 3). These measures could be countable numbers or a set of verbs describing a state of progress (like “*finish*” or “*complete*”).

In the following two subsections (4.1.1 and 4.1.2) we present some example structures which enable a SMART analysis of learning goals. They are exemplarily designed for one week, in order to create a useful SRL cycle. The assessment of the closeness to reality can only be covered by an additional input field. As mentioned in chapter 3, this information is not included in the goal formulation.

Time-period-based Goal Formulation

With a time-period-based goal formulation, learners can set an action to a purpose. It can be formulated as follows and is a flexible structure for one-time conditions:

“ This week I want to **[action]** to **[purpose]**. “

Event-based Goal Formulation

Through an event-based goal formulation, learners can define focus events within a time period. Every time this event occurs, the learner defines a specific action to perform. The wording can be chosen as follows:

“ Every time I **[event]** this week, I want to **[action]** to **[purpose]**. “

In contrast to time-period-based extraction, an additional condition (event) is involved. It should, therefore, be chosen in such a way that it occurs frequently in the time period. A predefined selection of events can, therefore, be considered as a simple solution.

Stage 2 - Continuous Text Goal Extraction

This stage is concerned about goal extraction from continuous text. By further opening the goal formulation, it extends the structured text goal extraction through goal extractions from textboxes. This enables learners to freely define goals in their preferred sentence structures. The Continuous Text Goal Extraction stage has to deal with more varieties of SMART learning goals and should include sentence analysis, POS analysis, and entity extraction. It should include feedback in the form of recommendations to improve learning goals (Verbert et al., 2012), which can be achieved by a set of tips. These can be shown if a particular feature of the SMART guideline could not be found in the goal formulation.

Stage 3 - Dialogue-based Goal Extraction

This stage is the end of the roadmap and marks the dialogue-based goal extraction. It defines a conversational extension to the continuous text goal extraction, which is able to extract goals from a conversational dialogue, question on goal formulations and provide examples how to define SMART goals. This stage should ideally be integrated in a chatbot-system that tries to model goals as described in (Brusilovsky & Millán, 2007).

5 USE CASE SCENARIO

Our roadmap of goalsetting and applied goals could be integrated into SRL diaries. They should enable not only to document one's own learning progress but also to set goals and evaluate their achievement. A dialogue-based goal extraction with an intuitive interface would enhance these systems. It would help students in defining meaningful goals for their SRL cycle by asking questions and recommending improvements.

6 OUTLOOK

The goal extraction mechanisms proposed in this paper can help learners to define and keep track of meaningful goals. In the next step, we plan to follow our research roadmap in order to implement such a system and study its effects. It should show insights about the possibilities and limitations of its use, which result from the entire roadmap process.

REFERENCES

- Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web* (pp. 3–53). Springer.
- Cropanzano, R., James, K., & Citera, M. (1993). A goal hierarchy model of personality, motivation, and leadership. *Research in Organizational Behavior*, 15, 267.
- Doran, G. T. (1981). There's a SMART way to write management's goals and objectives. *Management Review*, 70(11), 35–36.
- Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (Eds.) 2017. *Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice*: Springer.

- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*: Prentice-Hall, Inc.
- Rideout, V., & Robb, M. B. (2018). *Social media, social life: Teens reveal their experiences*. San Francisco, CA: Common Sense Media.
- Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: A survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4), 318–335.
- Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), 5–23.
- Winkler, R., & Söllner, M. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis.
- Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 311–328). Routledge.

Towards knowledge-transforming in writing argumentative essays from multiple sources: A methodological approach

Mladen Raković
mrakovic@sfu.ca

Zahia Marzouk
zmarzouk@sfu.ca

Daniel Chang
dth7@sfu.ca

Philip H. Winne
winne@sfu.ca

Simon Fraser University

ABSTRACT: Skillful essay writers successfully transform knowledge from multiple sources. However, when post-secondary writers draft essays after researching the articles, they often face challenges to engage in knowledge transforming, a complex process simultaneously involving reading comprehension, writing production and metacognitive monitoring (Bereiter & Scardamalia, 1987). We describe a two-facet methodological approach to model linguistic properties that distinguish knowledge-telling evidential sentences from knowledge-transforming ones in disciplinary argumentative writing. We collected and coded 40 post-secondary disciplinary argumentative essays based on an assigned argumentation framework and Bloom's taxonomy (Sadker & Sadker, 2006). We use these coded argumentation schemes to develop a computational tool to generate writing analytics to scaffold writers towards more knowledge transforming processes.

Keywords: Argumentation, writing, text analysis, knowledge telling, knowledge transforming

1 INTRODUCTION

To develop well-structured arguments in essays, students need to form and present claims and adjoin credible evidence to support arguments. This entails successfully navigating between a rhetorical problem space and a content problem space (Bereiter & Scardamalia, 1987). In the rhetorical problem space, students work to design, structure, and precisely and coherently communicate claims and supportive evidence. Solving rhetorical problems accomplishes argumentative goals. Simultaneously, in the content problem space, students process information they identify and mine from multiple sources. As they compare facts, reasons and explanations, evaluate and generalize findings, and establish semantic relationships among key concepts, opportunities arise to coordinate evidence relating to claims positioned in the rhetorical space.

In this process, students actively rework drafts to fit parameters of the writing task and its goals. Bereiter & Scardamalia (1987) modeled interactions among discourse and content processing, and metacognitive monitoring as a composite process called knowledge transforming. Because this process triggers reflective thinking while writing, Bereiter & Scardamalia (1987) argue that knowledge transforming promotes learning.

Producing knowledge-transforming texts is a challenge for many post-secondary writers. Research indicates student writers often fail to paraphrase, interpret, and evaluate content in sources; construct novel associations across multiple sources; and integrate multiply-sourced information into a coherent structure (Bereiter & Scardamalia, 1987; Aull, 2015; Boscolo, Ariasi, Favero, & Ballarin, 2011; Dong, 1996; Flower et al, 1990; Petrić, 2007). As a result, under-skilled post-secondary writers often engage in a more limited text production process termed knowledge telling. Writers who generate knowledge-transforming text typically use monitoring and planning strategies that develop

a coherent text. In contrast, writers who produce knowledge-telling texts focus overly on generating basic text, e.g., staying on topic and repeating facts from sources. In the knowledge-telling process of writing, interactions between the content problem space and the rhetorical problem space are few, limited in complexity and unproductive. We hypothesise writing analytics can be generated to help struggling writers move from knowledge telling toward knowledge transforming. Such analytics should invite writers to engage in knowledge transforming processes while practicing writing, reading, and arguing strategies that help them navigate between the content and rhetorical spaces.

We present a methodological approach to identify knowledge transforming in evidential sentences situated in disciplinary argumentative essays generated by post-secondary students. Specifically, we seek to identify when students transform source information by applying evidence to promote argumentative claims. Hemberger, Kuhn, Matos, & Shi (2017) posited that coordinating evidence with claims is essential to skilled argumentative writing. Thus, the final goals of our research are (a) to develop an ensemble of computational algorithms to analyze linguistic properties of evidential sentences in an argumentative essay relative to information available in sources, and (b) generate learning analytics that scaffold knowledge transforming as writers bring evidence to support claims. The computational tool will use linguistic properties of evidential sentences as standards for tailoring learning analytics in form of metacognitive prompts to writers helping them go beyond merely restating information borrowed from sources to engage in knowledge transforming.

2 RELATED WORK AND THEORETICAL MODEL

Citations in an essay – references to and quotes of source information – have been classified with respect to various linguistic functions (see Petrić, 2007). We elaborated Bereiter and Scardamalia's (1987) model contrasting knowledge telling and knowledge transforming by additionally categorizing evidential sentences in argumentative writing in terms of Bloom's taxonomy of the cognitive domain (Sadker & Sadker, 2006; Table 1). The taxonomy describes a progression of thinking processes across knowledge, comprehension, application, analysis, synthesis and evaluation. While not without criticism (e.g., see Darwazeh, 2017) it has potential to supply an underlying framework for developing informative, specific and useful learning analytics to guide learners in advancing from knowledge-telling to knowledge transforming. According to Bereiter and Scardamalia's (1987) writing model, students engaged in knowledge telling neglect cognitive and metacognitive operations that transform knowledge. Using Bloom's taxonomy to classify writers' evidential sentences could reflect underlying cognitive and metacognitive processes writers engage in. Bloom's knowledge classification aligns with Bereiter and Scardamalia's knowledge-telling model where writers focus on generating basic text. Bloom's comprehension, application, analysis, synthesis and evaluation categories reflect Bereiter and Scardamalia's knowledge transforming category where writers coordinate and create knowledge. Thus, classifying students' evidential sentences in terms of

Table 1: Framework for classifying evidential sentences in argumentative writing

Category	Operationalization	Writing Mode
Knowledge	paraphrased/copied information from a source	Knowledge telling
Comprehension	elaborated source information	Knowledge
Application	source information applied to the real-world context	Knowledge
Analysis	inferential additions to information mentioned in sources	Knowledge
Synthesis	integrating information from different sources or a proposition	Knowledge
Evaluation	evaluating or discrediting source information	Knowledge

3 METHOD

3.1 Corpus and writing task

Our corpus was 40 argumentative essays written by undergraduates enrolled in various disciplinary majors and registered in an introductory educational psychology course in a Western Canadian university. Students were assigned a 1500-2000 word argumentative essay on a specific disciplinary issue of their choice. Essays were required to present (a) at least three arguments supported with evidence gathered from 5-7 sources students selected from 160 sources in the course repository, (b) at least one counterargument with evidence, and (c) rebuttal(s) to the counterargument(s).

3.2 Hand coding – codebook

Sentences were sampling units. Since we focus on analyzing arguments and evidence, we coded sentences in the essay body (excluding the introduction paragraph, conclusion paragraph, and headings) in terms of argumentation, writing mode and relationality.

For argumentation, we coded sentences in one of five categories: Argument (A), a sub claim supporting the thesis statement (main claim); Evidence (E), sentences providing support to the argument; Counterargument (C), counter claims; and Rebuttal (R), sentences discrediting the counterargument; Not applicable (NA), a sentence that did not fit any argumentation category, e.g., definition or background information. For Writing mode, categories (Table 1) referred to Bereiter and Scardamalia's knowledge transforming model (1987) elaborated by Bloom's taxonomy of the cognitive domain following Sadker & Sadker (2006). A 3-point scale quantified relationality in terms of each argument's (or sub argument's) linkage to the thesis statement (or main argument), and the relation of evidence to arguments (sub arguments): 0 indicated not related, 1 described far-fetched, and 2 described related. The coding method is illustrated in the Figure 1. The sentence coded as argument (A) receives a rating on its relation to thesis statement. The sentence coded as evidence receives a rating on its relation to the preceding argument.

3.3 Hand coding – codebook

To reach high interrater agreement among three coders, coding proceeded in three rounds of train together → code independently → calculate reliability. In round 1, two randomly selected essays were collaboratively coded followed by independently coding four randomly selected essays.

Sentence	Macro Structure	Argumentation	Writing mode	Relation to thesis statement/ argument
Meeting the different needs of learners and allowing them to be included in classrooms can result in children achieving educational success.	Intro			
Learner differences should be a primary concern when it comes to educating teachers and achieving inclusion, as the failure to incorporate learning needs can be disastrous for all students.	Intro			
While most schools focus on bringing underachieving students up, individuals who are of high ability are neglected.	Body	A		2
According to Northwestern University (2017), children are then left to rely on their parents to provide them with advanced instruction.	Body	E	Knowledge	2
Therefore, many students miss out on opportunities for achievement as many families cannot provide them with the resources such as tutoring services or enrichment activities.	Body	E	Comprehension	2
When teachers are given appropriate instruction, they are able to teach learners who need support.	Body	NA		

Figure 1: Codebook

Altogether, those four essays comprised 28 paragraphs (per text: $M=7$, $SD=1.41$) and 245 sentences (per paragraph: $M=8.75$, $SD=3.63$). After independent coding, we calculated reliability using the AC1 statistic (Gwet, 2002) as this method corrects agreement among raters for the probability of chance agreement. Although inter-rater reliability was lower for Argumentation and Writing mode (0.67 and 0.77, respectively), differences arose in identifying argumentation categories because coders' failed to reliably identify evidential sentences. In addition, for Writing mode, coders struggled to discriminate synthesis from analysis, and analysis from comprehension. For round 2, we sharpened coding of Argumentation and Writing mode. In round 2, three coders coded two randomly selected student essays collaboratively followed by independently coding four randomly selected essays. Altogether, the four essays comprised 26 paragraphs (per text: $M=5.2$, $SD=1.3$) and 247 sentences (per paragraph: $M=9.27$, $SD=2.47$). Reliability of the argumentation mode was still low (0.76). Round 3 included collaboratively coding two randomly selected student essays followed by independently coding six randomly selected essays. Table 2 presents final inter-rater reliability results.

Table 2: IR reliability after the 3 rounds of “train together-code independently-calculate reliability”

Code	AC1 Reliability	Standard Error	95% CI
Macro-structure	0.97	0.01	[0.95, 0.99]
Argumentation	0.81	0.02	[0.77, 0.84]
Writing mode	0.83	0.02	[0.78, 0.87]
Relation to arguments/thesis	0.82	0.02	[0.78, 0.86]

In the Appendix, we illustrate codes within the Writing mode for each category of Bloom's taxonomy (Sadker & Sadker, 2006).

3.4 Extracting linguistic indices for sentences coded in Writing mode scheme

We propose modeling the following linguistic indices for each identified evidential sentence. The variables are grouped into: anaphoric devices, semantic overlap, and rhetorical connectives.

First, high accessibility (unstressed pronouns) and low accessibility anaphoric devices (full noun phrases and indefinite articles) will be computationally extracted. Sanders & Spooren (2007) pinpoint high accessibility markers in a sentence indicate continuation with previous topic, or the writer's tendency to stay on topic. Both are signs of knowledge-telling. Low accessibility markers, on the other hand, signal termination of current and activation of other topics. They indicate knowledge-transforming.

For each evidential sentence we will compute its semantic overlap with source text and with the preceding sentence (argument/counterargument/rebuttal/evidence). We hypothesize knowledge-telling sentences have higher semantic overlap with a source while knowledge-transforming sentences have lower semantic overlap with the source and the preceding sentence.

Seventeen rhetorical connectives will be calculated using the TAACO tool (see Crossley, Kyle & McNamara, 2016). We anticipate subsets of rhetorical connectives will predict knowledge telling and transforming. The analysis will provide substantial details.

REFERENCES

- Aull, L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*. Springer.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Routledge.
- Boscolo, P., Ariasi, N., Del Favero, L., & Ballarin, C. (2011). Interest in an expository text: How does it flow from reading to writing? *Learning and Instruction*, 21(3), 467-480.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4), 1227-1237.
- Darwazeh, A. N. (2017). A new revision of the [Revised] Bloom's taxonomy. *Distance Learning*, 14, 13-28.
- Dong, Y. R. (1996). Learning how to use citations for knowledge transformation: Non-native doctoral students' dissertation writing in science. *Research in the Teaching of English*, 428-457.
- Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6), 1-6.
- Hemberger, L., Kuhn, D., Matos, F., & Shi, Y. (2017). A dialogic path to evidence-based argumentative writing. *Journal of the Learning Sciences*, 26(4), 575-607.
- Petrić, B. (2007). Rhetorical functions of citations in high- and low-rated master's theses. *Journal of English for Academic Purposes*, 6(3), 238-253. <https://doi.org/10.1016/j.jeap.2007.09.002>
- Sanders, T., & Spooren, W. (2007). Discourse and text structure. *Handbook of cognitive linguistics*, 916-941.
- Sadker, M. & Sadker, D. (2006). Questioning skills. In J. Cooper (Ed.), *Classroom teaching skills* (8th ed., pp. 104 – 150), Boston, MA

APPENDIX**Sample coded sentences**

Category	Example Sentence
Knowledge	<i>When institutions and classrooms integrate self-directed learning into their curriculum, long term benefit have been observed through increased student retention and graduation rates (University of Texas at Austin, 2016).</i>
Comprehension	<i>With this type of learning, students can fully control their educational experience and focus on information they would like to explore.</i>
Application	<i>Having different interpretations based on cultural differences is a concern, particularly for schools in British Columbia and other Canadian metropolitan centers where we have and are projected to receive more international students particularly from Asia.</i>
Analysis	<i>Meaning engagement in some form of unstructured play could also result in an increase in academic performance.</i>
Synthesis	<i>However, this is not the case, because praise is not overly useful feedback, and if it is undeserved, it can cause students to feel like their teachers do not expect much from them.</i>
Evaluation	<i>One of the limitations is that the research is centered on a questionnaire survey which may result in certain biases including social desirability bias.</i>

Analyzing learners' online behaviour for student success and course enhancement: Case-studies from Blackboard

Christine Armatas

The Hong Kong Polytechnic University

christine.armatas@polyu.edu.hk

Chun Sang Chan

The Hong Kong Polytechnic University

chun.sang.chan@polyu.edu.hk

Ada Tse

The Hong Kong Polytechnic University

ada.sk.tse@polyu.edu.hk

ABSTRACT: The large amount of data recorded about student behavior in a learning management system (LMS) is only useful if it can be accessed, analysed and interpreted easily and on demand. Through a series of case-studies, we demonstrate an easy to use Excel tool developed specifically for teachers to understand their students' online activity and enhance their teaching. Activities based on the case-studies illustrate how to use the tool to conduct analysis of a sample LMS data-set (which is also provided) to produce tables, figures and visualizations about student engagement in the LMS. The case-studies demonstrate how the indicators provided in the tool are informative and actionable for enhancing online teaching. The analyses used in the case-studies focus on helping students be more successful while studying, as well as how to use analysis of LMS data to enhance learning and the student experience for future course delivery. The tool gives users autonomy in accessing and analyzing students' online activity which can be used for evidence-based teaching enhancement.

Keywords: analysis of online behavior; enhancing outcomes for students; case-studies in learning analytics

1 BACKGROUND

With the increasing adoption of blended and online learning, teachers want to know what students are doing online and how this impacts on their learning. As a result, teachers want learning analytics tools that allow them to easily conduct analyses on their course data so they can understand students' activity in the learning management system (LMS) and gain actionable insights for enhancing teaching and learning. We have developed an Excel-based tool that allows teachers to do this and which puts powerful analyses and visualisations of LMS usage data in the hands of teachers.

This tool and the case-studies used to illustrate its application provide a structured means for exploring how LMS data can be used to help students at the time they are studying a course by

understanding students' engagement with the online environment and the impact this has on their achievement of learning outcomes. This in turn can assist with promoting student success. How analysis of LMS data can do this is explored through case-studies based on analysis of students' online behavior as recorded in de-identified Blackboard logs for courses taught at our university. The case-studies include using analytics while the course is being taught as well as after the course is finished to make improvements for future delivery. After working through the activities for each of the case-studies, users with access to Blackboard courses will be able to use the tool to access, analyse and visualize data from their own students using the case-studies as a reference. It is expected that users will find this user-friendly and on-demand approach to understanding students' online behavior informative and useful.

1.1 About the analysis tool

As Hackbarth (2017) notes, learning analytics tools for classroom teachers should be easy to use and can stimulate and satisfy teachers' curiosity about student learning. In keeping with this, our Excel tool is supported by VBA code and Excel add-ins and has been developed to extract Blackboard usage data from course archives, analyze students' LMS data at course level and generate useful tables and visualizations for teachers to understand what students do online. The advantage of developing this learning analytics tool in Excel is that Excel is an environment which many teachers already know and use and almost all teachers already have installed on their computers. This in turn helps with acceptance and adoption of the tool. The tool makes it easy for teachers using Blackboard to obtain and analyse LMS data in their own time and without specialist assistance. Teachers only need to go to their Blackboard course and make use of the archive course function to produce a zipped file which is a permanent record of a course including all the content and user interactions. After downloading the zipped file, teachers simply press a button in the Excel tool to import the data from the zipped file and the tables and visualizations are automatically generated for them. Users can vary the period of analysis and can conduct their own analyses using additional modules provided in the tool. Overall, the tool gives teachers the ability to monitor students' learning and to study the impact of, or evaluate, their teaching strategy during specific time periods.

1.2 About the case-studies

The tool and case-studies can be requested from the authors. Instructions on how to extract the data from Blackboard and import it into the tool are provided for those wanting to work with their own Blackboard data. For the case-studies, an LMS data set is provided which can be imported into the tool and used to explore several scenarios we have developed.

1.2.1 Case-study A: Using usage data to support students while studying

The tool automatically generates usage summaries and individual student usage profiles as shown in Figures 1 and 2. A correlation matrix showing the relationship between entries in the Grade Center and feature usage is also generated automatically based on the data from the course archive (see Figure 3.) The first case-study explores how to use the information generated for "just-in-time" identification and support of students who may be at risk or not engaging online by looking at the information in these three worksheets.

Dashboard - Overview

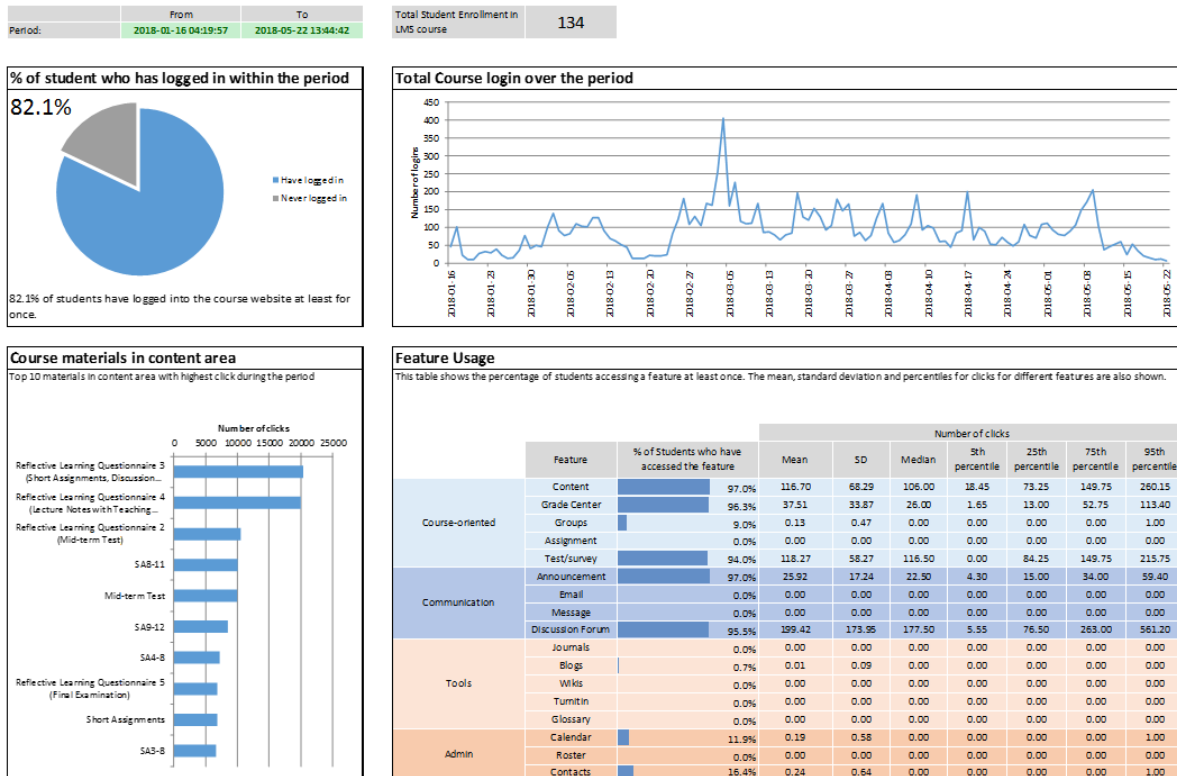


Figure 1: Worksheet "Overview" showing the students' LMS activity in a class

Student Usage on Blackboard features and course materials																			
User ID	User Name	Course-oriented					Communication				Tools					Admin			
		Content	Grade Center	Groups	Assignment	Test/survey	Announcement	Email	Message	Discussion Forum	Journals	Blogs	Wikis	Turnitin Assignments	Glossary	Calendar	Roster	Contacts	
123456	Stud1	119	26	0	0	84	32	0	0	16	0	0	0	0	0	0	0	0	
123457	Stud2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
123458	Stud3	127	42	0	0	112	32	0	0	158	0	0	0	0	0	0	0	0	
123459	Stud4	96	21	0	0	120	14	0	0	260	0	0	0	0	0	0	0	0	
123460	Stud5	184	32	0	0	87	40	0	0	185	0	0	0	0	0	0	0	0	
123461	Stud6	79	24	0	0	121	19	0	0	29	0	0	0	0	0	0	0	0	
123462	Stud7	72	15	0	0	117	16	0	0	164	0	0	0	0	0	0	0	1	
123463	Stud8	69	7	0	0	99	9	0	0	284	0	0	0	0	0	0	0	0	
123464	Stud9	111	17	0	0	115	25	0	0	127	0	0	0	0	0	0	0	0	

Figure 2: Worksheet showing individual student's activity in different features of the course

Import Data and grades
Export forum posts
Prediction
Generate student list
Email via outlook
Email Selected Students
About

D20

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Correlation between clicks on features and assessment grades														
2															
3		Course-oriented					Communication				Tools				
4	Grade Center Items	Content	Grade Center	Groups	Assignment	Test/survey	Announcement	Email	Message	Discussion Forum	Journals	Blogs	Wikis	Turnitin Assignments	Glossary
5	Week 1 Test	-0.008	-0.380	0.718	0.755	-0.752	0.115	0.115						-0.907	
6	Manually created column	0.269	0.154	-0.808	-0.808	0.655	0.115	0.115						-0.078	
7	Week 3 test	0.027	0.037	-0.515	-0.446	0.123	-0.515	-0.515	-0.027	0.181	0.217	0.461	-0.250	0.217	0.461
8	Writing task 1	-1.000	1.000	1.000	1.000	1.000					-1.000		-1.000		

Figure 3: Worksheet showing the correlation between LMS usage and students' achievement

1.2.2 Case-study B: Analysing discussion post data

Analysis of patterns of participation in Blackboard discussion forums is also automatically generated by the tool to show who is posting and when (see Figure 4.). How to make use of this information to support students when engaging in discussion for the course is explored in the second case-study.

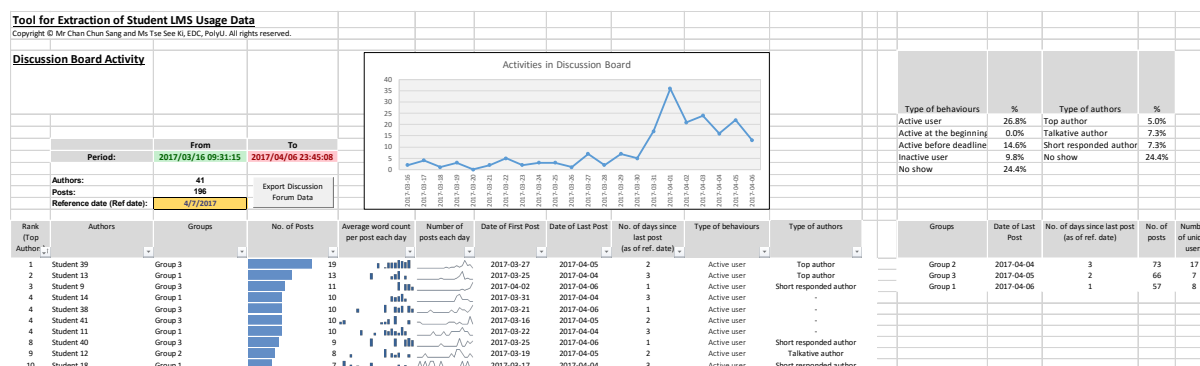


Figure 4: Worksheet about activities in the discussion board

1.2.3 Case-study C: Understanding what predicts student success

The final case-study demonstrates using the tool to analyse student LMS usage data and patterns to understand what predicts student success and how to enhance the course for future delivery. Built-into the tool is a prediction module which allows users to easily build and test predictive models based on results in the Grade Center and Blackboard usage data. How to use the prediction module to understand what predicts student success in the course is explored via the case-study, together with how to use this information to inform and enhance future delivery of the course.

2 OBJECTIVES AND OUTCOMES

Completing the activities in the case-studies allows users to:

- Develop an understanding of how analysis of usage logs can be used to support students while studying a course and for enhancing the course for future offerings;
- use the tool provided to analyse and interpret LMS usage data from one of their own courses, both during and after the course, to inform student support and course enhancement.

REFERENCES

Hackbarth, A. J. (2017, March). Are We Losing Sight of the Trees for the Forest? A Case for Localized Longitudinal Analytics. In Shehata, S. & Tan, J.P-L. (Eds.), *Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK17)* (pp.77-82). Simon Fraser University, Vancouver, Canada: SoLAR.



The 9th International

Learning Analytics & Knowledge Conference

Tempe, Arizona

March 4-8, 2019 #LAK19

hosted by



Learning Analytics Deployment Tactics: A meta-workshop

Pablo Munguia

RMIT University

pablo.munguia@rmit.edu.au

ABSTRACT: Learning analytics is a young field and beyond its research space its uptake has been slow across academics. Often, top down strategies are not easily adopted or focus on metrics that may not align across all disciplines in a university while bottom up approaches, while well focused have difficulty increasing their reach and capacity. Ultimately, designing a professional development plan in a university is not enough at best, and incorrect at worst. This workshop focuses on developing strategies on how create interest with academics and other units to help improve the student experience. The workshop is split into two half-day sections. The first focuses on the components of that strategy such as the data sets needed, the visualization tools and the analytical solutions, and how to combine these to ensure they can cater to different disciplines. The second focuses on developing tactics for increasing engagement with learning analytics solutions across a university or large unit. The workshops will be run as a blended course where participants are encountering the material first hand, and their reflections provide solutions for designing the engagement strategies in their respective institutions.

Keywords: outreach, up-skilling, professional development, university strategy, analytics for academics

Organisational details of proposed event:

Type of event: workshop

Proposed schedule and duration: full-day split into two half day sessions.

Type of participation: 'open' workshop (i.e., any interested delegate may register to attend)

The workshop/tutorial activities that participants should expect: small group activities, discussion groups, interactive.

Expected participant numbers: 10-15

Planned dissemination activities to recruit attendants: LAK Newsletter, targeted emails,

Required equipment: Screens to project from laptops.

1 INTRODUCTION

Learning analytics is coming to town. Many universities have now adopted policies to ensure proper use and storage of student generated data (Tsai and Gasevic 2017), and research within the learning analytics field is maturing to scope mechanisms that improve the student experience beyond a

single course into program and even school levels (Knight et al 2016, Deakin Crick et al 2017). Practitioners of learning analytics have emphasized the importance of stakeholder engagement, whether teachers or services (Greller and Drachsler, 2012, Colvin et al., 2015, Arnold et al., 2014, Tsai et. al., 2018). The rationale is clear, the learning analytics field is not just research-focused but it provides a pathway to improve the teaching service and as such it requires willing customers.

Staring at data or analyses of your own performance as a teacher can be confronting and challenging, and perhaps daunting if you are not numerically inclined. University-wide strategies such as initiating “professional development” courses may work, but often encounter obstacles. These challenges are scale-related, a solution needs to help different disciplines in a university, ensure the metrics are well understood, and allow for diverse feedback to help improve the analytical solution. Ultimately, designing a professional development plan is not enough at best, and incorrect at worst. The alternative approach stems from individual academics sharing their learning analytics practice with fellow teachers. Here, the challenges include the rate of adoption across the university, and the generation of support to help disseminate the uptake.

How can we equip academic staff with the right tools and increase their engagement (or design better tools)? This workshop will be run as a meta-workshop, where participants will be experiencing a simulation of how academics could engage with activities designed to equip academic staff with the knowledge(s) needed to engage with learning analytics. In turn, the insights by the workshop participants as subjects of the exercises will help generate strategies that can then be shared with their home institutions.

This workshop is split into two sessions, and participants are welcome to attend one or both. The first focuses on the engagement strategy components such as the data sets needed, the visualization tools and the analytical solutions, and how to combine these to ensure they can cater to different disciplines. The second focuses on developing tactics for increasing engagement with learning analytics solutions across a university or large unit. Information will be provided before the session and will involve simple activities to initiate reflection that will help drive the workshop. The workshop will rely on a series of sessions designed in Canvas LMS for program managers at RMIT University to help them engage with data at course and program level. This workshop will also gain insights from a round table discussion on how to engage with academics taking place during the Australian Learning Analytics Summer Institute in November 2019.

INTENDED OUTCOMES

The takeaway insights from participants are: (a) identifying datasets that may be useful to different analytical skill levels (b) methods to present the analyses to ensure all starting skill levels can successfully engage (c) developing communications and workflow strategies to increase the scale of uptake (d) develop feedback plans to ensure a sustainable institutional model is in place. The sections below outline the proposed activities.

MODULE DESIGN AND PRE-SESSION ACTIVITIES

We will be relying on a course designed in Canvas for this purpose, and participants will be enrolled as ‘teaching staff’ interacting with information that is typically available for teaching staff. The modules of activities can be broken into sections such as: where to find the data and what it represents (Figure 1), data associated with course level learning analytics, or data associated with program level learning analytics (Figure 2).

The pre-session activities involve a short list of tasks to help participants engage with the Canvas module and reflect on what is useful and what is missing if they were to be teachers engaging for the first time with learning analytics. These activities are expected to take one hour.

Workshop A (i.e., morning session)

The first workshop focuses on the components of an engagement strategy such as the data sets needed, the visualization tools and the analytical solutions, and how to combine these to ensure they can cater to different disciplines. There are three main sections here, first introducing data to academics (e.g., Figure 1). Second, how to use the tools and datasets available (e.g., Figure 2). Particular attention is designing tools that can be used by people with interest in a quick and shallow understanding of what is happening in a course, and those with a deep dive that will be seeking raw data or coding scripts. Third, how to generate self-reflection, sharing information and insights amongst colleagues (Figure 3). Different tools that have been currently developed will be shared amongst workshop participants including their instruction manuals, allowing for ways to improve the visuals needed in the teaching practice (e.g., Figure 4).

Given these objectives, the workshop will be ran through small teams that will quickly triage strategies to engage within the activities for academics but also how to increase the span amongst academics.

Workshop B (i.e., afternoon session)

The second workshop focuses on developing tactics for increasing engagement with learning analytics solutions across a university or large unit (e.g., faculty or college). These tactics involve marketing solutions, top-down, and bottom-up approaches. Particular attention will be given to policy within the university and ways to reduce anxiety in academics when sharing their insights and ways to improve their practice. Session two will be relying on small team development of ideas and then testing these amongst the broader group.

REFERENCES

- Arnold, K. E., Lynch, G., Huston, D., Wong, L., Jorn, L. & Olsen, C. W. (2014). Building institutional capacities and competencies for systemic learning analytics initiatives. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. Indianapolis, Indiana: ACM.
- Colvin, C., Rogers, T., Wade, A., Dawson, S., Gasevic, D., Buckingham Shum, S., Nelson, K., Lockyer, L., Kennedy, G., Corrin, L. & Fisher, J. (2015). Student retention and learning analytics: A

- snapshot of Australian practices and a framework for advancement. Sydney: Australian Government Office for Learning and Teaching.
- Deakin Crick, R., Knight, S. and Barr, S. (2017). Towards Analytics for Wholistic School Improvement: Hierarchical Process Modelling and Evidence Visualization. *Journal of Learning Analytics* 4(2), 160–188. <http://dx.doi.org/10.18608/jla.2017.42.13>
- Greller, W. & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Journal of Educational Technology & Society*, 15, 42-57.
- Knight, S., Dawson, S., Gasevic, D., Jovanovic, J., HersHKovitz, A. (2016). Learning analytics: Richer perspectives across stakeholders. *Journal of Learning Analytics*, 3(3), 1–4. <http://dx.doi.org/10.18608/jla.2016.33.1>
- Tsai, Y. S., and Gasevic D. (2017). Learning Analytics in Higher Education – Challenges and Policies: A review of eight learning analytics policies. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge*. ACM.
- Tsai, Y. S., Moreno-Marcos, P. M., Tammets, K., Kollom, K., & Gašević, D. (2018). SHEILA policy framework: informing institutional strategies and policy processes of learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 320-329).

APPENDICES

The screenshot displays a course interface with a top navigation bar containing 'View Progress' and '+ Module' buttons. The main content area is divided into two sections: 'Welcome' and 'Session 1: Data sets and where to find them'. Each section has a green checkmark, a plus icon, and a three-dot menu icon. The 'Welcome' section contains one item: 'Introduction to 'Finding Actionable Insights''. The 'Session 1' section contains eight items, each with a green checkmark and a three-dot menu icon.

Section	Item	Progress	Menu
Welcome	Introduction to 'Finding Actionable Insights'	✓	⋮
Session 1: Data sets and where to find them	Introduction to Session 1: Data sets and where to find them	✓	⋮
	Your Data - BI Program Dashboard	✓	⋮
	Your Data - Data Files	✓	⋮
	Your Data - Qualitative Survey Data	✓	⋮
	Your Data - Constellations	✓	⋮
	Your Data - CES Program Summary Dashboard 1	✓	⋮
	Your Data - CES Program Summary Dashboard 2	✓	⋮
	Your Data - Thoughts on Initial Data Exploration	✓	⋮

Figure 1. Welcome and session 1, how to work with course level-datasets. This is also an example of what users in the workshop would see – as academics would engage with a learning analytics workshop.

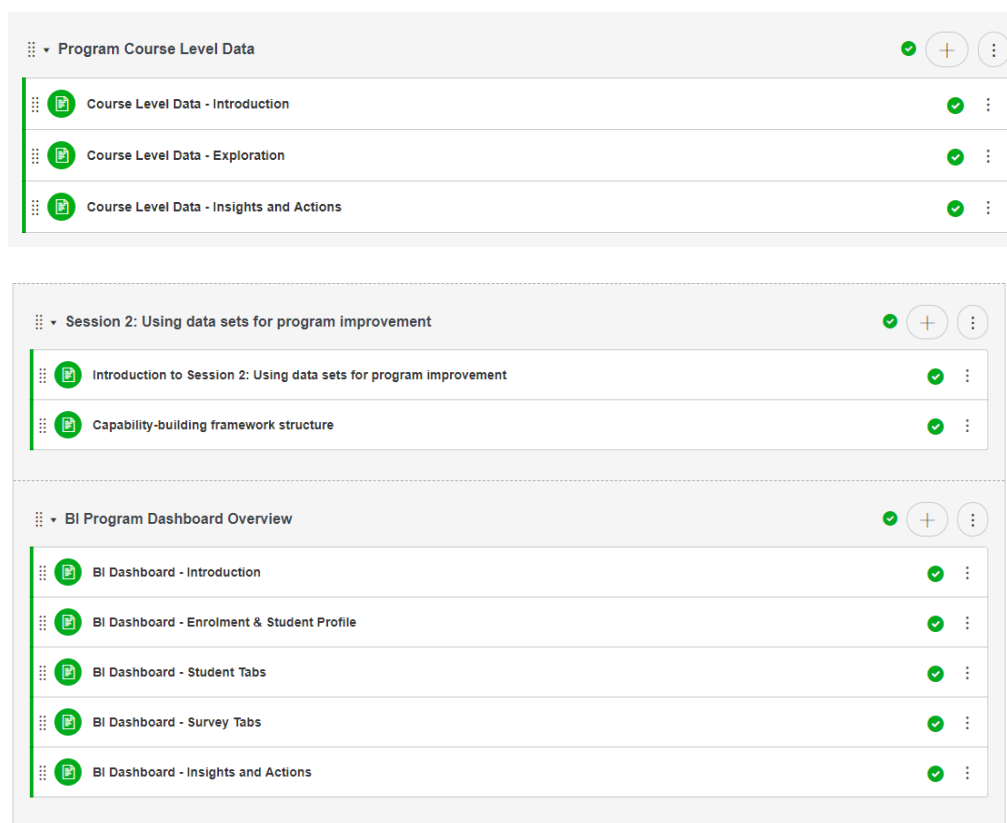


Figure 2. Course and Program-level datasets and introduction to dashboards and data. The proposed workshop for LAK already contains the modules to be tested – and these are agnostic to disciplines as they are designed for an entire university.

Module 6 - Final Insights and Actions

Now is time to bring everything together.

You should have a list of insights and actions at both the program and course level. Check over your list for interrelated themes. Are the same issues apparent from different sources. Review which insights are actionable and which aren't? Which actions will have the biggest impact? What data could help you? Do you know anyone who has experience with any of the action you wish to take? Look at the [Capability-building framework](#) and check for Modules that could help your program and teachers.

Once again you are decision makers. You have the best knowledge of and hence are best placed to make valuable insights about your program. The data is here to help and guide you.

Finally talk to other people you know who have completed this process. Work together on common insights and actions.

Activity: Summarise your findings

Purpose

Finalise your list of insights, questions and actions

Time

Unknown

Task

Step 1: Review all your notes

Step 2: Finalise your list of your most important/valuable insights. Double check that they are actionable?

Step 3: Finalise your list of questions and additional data sets you need. Check off any that you have already answered?

Step 4: Finalise your list of actions to take. Are they achievable? Will they have an impact?

Step 5: Start doing them.

Figure 3. Module addressing reflection on analysis and focusing on creating actions. This is for illustration purposes as the workshop will work on finessing this section.



Program Course Constellations User Guide

Purpose

The Program Course Constellations are a new way of visualising data about students enrolled in your program in a given term. They combine student data, such as enrolments, results and stage in program with survey data to give a snapshot of what courses your students were enrolled in, their performance and what they think.

If any student from your program was enrolled in a course in the semester, that course will appear in the program constellation, even if that course is from a different school or college. The important thing to remember is that this is a picture of what your students did; not what you expected them to have done.

Production

The Program Course Constellations is produced primarily from student course enrolments data. The process is as follows:

- 1) Get all student course enrolments for the term
- 2) Add additional student data
- 3) Calculate course level metrics for all students
- 4) Limit students to those enrolled in a program
- 5) Recalculate course level metrics for students in the program
- 6) Use course enrolments to determine the number of students taking courses concurrently
- 7) Plot all courses that students from the program were enrolled in within the single term.

Constellation Features

Line thickness

The thickness of the line between courses is determined by the number of students from the program enrolled in both courses.

Size

The size of the course is determined by the number of students from the program enrolled in that course.

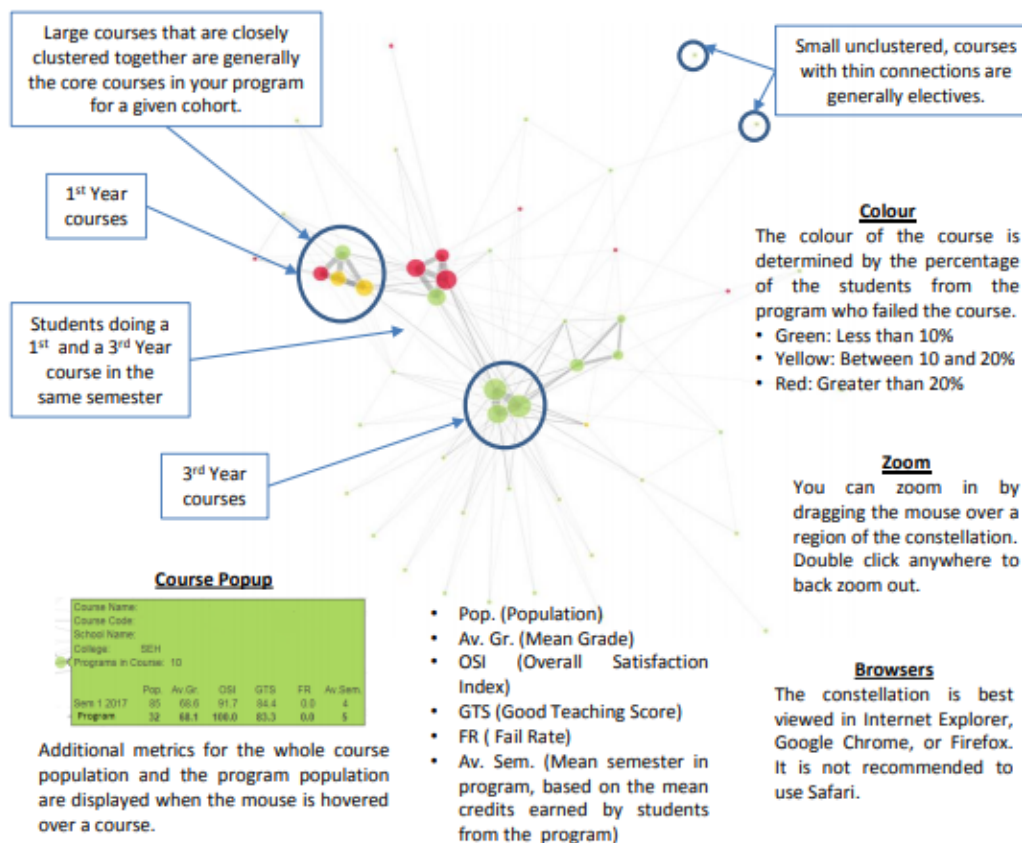


Figure 4. Example of instruction guides to use analytical tools. In this case, the *Constellation* is a tool to help visualize courses within programs. It is an interactive tool that can provide granularity on different dimensions such as failure rates and learning outcomes within courses and within programs.

Supporting Feedback Processes at Scale with OnTask A Hands-on Tutorial

Abelardo Pardo

University of South Australia
Abelardo.Pardo@unisa.edu.au

Dragan Gašević

Monash University
Dragan.gasevic@monash.edu.au

Shane Dawson

University of South Australia
Shane.Dawson@unisa.edu.au

George Siemens

University of South Australia
gsiemens@gmail.com

ABSTRACT: Although we have seen significant progress in the wealth of data captured in learning environments and the tools and techniques to deploy learning analytics methods, the true impact on the overall quality of a learning experience needs further study. We posit that the provision of personalized feedback for large student cohorts offers an ideal context to connect the variety of data sources currently emerging in institutions, the connection with a learning design, and the specifics to connect derived knowledge with tangible student support actions. The half-day session is targeted to researchers and practitioners interested on the use of data to adapt their design to provide personalized support to learners. Attendees will be offered the possibility of exploring this context using the open-source tool OnTask with a synthetically generated data set. Additionally, the session includes a discussion on how the proposed paradigm is being used in used in various educational institutions throughout the world.

Keywords: Feedback, instructional design, institutional adoption, student support

1 BACKGROUND

Although technology mediation in learning experiences offers a wealth of data about the events that take place while learners interact in an educational environment, data availability is only the first stage of a long journey that should conclude with tangible increases in either the understanding or the quality of a learning experience. Educational institutions have recognized the potential of the current data being captured through a variety of methods and processes. However, the evidence of impact of the overall paradigm still remains elusive (S. Dawson et al., 2018). Early use of data to tackle retention problems (e.g. Arnold, Hall, Street, Lafayette, & Pistilli, 2012; Colvin et al., 2016; Jayaprakash, Moody, Eitel, Regan, & Baron, 2014) have given way to a wider application of these methods to areas such as writing assignments (Gibson et al., 2017), epistemic network analysis (Shaffer, Collier, & Ruis, 2016), learning strategies (Fincham, Gašević, Jovanović, & Pardo, In Press), or the evolution of communities of enquiry (Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015).

But equally important to understanding learning experiences is the adoption of actions in this data-rich context. Wise, Vytasek, Hausknecht, and Zhao (2016) point out that the sensemaking stage has received more attention than the one about decision-making. One of the challenges in the latter resides on how to *enact change*. There has been numerous reviews on how dashboards are used as a

vehicle to enact this change (e.g. Schwendimann et al., 2017), however, their relation with student achievement remains unclear (Corrin & de Barba, 2015; Kahn & Pardo, 2016)

Feedback has been identified as one of the top ten aspects of learning to enhance student achievement in a learning experience (Hattie & Gan, 2011). In the context of higher and professional education Boud and Molloy (2013) explore two feedback models. The first one (Mark 1) is closer to the conventional perception of feedback as a process mediated by the instructor and typically requiring several stages for verification and student engagement. One of the challenges for this model is its scalability because most of its elements depend on the number of students. It is in this space where the proposed tutorial is situated. P. Dawson et al. (2018) summarized the key findings of a systematic literature review about the use of technology in feedback processes. The authors describe the approaches based on *digital text* as simple and convenient to the instructors, having the possibility of including specific comments, increased legibility, and more comfortable for the students. At the same time, they point to the challenges of providing detailed and personalized comments. Recent studies have shown the feasibility of personalized feedback deployed in learning environment with large number of students (Pardo, Jovanović, Dawson, Gašević, & Mirriahi, 2018). Additionally, Pardo et al. (In Press) have proposed and implementation OnTask¹, an architecture and set of open-source tools that support instructors to capture the connection between data and support actions and deploy them through a variety of methods.

We believe that the area of learning analytics could benefit from a more explicit connection between the data collected in learning environments and specific actions supporting all students throughout their experience. The combination of data, a learning design, and the provision of personalized feedback offer the ideal context to explore the requirements derived from the use of learning analytics methods at the institutional level how the relationship among its various elements

The goal of the workshop is to use OnTask and a synthetic data set to allow researchers and practitioners to explore how to articulate this connection between the elements in a data set, a learning design, and actions to support students during the experience. The attendees will also have the opportunity to discuss adoption scenarios of this paradigm at various levels in educational institutions (course, program, overall student experience).

2 ORGANISATIONAL DETAILS

The tutorial offers a BYOT, hands-on experience to use OnTask to manipulate a synthetic data set with information about: demographics information (type of enrolment, type of program), participation in online activities (videos, multiple-choice questions, discussion forum), and results of summative assessment (a midterm examination). The structure of the dataset is shown in the following Figure 1.

The session starts with an overview of the data set and a first task to upload and explore its components. The following task is to explore how to use the data to create personalized messages. Attendees will first identify within the learning context the possibility to support students through the provision of text messages and then articulate these messages depending on the information existing

¹ More information in ontasklearning.org

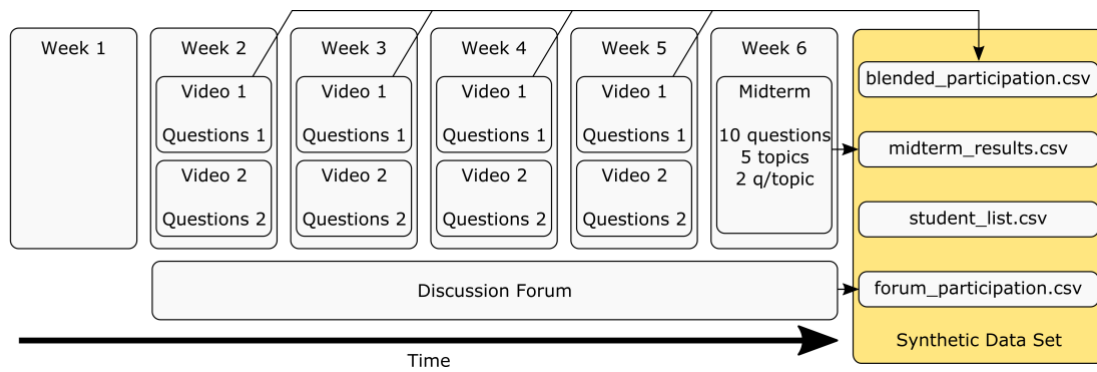


Figure 1 Structure of the synthetic data set used in the tutorial

in the data set. The third task shows how to combine self-reported data to refine further the personalized messages. Attendees will identify a set of questions to be sent to the students so that the provided answers complement the existing observational data and is used to further refine the personalized messages. The session concludes exploring how the approach has been used in several educational institutions and the challenges to adopt it. Attendees will be given an account to work in an OnTask web-based server where they can upload the given data set, create the personalized actions and preview the resulting text. Attendees must bring their own computer to the tutorial. Due to the combination of hands-on and discussion activities, attendance will be limited to 30 persons.

3 OBJECTIVES

The topic of the tutorial will appeal to researchers and practitioners in the area of learning analytics with some basic previous knowledge of data sources typically present in technology-mediated experiences. The topic may also appeal to educational data mining experts and academic designers interested on improving the connection between data and decision-making to increase student support. No programming language experience is required, but proficient management of data files (excel or CSV format) is desirable. The objectives of the hands-on tutorial are:

- Identify student support actions driven by data to deploy them in real time.
- Express the connection between data and actions in a formalism suitable to produce personalized support actions at scale.
- Discuss the deployment of these actions in educational institutions at various levels within the organization.

The attendees will have access to both the tool and the dataset beyond the tutorial session.

REFERENCES

- Arnold, K. E., Hall, Y., Street, S. G., Lafayette, W., & Pistilli, M. D. (2012). *Course Signals at Purdue: Using Learning Analytics to Increase Student Success*. Paper presented at the International Conference on Learning Analytics and Knowledge.
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698-712. doi:10.1080/02602938.2012.691462

- Colvin, C., Rogers, T., Wade, A., Dawson, S., Gašević, D., Buckingham Shum, S., . . . Fisher, J. (2016). Student retention and learning analytics: a snapshot of Australian practices and a framework for advancement. Caberra, ACT: Australian Government Office for Learning and Teaching.
- Corrin, L., & de Barba, P. (2015). *How Do Students Interpret Feedback Delivered via Dashboards?* Paper presented at the International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, USA.
- Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M., & Molloy, E. (2018). Technology and Feedback Design. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, Design, and Technology. An International compendium of theory, research, practice and Policy* (pp. 1-45). Switzerland: Springer International Publishing.
- Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., & Gasevic, D. (2018). *Rethinking learning analytics adoption through complexity leadership theory*. Paper presented at the International Conference on Learning Analytics and Knowledge - LAK '18, Sydney, Australia.
- Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (In Press). From Study Tactics to Learning Strategies: An Analytical Method for Extracting Interpretable Representations. *IEEE Transactions on Learning Technologies*.
- Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). *Reflective writing analytics for actionable feedback*. Paper presented at the International Conference on Learning Analytics and Knowledge, Vancouver, Canada.
- Hattie, J., & Gan, M. (2011). Instruction Based on Feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning Instruction* (pp. 249-271). New York: Routledge.
- Jayaprakash, S. M., Moody, E. W., Eitel, J. M., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students : An Open Source Analytics Initiative. *Journal of Learning Analytics, 1*, 6-47.
- Kahn, I., & Pardo, A. (2016). *Data2U: Scalable Real time Student Feedback in Active Learning Environments*. Paper presented at the International Conference on Learning Analytics and Knowledge, Edinburgh, UK.
- Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education, 27*, 74-89. doi:10.1016/j.iheduc.2015.06.002
- Pardo, A., Bartimote-Aufflick, K., Buckingham Shum, S., Dawson, S., Gao, J., Gašević, D., . . . Vigentini, L. (In Press). OnTask: Delivering Data-Informed Personalized Learning Support Actions. *Journal of Learning Analytics, In Press*.
- Pardo, A., Jovanović, J., Dawson, S., Gašević, D., & Mirriahi, N. (2018). Using Learning Analytics to Scale the Provision of Personalised Feedback. *British Journal of Educational Technology*. doi:10.1111/bjet.12592
- Schwendimann, B., Rodriguez-Triana, M., Vozniuk, A., Prieto, L., Boroujeni, M., Holzer, A., . . . Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1). doi:10.1109/tlt.2016.2599522
- Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. *Journal of Learning Analytics, (In press)*.
- Wise, A. F., Vytasek, J. M., Hausknecht, S., & Zhao, Y. (2016). Developing Learning Analytics Design Knowledge in the "Middle Space": The Student Tuning Model and Align Design Framework for Learning Analytics Use. *Online Learning Journal, 20*(2).

2nd Educational Data Mining in Computer Science Education (CSEDM) Workshop

David Azcona

Dublin City University
david.azcona2@mail.dcu.ie

Yancy Vance Paredes, Sharon Hsiao

Arizona State University
yvmparedes@asu.edu, sharon.hsiao@asu.edu

Thomas Price

North Carolina State University
twprice@ncsu.edu

ABSTRACT: The objective of this workshop is to facilitate a discussion among our research community around Artificial Intelligence (AI) in Computer Science Education. The workshop is meant to be an interdisciplinary event. Researchers, faculty and students are encouraged to share their data mining approaches, methodologies and experiences where AI is transforming the way students learn Computer Science (CS) skills. This year, we are introducing a Dataset Challenge, an activity which attempts to discuss some challenges when doing Data Mining in Computer Science Education.

Keywords: Computer Science Education, Learning Analytics, Educational Data Mining.

1 WORKSHOP BACKGROUND

Computer Science (CS) has become ubiquitous and is part of everything we do. Studying CS enables us to solve complex, real and challenging problems and make a positive impact in the world we live in.

Yet, the field of CS education is still facing a range of problems from inefficient teaching approaches to the lack of minority students in CS classes and the absence of skilled CS teachers. One of the solutions to these problems lies with effective technology-enhanced learning and teaching approaches, and especially those enhanced with AI-based functionality.

Providing education in Computer Science requires not only specific teaching techniques but also appropriate supporting tools. The number of AI-supported tools for primary, secondary and higher CS education is small and evidence about the integration of AI-supported tools in teaching and learning at various education levels is still rare.

In order to improve our current learning environments and address new challenges we ought to implement new AI techniques, collaborate and share student data footprints in CS. Data is the

driving force for innovation at this time and new approaches have been implemented in other fields of innovation and research like Computer Vision and Image Classification. New data-driven learning algorithms and machines to process them are now widely accessible such as Deep Neural Networks and Graphical Processing Units (GPUs).

We want to keep the momentum and support the Computer Science Education community by organizing a workshop focusing on how to mine the rich student digital footprint composed by behavioral logs, backgrounds, assessments and all sort of learning analytics. We aim to create a forum to bring together CS education researchers from adjacent fields (EDM, AIED, CSE) to identify the LAK challenges and issues in the domain-specific field, Computer Science Education.

This workshop will follow on Educational Data Mining in Computer Science Education (CSEDM 2018) and AI-supported Education for Computer Science (AIEDCS) 2013 and 2014 which had an increasing number of participants, submissions and presentations. These workshops and the conferences on this field such as the ACM Technical Symposium on Computer Science Education (SIGCSE) demonstrate the strength of a community that leverages AI techniques to build its innovations.

The workshop encourages contributions from the following topics of interest:

- Predictive student modelling for Computer Science courses and learning
- Adaptation and personalization within Computer Science learning environments
- Intelligent support for collaborative Computer Science problem solving
- Deep learning approaches to massive Computer Science datasets and courses
- Online learning environments for Computer Science: implementation, design and best practices
- Multimodal learning analytics and combination of student data sources in Computer Science Education
- Affective, emotional and motivational aspects related to Computer Science learning
- Explanatory predictive models in Computer Science Education
- Adaptive feedback, adaptive testing for Computer Science learning
- Discourse and dialogue research related to classroom, online, collaborative, or one-on-one learning of Computer Science
- Peer-review, peer-grading and peer-feedback in Computer Science
- Teaching approaches using AI tools
- Visual Learning Analytics and Dashboards for Computer Science
- Learning approaches using AI tools
- Network Analysis for programming learning environments
- Self-Regulated learning for Computer Science environments
- Writing and syntax analysis for programming design learning
- Natural Language Processing for Computer Science forums and discussions
- Analysis of programming design and trajectory paths
- Linked Data for Computer Science knowledge mapping
- Recommender systems and in-course recommendations for Computer Science learning

2 ORGANIZATIONAL DETAILS

This event is a full-day workshop. During the morning session, there will be a panel discussion among some invited speakers who are currently doing educational data mining. It will be done in an interactive manner. It will then be followed by the presentation of the research papers. During the afternoon session, there will be a discussion on some of the submissions for the dataset challenge. Participants will be given the chance to briefly discuss their techniques and findings. Finally, we will culminate with a breakout session and final discussion to wrap things up. Participation to the workshop is open to anyone interested. Participants are invited to submit their original and unpublished work for presentations and discussions. Submissions must be formatted using the Learning Analytics & Knowledge (LAK)'s Companion Proceedings Template. There will be three types of submissions, each having their own deadlines:

- Research Papers addressing any of the topics above. Accepted papers will be published in the LAK Companion Proceedings (max. 6 pages).
- Presentation Abstracts providing an overview of the presentation of a researcher. This will be in a conversational format. Accepted papers will NOT be published in the LAK Companion Proceedings (max. 2 pages). Presentations might include:
 - Descriptions of shareable Computer Science (CS) datasets
 - Descriptions of data mining / analytics approaches applied to specifically Computer Science datasets
 - Descriptions of tools or programming environments that use/produce data
 - Case studies of collaboration where reproducible practices were used to integrate or compose two or more data analysis tools from different teams
 - Descriptions of infrastructures that could collect and integrate data from multiple learning tools (e.g. forum posts, LMS activity and programming data)
 - Calls for Conversation (i.e. Birds of a Feather)
- Dataset Challenge entries attempting to solve a CS education problem. This must discuss in detail the methods used to make the predictions (max. 6 pages).

We expect around 15 participants and would need a room with a workstation, a microphone, and a projector for the presentations.

3 WORKSHOP OBJECTIVES

The objective of the workshop is to invite researchers who are interested in further exploring, contributing, collaborating and developing AI techniques for building educational tools for Computer Science to submit present their work for discussion. These outcomes will be disseminated through the official workshop website. We also intend to use the hashtag #CSEDMatLAK19.

We have accepted six research papers which will be presented during the workshop:

1. Creativity Inside and Outside Programing Learning
2. ProgSnap2: A Flexible Format for Programming Process Data

3. How does Performance in an Online Primer Predict Achievement in a Future Computer Science Course?
4. Analyzing Score and Time Trails in Data Collected by Tutors
5. A Comparison of Two Designs for Automated Programming Hints
6. Using Legacy Data to Build Bayesian Knowledge Tracing Model and Evaluate Its Effectiveness

Creativity Inside and Outside Programming Learning

Arnon HersHKovitz, Raquel Sitman, Rotem Israel-Fishelson

Tel Aviv University, Tel Aviv, Israel
{arnonhe,rocky,rotemisrael}@tauex.tau.ac.il

Andoni Eguíluz, Pablo Garaizar, Mariluz Guenaga

University of Deusto, Bilbao, Spain
{andoni.eguiluz,garaizar,mlguenaga}@deusto.es

ABSTRACT: Both creativity and computational thinking are considered as crucial skills for future citizens. We studied the associations between these two constructs among middle school students (N=57), considering two types of creativity: general creative thinking, and specific computational creativity. We find some similarities between creative thinking and computational creativity, and interesting associations between the latter and computational thinking acquisition.

Keywords: Creativity; computational thinking; game-based learning

1 INTRODUCTION

As Computational Thinking has been recognized as a key skill in today's digital era, it has been integrated into school curricula around the world, and many online platforms, especially game-based learning environments, now promote its development. Despite their popularity, research on these latter environments is meager; it is mainly qualitative and based on limited data.

Creativity is closely related to computer science and has a central role in fostering motivation and interest in this field of study. Studies have found a bi-directional connection between creativity and computer science. On the one hand, creativity may serve as a catalyst to solving algorithmic problems, creating computational artifacts, and developing new knowledge. As was previously shown, scores from a standardized creativity test (the one that we used in the current study) predicted creativity in problem solving in computer programming, among undergraduate students (Liu & Lu, 2002). On the other hand, practicing the skills required for computer science—e.g., observation, imagination, visualization, abstraction, and creation and identification of patterns—can support the development of creative thinking (Clements & Gullo, 1984; Seo & Kim, 2016; Yadav & Cooper, 2017). Indeed, engaging with rich digital environments was shown to promote creativity (Lau & Lee, 2015; Psotka, 2013). It is not surprising, then, that software engineering—in which Computational Thinking inherently, conveniently resides—has been identified as a field that can benefit from creativity (Díaz, Aedo, & Cubas, 2014; Zhou, 2016).

Research on creativity in Computational Thinking (or programming) usually employs one of two possible types of exploration. The first type focuses on creativity within the scope of Computational Thinking, that is, on creative artifacts, which are products of the Computational Thinking learning process (usually programs written by learners). Yadav and Cooper say platforms like Alice or Scratch

provide opportunities for students "to extend their creative expression to solve problems, create computational artifacts" (Yadav & Cooper, 2017, p. 31). Such studies argue that creativity enabled by programming environments may act as a driving force for learning (Knobelsdorf & Romeike, 2008; Romeike, 2007; Roque, Rusk, & Resnick, 2016). In this category, we can also include studies looking for associations between creativity and other variables that refer to constructs out of the learning environment. For example, Doleck et al. (2017) examined associations between creativity as an inherent component of computational thinking and academic achievement. It is important to note that some studies have used an automatic method for detecting creativity in programming (Bennett, Koh, & Repenning, 2010; Manske & Hoppe, 2014). We took a similar approach.

The second type of study explores the relationship between measures of creativity outside the scope of Computational Thinking and variables associated with the acquisition of Computational Thinking. The main questions raised are whether creativity supports the acquisition of Computational Thinking (Pérez Poch, Olmedo Torre, Sánchez Carracedo, Salán Ballesteros, & López Álvarez, 2016), and whether teaching Computational Thinking can improve creativity (Chao, Liu, & Chen, 2014; Seo & Kim, 2016). For example, Knochel and Patton (2015) argue that presenting creativity in programming to design students promotes better creative design.

Therefore, associations between Computational Thinking and creativity—either within or outside the scope of Computational Thinking —have been recently studied, and preliminary evidence suggest some interesting links between these constructs. Still, a gap exist, as only little has been studied regarding the relationship between the two types of creativity. Also, most of the relevant studies have only focused on aggregated measures of creativity. We aim at bridging this gap by operationalizing a "continuous" (rather than aggregated) measure of Computational Thinking - related creativity, and to test for its associations with a standard, aggregated, non- Computational Thinking-related measure of creativity.

Note that we refer to creativity both inside and outside the learning environment. Inside the learning environment, we refer to Computational Creativity, specifically, the extent to which a correct solution is original (compared with all the solutions submitted by the whole research population). Outside the learning environment, we refer to Creative Thinking, which will be measured by classical creativity definitions. Computational Thinking is defined as the efforts and success demonstrated by the students in the game-based learning environment.

2 METHODOLOGY

To explore the role of creativities—both inside and outside the learning process—in the acquisition of computational thinking Computational Thinking, we formulated the following research questions: (1) What are the associations between Creative Thinking and the acquisition of Computational Thinking?; (2) What are the associations between Computational Creativity and the acquisition of Computational Thinking?; (3) What are the associations between Creative Thinking and Computational Creativity?

2.1 Population and Research Process

Data was collected in April 2017 from a population of N=131 secondary school students. The students used Kodetu, an online, block-based game for teaching basic programming skills, for about 50 minutes. Each of Kodetu's 15 levels presents the user with a maze in which an astronaut should get to a marked destination. Progressing to the next level is possible upon successfully completing the current level. We included only participants with no previous experience in programming or in using Kodetu (based on their self-reports), N=57.

Students also filled-up the Torrance Test for Creative Thinking (TTCT) – Figural Test (Torrance, 1974). In this pen-and-paper test, each participant was presented with a sheet with 12 identical, empty circles, and was asked to make as many drawings as possible using the circles as part of the drawings. An eligible drawing used the circle as part of the drawing. See Figure 1 for examples.

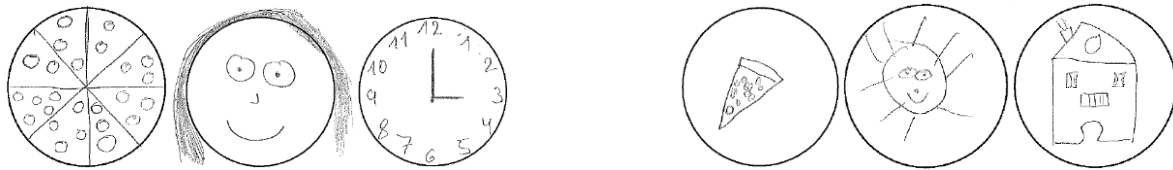


Figure 1. Examples of 3 eligible (left) and 3 non-eligible (right) drawings from TTCT – Figural Test

2.2 Dataset and Preprocessing

The full log file included 101,728 rows, each representing an action taken by a user, including its timestamp, the level in which it was taken [1-15], its result [Success, Failure, Timeout, Error, Unset], and the written code associated with this action. For our analysis of creative solutions, we only referred to correct solutions, which left us with 1332 rows from Levels 1-14 (no correct solution was logged for Level 15). As Levels 13 and 14 were only completed by a few students, we omitted them.

2.3 Research Variables

These variables are calculated at the student level.

Creative Thinking. Based on TTCT scoring guidelines, we defined the following four variables for each participant: *Fluency* (M=6.12, SD=3.85) - number of eligible drawings; *Flexibility* (M=2.75, SD=2.08) - number of drawing categories; *Originality* (M=0.75, SD=0.09) - the average frequency of the drawing categories, across all drawings, inversed (that is, the higher this value is, the more original the student is); finally, *Credibility Index* (M=0.14, SD=0.82) - average of standardized fluency, flexibility, and originality.

Computational Creativity. Calculated separately for each level, that is, each student has a set of 14 Computational Creativity measures; for each level, calculation is done by taking the complementary to 100% of the frequency of the student's correct solution among all the correct solutions for that level (averaged across the student's multiple correct solutions, if relevant).

Computational Thinking. We focused on two variables to measure the acquisition of Computational Thinking: *Max Level* (M=10.8, SD=1.97) - maximum level reached (not necessarily completed successfully); *Solution Attempts* (M=5.6, SD=3.4) – average number of attempts to solve each of Levels 1-12 (also calculated separately for each level).

3 FINDINGS

3.1 Computational Creativity along the Game

We ran 55 pair-wise between-level correlations, correcting for multiple comparisons using the post-hoc False Discovery Rate (FDR) method. We found **significant, positive, moderate to strong correlations between the pairs of almost all consecutive levels**; exceptions were the pairs of Levels 2-3 and 10-11; we also found significant, positive, moderate to strong relations between the non-consecutive pairs of Levels 3-5, 4-6, 4-7, 5-7. Significant p values ranged between 0.32 and 0.66.

3.2 Creative Thinking and Acquisition of Computational Thinking

We tested for correlations between the Creative Thinking variables and both Max Level and Average Solution Attempts, and found no significant correlations. There was, however, a marginally significant positive correlation between originality and Max Level, with $p=0.27$, at $p=0.052$ (N=51). Findings are presented in Table 1.

Table 1. Correlations between Creativity Thinking (columns) and Computational Thinking (rows).

	Fluency (N=56)	Flexibility (N=56)	Originality (N=51)	Creativity Index (N=51)
Max Level	$\rho=-0.0$ $p=0.95$	$\rho=0.22$ $p=0.11$	$\rho=0.27$ $p=0.052$	$\rho=0.10$ $p=0.50$
Average Solutions Attempts	$r=-0.07$ $p=0.62$	$r=-0.03$ $p=0.84$	$r=-0.06$ $p=0.68$	$r=0.03$ $p=0.84$

3.3 Computational Creativity and Acquisition of Computational Thinking

We found no associations between Computational Creativity and Level Solution Attempts (with one data point for each variable at each level except Level 8), with $p=0.18$, at $p=0.60$. That is, **overall, Computational Creativity was not linearly associated with level difficulty**. Testing correlations of these two variables in each level separately, we found only one case with a significant correlation: in Level 2, Computational Creativity was significantly negatively correlated with Level Solution Attempts, with $\rho=-0.28$, at $p<0.05$ (N=57). Therefore, **the more original a participant's solution was in level 2, the fewer attempts she or he needed to complete this level**.

Taking a more aggregated view of the data, we tested for correlations between Computational Creativity in each level and both Max Level and Average Solution Attempts. In this case, we found a significant negative correlation between level 2 Computational Creativity and Max Level, with $\rho=-0.37$, at $p<0.01$ (N=57). That is, **providing an original solution in an early stage of the game was negatively associated with progressing farther in the game**.

3.4 Creative Thinking and Computational Creativity

In the next step, we tested for associations between the creativity-related measures outside and inside the learning environment. As we were not assuming dependence within the level-based Originality measures, we correlated each of the Creative Thinking measures with each of the level-based Originality variables. **In four cases – levels 4, 9, 11, and 12 – we found significant correlations between the two types of creativity measures**, with Spearman's ρ taking values between 0.30-0.55. In these levels, Creative Thinking's Fluency, Flexibility, and Creativity Index were positively correlated with the level-based Originality.

4 DISCUSSION

Overall, we found no correlations between Computational Creativity and task difficulty. Other recent studies argue for a direct, positive relationship between difficulty and creativity (Chae & Seo, 2015; Espedido & Searle, 2018), however with a global measurement of difficulty. In our study, the same task may have been difficult for one student and easy for another, therefore we suggest **no correlation between Computational Creativity and acquisition of Computational Thinking**. This may be explained by the tension between knowledge and time constraints.

We also find some striking associations between the two measures of creativity. In four out of 11 levels, level-based Computational Creativity was positively associated with two dimensions of Creative Thinking—Fluency and Flexibility—and with the overall Creative Index. This finding supports the hierarchical model of creativity, which integrates both domain-general and domain-specific types of creativity (Baer, 2010); also, it resonates previous findings of associations between TTCT score and creativity in problem-solving in programming (Liu & Lu, 2002). Surprisingly, Computational Creativity—which has to do with Originality—is mostly associated with the non-originality dimensions of Creative Thinking. That is, Originality (and Fluency and Flexibility) may have different meanings in different contexts.

This study contributes to the growing body of literature on creativity, and to the scarce knowledge about creativity in programming. Taking a log-based approach allows us to study this phenomenon on a larger scale, and we plan to do so. Many learning environments for computational thinking seek efficiency and penalize original solutions. We seek the formula for promoting both. Overall, this research raises many questions that we hope will ignite many more studies in the field.

REFERENCES

- Baer, J. (2010). Is creativity domain specific? In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge Handbook of Creativity* (pp. 321–341). New York, NY: Cambridge University Press.
- Bennett, V. E., Koh, K. H., & Repenning, A. (2010). Computing creativity: Divergence in computational thinking. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education* (pp. 359–364). Denver, CO. <https://doi.org/10.1145/2445196.2445302>
- Chae, S. W., & Seo, Y. W. (2015). Task difficulty and team diversity on team creativity: Multi-agent simulation approach. *Computers in Human Behavior*, 42, 83–92. <https://doi.org/10.1016/J.CHB.2014.03.032>
- Chao, J.-Y., Liu, C.-H., & Chen, J.-Y. (2014). The influence of courses integrating Atayal culture and LEGO Dacta on the programming ability and creativity of Aboriginal children. *Global Journal of Computers & Technology*, 1(2), 34–43. Retrieved from

- <http://www.gpcpublishing.org/index.php/gjct/article/view/10>
- Clements, D. H., & Gullo, D. (1984). Effects of computer programming on young children's cognition. *Journal of Educational Psychology*, 76(6), 1051–1058. <https://doi.org/10.1037/0022-0663.76.6.1051>
- Díaz, P., Aedo, I., & Cubas, J. (2014). CoDICE: Balancing software engineering and creativity in the co-design of digital encounters with cultural heritage. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14* (pp. 253–256). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2598153.2598190>
- Doleck, T., Bazelais, P., Lemay, D. J., Saxena, A., & Basnet, R. B. (2017). Algorithmic thinking, cooperativity, creativity, critical thinking, and problem solving: exploring the relationship between computational thinking skills and academic performance. *Journal of Computers in Education*, 4(4), 355–369. <https://doi.org/10.1007/s40692-017-0090-9>
- Espedido, A., & Searle, B. J. (2018). Goal difficulty and creative performance: The mediating role of stress appraisal. *Human Performance*, 3(13), 179–196. <https://doi.org/10.1080/08959285.2018.1499024>
- Knobelsdorf, M., & Romeike, R. (2008). Creativity as a Pathway to Computer Science. *ACM SIGCSE Bulletin*, 40(3), 286. <https://doi.org/10.1145/1597849.1384347>
- Knochel, A. D., & Patton, R. M. (2015). If art education then critical digital making: Computational thinking and creative code. *Studies in Art Education*, 57(1), 21–38.
- Lau, K. W., & Lee, P. Y. (2015). The use of virtual reality for creating unusual environmental stimulation to motivate students to explore creative ideas. *Interactive Learning Environments*, 23(1), 3–18. <https://doi.org/10.1080/10494820.2012.745426>
- Liu, M.-C., & Lu, H.-F. (2002). A study on the creative problem-solving process in computer programming. In *Proceeding of the International Conference on Engineering Education*. Manchester, UK. Retrieved from <https://pdfs.semanticscholar.org/057e/f657236382b17b7b3e9865178709def3296b.pdf>
- Manske, S., & Hoppe, H. U. (2014). Automated Indicators to Assess the Creativity of Solutions to Programming Exercises. In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 497–501). IEEE. <https://doi.org/10.1109/ICALT.2014.147>
- Pérez Poch, A., Olmedo Torre, N., Sánchez Carracedo, F., Salán Ballesteros, M. N., & López Álvarez, D. (2016). On the influence of creativity in basic programming learning at a first-year Engineering course. *International Journal of Engineering Education*, 32(5(B)), 2302–2309. Retrieved from <https://upcommons.upc.edu/handle/2117/97382>
- Psotka, J. (2013). Modeling, simulations and education. *Interactive Learning Environments*, 21(4), 319–320. <https://doi.org/10.1080/10494820.2013.808880>
- Romeike, R. (2007). Applying creativity in CS high school education – Criteria , teaching example and evaluation. In *Proceedings of the Seventh Baltic Sea Conference on Computing Education Research-Volume* (pp. 87–96). Australian Computer Society, Inc.
- Roque, R., Rusk, N., & Resnick, M. (2016). Supporting Diverse and Creative Collaboration in the Scratch Online Community. In *Mass Collaboration and Education* (pp. 241–256). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-13536-6_12
- Seo, Y.-H., & Kim, J.-H. (2016). Analyzing the effects of coding education through pair programming for the computational thinking and creativity of elementary school students. *Indian Journal of Science and Technology*, 9(46), 1–5. <https://doi.org/10.17485/ijst/2016/v9i46/107837>
- Torrance, E. P. (1974). *Torrance tests of creative thinking*. Bensenville, IL: Scholastic Testing Service.
- Yadav, A., & Cooper, S. (2017). Fostering creativity through computing. *Communications of the ACM*, 60(2), 31–33. <https://doi.org/10.1145/3029595>
- Zhou, C. (2016). Developing creativity as a scientific literacy in software engineering education towards sustainability. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 2257–2261). IEEE. <https://doi.org/10.1109/FSKD.2016.7603533>

ProgSnap2: A Flexible Format for Programming Process Data

Thomas W. Price¹, David Hovemeyer², Kelly Rivers³, Austin Cory Bart⁴,

Andrew Petersen⁵, Brett A. Becker⁶, Jason Lefever²

¹North Carolina State University, ²York College of Pennsylvania, ³Carnegie Mellon University,

⁴University of Delaware, ⁵University of Toronto, ⁶University College Dublin

twprice@ncsu.edu, dhovemey@ycp.edu, krivers@andrew.cmu.edu, acbart@udel.edu,

petersen@cs.toronto.edu, brett.becker@ucd.ie, jlefever@ycp.edu

ABSTRACT: In this paper, we introduce ProgSnap2, a standardized format for logging programming process data. The goal of this common format is to encourage collaboration among researchers by helping them to share data, analysis code, and data-driven tools to support students. We first highlight possible use cases for ProgSnap2 and give a high-level overview of the format. We then share two case studies of our experience using the format and outline goals for the future of ProgSnap2, including a call for collaboration with interested researchers.

Keywords: programming process data, data standards, data sharing, learning analytics

1 INTRODUCTION

Analysis of programming process data, logged as students complete programming tasks, has furthered the field of computing education research in many ways, including identifying common programming errors (Brown & Altadmri, 2014a) and detecting plagiarism (Hellas et al., 2017). However, there are few common standards for how such data should be collected, represented, or shared, making it more difficult for researchers to collaborate, replicate findings, and share tools. Initiatives such as the PSLC Datashop (Koedinger et al., 2010) provide a common data format and tools to store, analyze, and share *generic* educational data. However, programming datasets have a number of distinct, domain-specific features, which make it difficult to use generic formats. Programming datasets may track entire projects with multiple files, and interpreting them often requires specific metadata, such as the version of the IDE or compiler. Programming data collection tools, such as BlackBox (Brown et al., 2014b) and CloudCoder (Papancea et al., 2013), have addressed this problem by defining their own data formats, but these are system-specific and not widely adopted.

In this paper, we present ProgSnap2: a standardized format for logging programming process data, which we have developed and are currently refining. ProgSnap2 builds on the original Progsnap format (Hovemeyer et al., 2017)¹ by representing a richer set of event data types and using a “flat” representation more suitable for direct analysis by statistics software. The goal of ProgSnap2 is to support researchers in sharing and analyzing programming process data. The format was designed

¹ See the full specification for the original Progsnap at: <http://cloudcoderdotorg.github.io/progsnap-spec/>

to prioritize the needs of both the *data producer* and the *data consumer*. For the *data producer*, our goal is to make exporting data straightforward, with a default structure to encourage best practices (e.g. what to log and how), a small set of required elements, and extensibility to support a variety of datasets. For the *data consumer*, our goal is to make importing and analyzing data straightforward, while making explicit how the data were logged, and any caveats or oddities that might impact analysis.

We see three primary use cases for the ProgSnap2 format. **1) Sharing Data:** There is a high cost to sharing unstandardized data. Both parties must invest time for the consumer to understand and parse the new format. A common format lowers these barriers, while improving the quality of new and existing logging systems by defining a standard set of events and attributes to log. Efforts to standardize the format and storage of learning data in other domains have led to datasets and research efforts that spanned multiple researchers and institutions (Koedinger et al., 2010). **2) Sharing Analysis Code:** A common format also allows researchers to write analysis code that can be shared and reused on new datasets that have the same format. This enables researchers to collaborate, even when sharing data is not possible (e.g. for privacy reasons). Publishing analysis code can also increase the replicability of computing education research and encourage the development of shared analysis libraries. For example, many researchers use the Error Quotient (Jadud, 2006) to quantify learners' compilation behavior. A shared implementation of the Error Quotient, capable of operating on any dataset in the common format, would save effort and ensure a consistent definition. **3) Sharing Tools:** A number of data-driven tools have been developed to support computing classrooms, such as student models (Yudelson et al., 2014) and on-demand hints (Rivers and Koedinger 2017; Price et al, 2017). A common input data format would allow these tools to be more easily shared, reused, and composed together. This raises the possibility of publishing these tools as *services* that any researcher can utilize, for example allowing any programming environment to employ an adaptive student model by sending its data to the appropriate service.

2 PROGSNAP2

A ProgSnap2² dataset consists of logs and relevant data that capture how users interacted with a programming or learning environment. A dataset includes a *main event table*, a *metadata table* and optional *link tables* to reference outside resources, all represented as CSV files. A dataset also contains a *code repository* containing sequential snapshots of students' code and optional auxiliary *resources* (e.g. assignment descriptions). We chose to define most elements of the dataset as directly parsable CSV files, rather than using a database, with the goal of making analysis as straightforward as possible.

2.1 Main Event Table

The central component of a dataset is the *main event table*, which represents a collection of events that took place in the programming environment. These events can represent both fine-grained interactions, such as individual keystrokes, and high-level actions, such as entire problem attempts,

² The full specification for ProgSnap2 is available at: <http://bit.ly/ProgSnap2>

depending on the granularity of the logging system. Each row in the table represents one event, and each column represents an event property. ProgSnap2 defines a small set of mandatory columns:

- **EventType**: an enumeration value indicating the type of event; examples include “Session.Start”, “File.Edit”, “Compile”, “Compile.Error”, “Submit”, and “Run.Program”
- **EventID**: the unique ID of the event
- **Order**: the chronological ordering of the event compared to others (may be approximate)
- **SubjectID**: the ID of the human subject (or group) associated with the event
- **ToolInstances**: a string indicating the names and versions of tools associated with the event
- **CodeStateID**: the ID of a snapshot of the source code and resources when the event occurred

ProgSnap2 also defines a variety of optional columns with standard names. Some columns may not apply to all datasets (e.g. CourseID) and can be omitted. Others apply to a specific subset of events, and can be included for only these events (e.g. CompileMessageType is only appropriate for “Compile.Error” events), creating a *sparse* table. Data producers are encouraged to include as many optional columns and as much detail as possible. They can also define new columns when needed. Examples of optional columns include:

- **ParentEventID**: the EventID of a “parent” event, to represent causal relationships; for example, a “Compile.Error” event would typically have a “Compile” event as its parent
- **TermID, CourseID, CourseSectionID, AssignmentID, ProblemID**: these provide contextual information for the associated event, which may be found in a “Link Table” (described below)
- **EditType, EditTrigger, CodeStateSection**: these respectively describe how code was edited (e.g. typing, paste, undo), the reason it was recorded, and where the edit took place
- **ProgramInput, ProgramOutput, CompileMessageType, CompileMessageData**: these record relevant information about how the code was compiled and run
- **ExperimentalCondition, InterventionType, InterventionMessage**: these record data about experimental conditions and interventions used in research studies

2.2 Metadata, Link Tables and Resources

The **Dataset Metadata** is a mandatory CSV file specifying the global properties of the dataset as key/value pairs. Currently, only a few global properties are defined, including the ProgSnap2 version number, whether event ordering is known, and which code state representation is used. **Link Tables** are *optional* files used to associate contextual ID values (or combinations of IDs) with a *resource* (defined below) providing more information. For example, a link table could associate a TermID/CourseID pair with the URL of a course website for that course and term. Another optional file, a **Resource**, is an arbitrary data blob, identified by a URL in a Link Table, which can be either external (accessed via the internet) or internal (local to the dataset). As with Link Tables, the inclusion of Resources is *optional* but encouraged. Where possible, we also encourage data producers to use *internal* resources (e.g. saving a static version of the course website in the example above) to ensure they are not lost or changed.

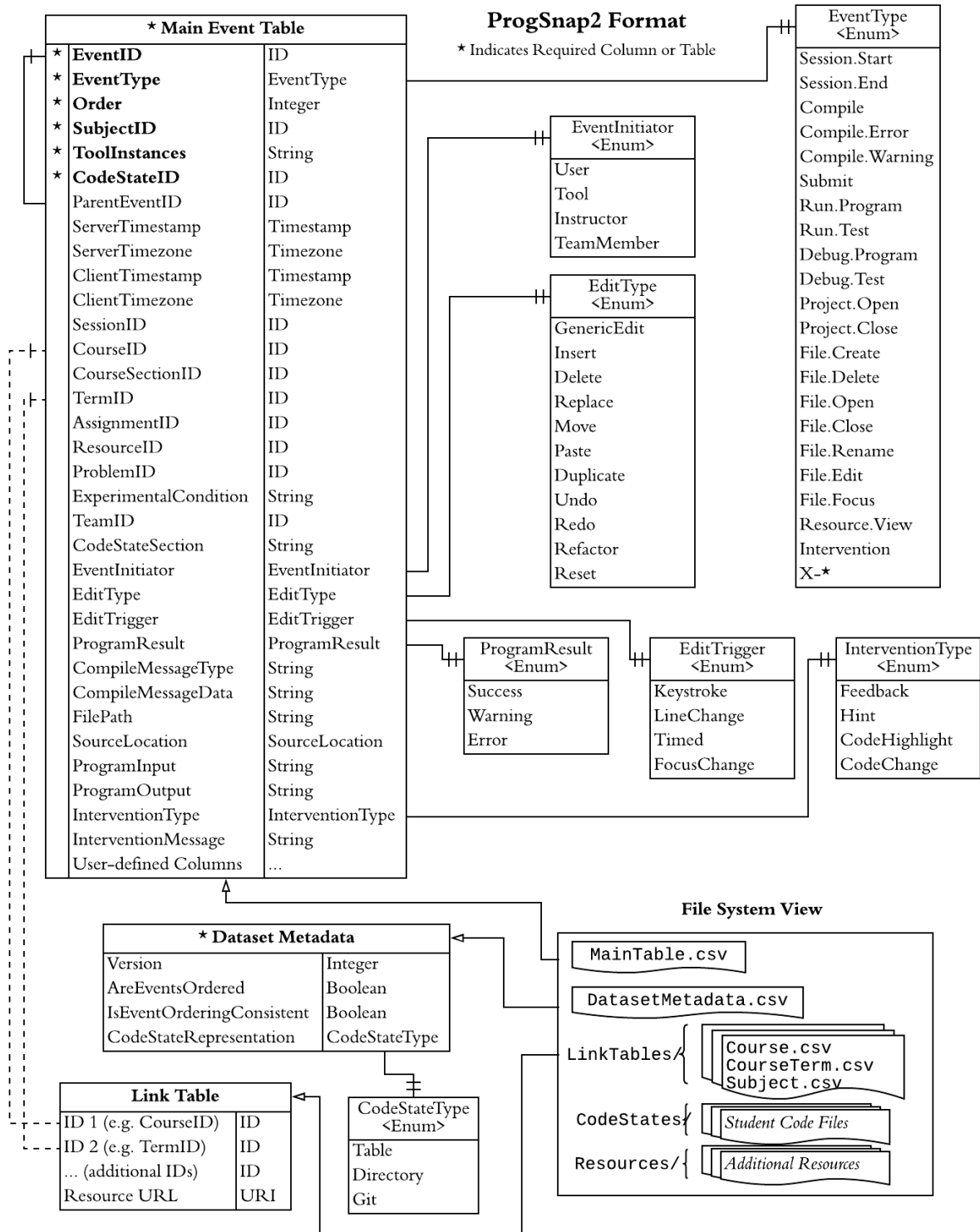


Figure 1: A diagram of the ProgSnap2 Format. Lines connect Enum types to their possible values, files to their respective tables, and IDs to their definitions in the Main Event Table.

2.3 Code State Representations

ProgSnap2 is intended to represent data from a variety of programming activities, from single-function exercises to complex final projects to block-based programs. To capture this diverse data, ProgSnap2 supports three source code representations: *Git*, *Directory*, and *Table*. This choice allows data producers to use the most appropriate representation, while constraining that choice to formats which are easily processed. Each format maps a CodeStateID value to a *code state*, which is simply a collection of one or more files with an optional directory structure. In the *Git* format, code states are represented as commits in a Git³ repository stored within the dataset. This format is appropriate for datasets where code states may consist of a relatively large number of files. In the *Directory* format, each CodeStateID maps to the name of a directory stored within the dataset which contains a collection of all files that are part of the code state. This format is appropriate for datasets where code states contain a small number of files. In the *Table* format, a dedicated CSV file maps CodeStateID values to text strings. This format is only appropriate for datasets where each code state consists of a single text file, and where the amount of data per code state is small.

3 CASE STUDIES

To explore and evaluate the standard, we implemented ProgSnap2 data exporters for two open source autograding systems, Virtual Programming Lab (VPL)⁴ and CloudCoder (Papancea et al., 2013).

Virtual Programming Lab (VPL): As learners make submissions and receive feedback, the VPL maintains a downloadable log stored as a zip file of directories, with each directory representing one student. These directories contain a timestamped series of paired folders representing student code submissions and their associated compilation information. Our tool⁵ consumes these logs and produces ProgSnap2 compliant archives. During conversion, each submission is decomposed into a sequence of events (“Submission”, “Compile”, “Compile.Error”, etc.). Each event is assigned a numerically ascending, unique Event ID, and necessary fields are assigned, such as the Server Timestamp, Event Type, and event-specific data like the code for a “Submission” event or the compiler’s output during a “Compile.Error” event.

We faced a few challenges during development, such as mapping VPL’s data to ProgSnap2 events. For example, a number of event types relate to compilation. Given that Python is not truly “compiled”, should we use a “Compile” event or a “Run.Program” event? When students run their code and receive autograder feedback, would a “Run.Test” event be appropriate, since autograding is more than just unit testing? If there is an error, VPL will still offer feedback to the student. Is this an “Intervention” or just the “ProgramResult”? When a Grade is assigned, is that also an “Intervention,” or does the standard need a new Event Type? Most of these challenges were easily resolved, though some led to ongoing conversations that may be settled as we develop other conversion tools.

³Standard libraries for extracting a commit from a Git repository (git-scm.com) can be found at libgit2.org

⁴ <http://vpl.dis.ulpgc.es/>

⁵ Source code for the tool is available at: <https://github.com/CSSPLICE/progsnap2>

CloudCoder: Exporting CloudCoder data to the ProgSnap2 format was fairly straightforward and mostly involved mapping CloudCoder's internal event representation to that of ProgSnap2. One challenge we encountered is that a single CloudCoder event can yield multiple ProgSnap2 events in some cases. To address this, the CloudCoder event IDs are multiplied by a constant to create a gap in the namespace where multiple derived events can be situated without conflict. We also had difficulty defining Session.Start and Session.End events, as CloudCoder does not directly record sessions. We considered defining them based on how much time elapsed between recorded CloudCoder events, but eventually decided to omit them. We felt it would be more appropriate for the data consumer to develop his or her own heuristics to reconstruct sessions during analysis.

4 FUTURE WORK AND CALL FOR COLLABORATION

We are currently working to refine ProgSnap2 by exporting datasets from additional programming environments, including PCRS (Zingaro et al., 2013) and iSnap (Price et al., 2017). However, our primary goal for the format is to facilitate collaboration through the sharing of data, analysis code, and data-driven tools. We plan to evaluate the utility of ProgSnap2 through these efforts, and we invite researchers interested in sharing programming data for collaboration to contact the authors.

5 REFERENCES

- Brown, N. C. C., & Altadmri, A. (2014a). Investigating Novice Programming Mistakes: Educator Beliefs vs Student Data. In Proceedings of the Tenth International Computing Education Research Conference (pp. 43–50). <https://doi.org/10.1145/2632320.2632343>
- Brown, N. C. C., Kölling, M., McCall, D., & Utting, I. (2014b). Blackbox: A Large Scale Repository of Novice Programmers' Activity. In Proceedings of the ACM Technical Symposium on Computer Science Education (pp. 223–228). <https://doi.org/10.1145/2538862.2538924>
- Hellas, A., Leinonen, J., & Ihantola, P. (2017). Plagiarism in Take-home Exams : Help-seeking , Collaboration , and Systematic Cheating. In Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education (pp. 238–243). <https://doi.org/10.1145/3059009.3059065>
- Hovemeyer, D., Hellas, A., Petersen, A., & Spacco, J. (2017). Progsnap: Sharing Programming Snapshots for Research. In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (p. 709).
- Jadud, M. C. (2006). Methods and tools for exploring novice compilation behaviour. In Proceedings of the Third International Workshop on Computing Education Research (pp. 73–84). <https://doi.org/10.1145/1151588.1151600>
- Koedinger, K. R., Baker, R. S. J., Cunningham, K., & Skogsholm, A. (2010). A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Sj. Baker (Eds.), Handbook of Educational Data Mining (pp. 43–55). CRC Press. <https://doi.org/doi:10.1201/b10274-6>
- Papancea, A., Spacco, J., & Hovemeyer, D. (2013). An Open Platform for Managing Short Programming Exercises. In Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research (pp. 47–52). New York, NY, USA: ACM. <https://doi.org/10.1145/2493394.2493401>

- Price, T. W., Dong, Y., & Lipovac, D. (2017). iSnap: Towards Intelligent Tutoring in Novice Programming Environments. In Proceedings of the ACM Technical Symposium on Computer Science Education.
- Price, T. W., Zhi, R., & Barnes, T. (2017). Evaluation of a Data-driven Feedback Algorithm for Open-ended Programming. In Proceedings of the International Conference on Educational Data Mining.
- Rivers, K., & Koedinger, K. R. (2017). Data-Driven Hint Generation in Vast Solution Spaces: a Self-Improving Python Programming Tutor. *International Journal of Artificial Intelligence in Education*, 27(1), 37–64. Retrieved from <http://link.springer.com/10.1007/s40593-015-0070-z>
- Yudelson, M., Hosseini, R., Vihavainen, A., & Brusilovsky, P. (2014). Investigating Automated Student Modeling in a Java MOOC. In Proceedings of the International Conference on Educational Data Mining (pp. 261–264).
- Zingaro, D., Cherenkova, Y., Karpova, O., & Petersen, A. (2013). Facilitating Code-writing in PI Classes. In Proceeding of the 44th ACM Technical Symposium on Computer Science Education (pp. 585–590). <https://doi.org/10.1145/2445196.2445369>

How does Performance in an Online Primer Predict Achievement in a Future Computer Science Course?

Soniya Gadgil

Eberly Center for Teaching Excellence and Educational Innovation
soniyag@andrew.cmu.edu

Steven Moore¹, John Stamper²

Carnegie Mellon University, HCII
StevenJamesMoore@gmail.com¹
jstamper@cs.cmu.edu²

ABSTRACT: We describe a feature engineering approach to predict future course performance based on students' interactions with an online math primer. To help incoming computer science freshman students gain competency on core discrete math concepts, we developed a primer course deployed in an interactive learning environment. The primer covered three foundational topics — logic, sets, and functions. Students completed this primer in the summer prior to their first semester as computer science undergraduates. We used random forest modeling and linear regression to understand which features predict performance in a subsequent face-to-face math course. Results indicated that students' performance on two of the three units (sets and functions) was positively associated with final grades, whereas total time spent in the course was negatively associated with final grades. We discuss implications for iterative course design as well as utility of educational data mining approaches for tracking preparation for future learning.

Keywords: Feature engineering · Computer Science Education · Educational Data Mining · Prediction Modeling · Interactive Learning Environment

1 INTRODUCTION

The proliferation of data on students interacting with online learning environments has opened up enormous possibilities for understanding student behavior within the last decade or two (Baker & Inventado, 2014). It has also enabled iterative improvements of these learning environments to promote student learning. However, a key challenge is to understand what aspects of students' behavior are most predictive of success in future learning situations.

In recent years, there have been calls to assess learning in terms of “robust” learning outcomes, going above and beyond traditional pretests and posttest which often measure only shallow encoding and retrieval (Koedinger, Corbett, & Perfetti, 2012). Robust learning refers to whether learning occurs in a way that transfers, prepares students for future learning, and is retained over time. While research on learning in online learning environments has been rapidly increasing, much less work has looked at how online courses prepares students for learning during future learning opportunities, including both online or in-person (Beaubouef, 2002). For example, if a student takes an online introductory course in mathematics, we can tell how the student performed

within the system itself, but whether this interaction with online learning prepared the student for future math courses is often unclear (Reilly & Emmett, 2011).

Prior research has attempted to predict student performance on tests of transfer. Specifically, such work has found that avoiding help seeking and making fast responses after bugs were negatively associated with transfer (Baker, Gowda, & Corbett, 2011). Hershkovitz et al. showed that student performance on a transfer test can be predicted by calculating moment-to-moment probabilities of learning a particular skill. Other research has focused on using early course data to predict future success, and develop early warning systems to students identified as at risk for failure (Costa et al., 2017; Dominguez, Bernacki, & Uesbeck, 2016). While prior work on online learning and transfer sheds important light on what attributes of student behavior are critical to transfer, it has largely focused on performance within a single online course. No prior studies to our knowledge have looked at the impact of student performance across multiple online sequential courses or on a future face-to-face course. In this paper, we analyze learning analytics data from an online math primer course and develop a prediction model for performance on an in-person computer science follow-up course students complete.

2 TOOLS & METHODS

2.1 Open Learning Initiative

The Open Learning Initiative (OLI) is an open-ended learning environments that allows instructors to develop online courses consisting of interactive activities and diverse multimedia content. Detailed student interactions with the course materials, such as watching videos or answering questions are logged in the course's database. OLI courses, such as the one used in this study, are often intended to be used asynchronously without an instructor. Prior research has compared student learning from a stand-alone OLI course on introductory statistics to face-to-face equivalent instruction, and found that students showed increased learning gains in half the time as compared to students with the traditional face-to-face instruction (Lovett, Meyer, & Thille, 2008). While this system has been proven to be effective, no studies around it have measured the transfer of the content to future in-person courses. This is true for many online learning environments, while they are proven effective for learning, studies do not look at their transfer and retention when the knowledge is required for a follow-up in-person course, such as a traditional undergraduate one.

2.2 Data Description

Our predictor data came from the Discrete Math Primer (DMP) OLI course, which was completed by incoming freshmen at Carnegie Mellon University during the summer of 2016. This course serves as a prerequisite for core computer science courses, providing students with a foundation for key concepts in the field, such as the notion of data structures. The course is divided into three units — Logic, Sets, and Functions, with which students interacted in a sequential manner. The final grades from the follow-up in-person course, Mathematical Foundations for Computer Science (MFCS), were used as our predictive variable. The final grade was calculated as a percent out of 100. This course was taken by the same students the following semester during Fall 2016 and was taught in a traditional in-person lecture and recitation format. From the syllabus of the follow-up course, proofs is one of the five listed key topics, which makes use of the Logic unit. Functions and Sets is another

key topic of the five listed covered in the follow-up course, which takes a deeper dive on the concepts than what is covered by the online DMP course. Performance in this online course is appropriate for predicting the performance on the follow-up as it directly builds upon the topics covered in the DMP course and is thusly a prerequisite of the MFCP course.

Our dataset consists of 34,999 transactions from 139 students. The transactions consist of student actions in the OLI course, such as selecting an answer in a multiple-choice question, requesting a hint, and submitting an answer. These data entries detail UI events, question correctness, time on task, performance on checkpoints, and hints where relevant. In total, the data spans 198.5 hours of student activity in the course. The course consists of twenty three pages, not including the three quizzes, and is comparable in length to a textbook page. Each page consists of instructional text that is interspersed with low-stakes questions that give detailed feedback intended to foster learning. The Logic unit consists of forty-three questions, Sets has twenty-three, and Functions consists of fifty-one for a total of 106 questions we had student data from in the course. Table 1 shows the variables that we used for our analysis.

Table 1: A description of each variable used in the dataset

Variable	Description
ID	A hashed string corresponding to the student
Duration	The time, in seconds, a student interacted with an element, such as a question
Student Response Type	Denotes the student's action, whether it be a hint request, question attempt, page view, or saving their question answer
Level (Module)	States which of the three units the transaction came from
Step Name	The unique name for the part(s) of a problem, each step contains an opportunity for a correct or incorrect response
Outcome	If applicable, whether the student got the problem correct or incorrect
Attempt at Step	Denotes the amount a student has attempted a given question step
Skill	The label for the skill associated with the particular problem step

2.3 Feature Engineering

We performed feature engineering to construct seven key predictors. Prior research has shown that students who perform at or below a failing grade level in an online course tend to have fewer interactions (Davies & Graff, 2005). Each entry in our dataset represents a student transaction, so we were able to count the numbers of transactions each individual student made through the course. Once the data was filtered on a per-student transaction basis, the total duration each transaction took could be summed to generate a student's total duration in seconds within the course.

As previously described, the course is divided into three units — Logic, Sets, and Functions. Each of these concludes with a summative quiz covering the core material covered in the unit. Each

quiz consisted of eight questions, and students were only allowed a single attempt per quiz question. The grade for each quiz was calculated by summing the number of correct questions out of eight possible points. This yielded three of our seven analyzed features, which were the final quiz grades for each unit.

For each student transaction that details the submission of a question, the OLI platform denotes if it is the student's first attempt at the question. Subsequently if they attempted the problem again, such as changing their answer and submitting, the following entry for the attempt would be marked with a two in the corresponding column. Using this attempt count in conjunction with the outcome, correct or incorrect of the problem attempt, we are able to determine the accuracy of a student's overall attempts as a percentage. Knowing the student's number of attempts at a question and its outcome also allows us to calculate their accuracy on the last attempt, our final feature. In total, this gives us the following seven features:

1. Number of transactions
2. Duration in course
3. Logic quiz grades
4. Sets quiz grade
5. Functions quiz grade
6. Accuracy of overall attempts
7. Accuracy on last attempt

2.4 Random Forest & Linear Regression

We used random forest model, implemented in the R programming language, to predict final grade performance in the follow-up in-person MFCS course. Random forest modeling is a classification and regression algorithm that estimates the amount of increase in mean squared error for each variable, when it is replaced by a set of random values. This provided us with a weighting of how important each of our seven defined features is in the prediction of the final grade. Following this, we used linear regression to predict the nature of the relationship of the predictor variables from our model and to estimate what percentage of variation in final grades was explained by each predictor variable.

3 RESULTS

The results of the random forest modeling indicated the following variables contributed to the increase in mean square error: total number of transactions, quiz grades for the Sets unit, quiz grade for the Functions unit, number of correct and incorrect attempts, and the duration of time spent in the course, see Figure 1.

A simple linear regression analysis was conducted to predict final grade based on the variables found to be associated with an increase in the mean square error. A significant regression equation was found, $R^2 = .43$, $F(7,120) = 12.55$, $p < .001$. Results indicated that the accuracy scores on the Sets ($t = 2.25$, $p = .02$) and Functions quizzes ($t = 2.10$, $p = .037$) had a significant positive association with final grade. The total number of transactions was negatively correlated with final grade ($t = -3.07$, $p = .002$). The regression performed on last attempt correct and incorrect was found to not be significant. It is interesting to note that while only the scores on the Functions and Sets quizzes were positively associated with the final grade on the subsequent course, it was not because

students were already performing at ceiling levels on the Logic module. Mean scores for the Logic and Sets quizzes were 78% and 77% respectively, whereas mean for the Functions quiz was significantly lower at 55%.

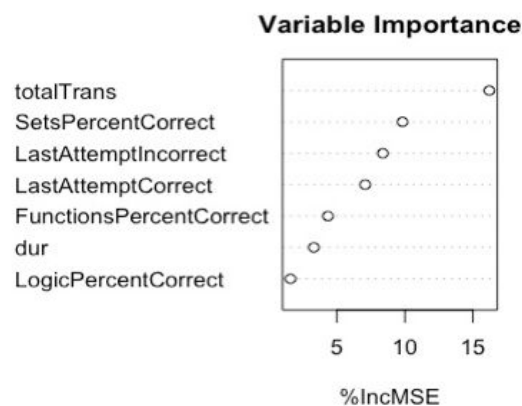


Figure 1: Variable importance plot using random forest modeling

4 DISCUSSION

In this paper, we describe some preliminary results on how students' performance in an online course can be used to predict their learning and performance in a future course. We found that students' performance on two of the three modules in the online OLI course significantly predicted final grades on the subsequent in-person course. The predictive power of the Sets and Functions units, but not the Logic unit, may be explained by the sequence they occur in for both the online DMP course and then in the in-person MFCP course. In the follow-up course, the Proofs section uses subject matter from the Logic unit, and occurs early on in the course. It may be the case that this is a minor section and not a heavily contributing portion of their final grade, since it is the very first part. However, Sets and Functions occurs in the middle of the follow-up course and is taught together. Since these two units are taught together in the follow-up course, it is likely that a student who did not perform well on these two units from the online DMP course will lack the required prior knowledge for this topic and vice-versa. Additionally as it falls in the middle of the course, it might be the case that midterms, an often large portion of a student's grade, occurs during this unit and contains a sizeable portion of material from Sets and Functions.

We found the total number of transactions was found to be negatively associated with final grades in the subsequent course. This is in contrast with prior work that showed that fewer interactions with the online learning system were associated with less learning (Rovai & Barnum, 2007). The system the course is implemented in, OLI, is intended for students to practice on low-stakes activities, not necessarily getting the questions right on the first attempt. However, if students read the accompanying instructional materials on the page, they should be able to answer the questions on the first try. This result of more transactions correlating to a lower grade could be attributed to guess-and-check behavior, where students omit reading the materials and attempt the questions until they achieve the correct answer. Attempting the problems in this system and many others is not technically discouraged, since they contain rich feedback that serves as an instructional moment. Unfortunately many students do not always read the feedback and believe they understand the content once the correct answer is achieved, even if it is by guessing.

Next, the evidence that the online discrete math primer helped students' performance in the subsequent course is only correlational. There are many factors that can come into play between the completion of the online course and conclusion of the follow-up one. However, these results demonstrate how online learning environments may make use of data they are already collecting, quiz scores and formative assessment answers, in a way that feed into a greater predictive system. Predictive modeling is a growing research area with many resulting systems suggesting interventions for at-risk students, based on the input data (Roblyer & Davis, 2008; Essa & Ayad, 2012). Such systems or similar methods could be integrated into the OLI platform, make use of this data, and provide interventions to the students that might fall into the at-risk category.

In sum, predicting future performance using student interaction data in an online course is a promising area of research, and should continue to be explored in the educational data mining literature. The insights gained will help improve student learning not only as measured by pre and post tests within the course, but will ensure that robust learning that prepares students for future learning opportunities is supported.

5 FUTURE WORK

As predictive modeling research continues and integrates with more systems, we hope to find trends across platforms that indicate a set of features that are continuously correlational. Future work in this area could also focus more on not only proving the effectiveness of the system for immediate learning, but for robust learning that transfers to later contexts where it is then prior knowledge. Looking at the transfer of this material from an online course context to an in-person one, like in this study, can help to indicate what makes online learning effective or not. With so many instructional materials and services online that claim to be effective, gauging the long term retention of what they teach is key to them truly being successful for learning. Additionally, future work in CS education should also consider courses in the curriculum that do not strictly rely on programming, such as this studies DMP course. Mathematical foundations are essential in certain aspects of programming and computational thinking, yet many transfer studies focus solely on programming contexts.

One limitation of the present work is that we did not have a measure of students' incoming mastery of the content of the DMP course. We are currently replicating the study with a new cohort of students, who took a short pretest at the beginning of the course, and the quizzes for each module included three items from the pretest to serve as a posttest. Analyses of pre and posttests will give a clearer window into what students learned from the online course, instead of simply measuring their performance on a test. We suggest future work in this area do the same, providing students with a concrete pretest and posttest to effectively evaluate their learning from the online materials. To further obtain a stronger causal evidence for its efficacy, a randomized controlled experiment, where one group of students completes the OLI course, whereas another completes a comparable activity of similar duration would be recommended.

REFERENCES

- Baker, R., Gowda, S., & Corbett, A. (2011). Towards predicting future transfer of learning. In *Artificial intelligence in education* (pp. 23-30). Springer Berlin/Heidelberg.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer, New York, NY.
- Beaubouef, T. (2002). Why computer science students need math. *ACM SIGCSE Bulletin*, 34(4), 57-59.
- Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247-256.
- Davies, J., & Graff, M. (2005). Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4), 657-663.
- Dominguez, M., Bernacki, M. L., & Uesbeck, P. M. (2016). Using Learning Management System Data to Predict STEM Achievement: Implications for early warning systems. Paper presented at the Educational Data Mining Conference, Raleigh, NC.
- Essa, A., & Ayad, H. (2012, April). Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 158-161). ACM.
- HersHKovitz, A., Baker, R., Gowda, S. M., & Corbett, A. T. (2013, July). Predicting future learning better using quantitative analysis of moment-by-moment learning. In *Educational Data Mining 2013*
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757-798.
- Lovett, M., Meyer, O., & Thille, C. (2008). JIME-The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*, 2008(1), 13-26
- Reilly, Christine F., & Emmett Tomai. An examination of mathematics preparation for and progress through three introductory computer science courses. "Frontiers in Education Conference (FIE), 2014 IEEE. IEEE, 2014.
- Roblyer, M. D., & Davis, L. (2008). Predicting success for virtual school students: Putting research-based models into practice. *Online Journal of Distance Learning Administration*, 11(4).
- Rovai, A. P., & Barnum, K. T. (2007). On-line course effectiveness: An analysis of student interactions and perceptions of learning. *International Journal of E-Learning & Distance Education*, 18(1), 57-73.

Analyzing Score and Time Trails in Data Collected by Tutors

Bishal Regmi and Amruth N Kumar

Ramapo College of New Jersey, Mahwah, NJ, USA

{bregmi1,amruth}@ramapo.edu

ABSTRACT: We analyzed the data collected by an intelligent tutor on *if/if-else* statements to see if we could find patterns in score trails and time trails that could be used for predictive purposes. We found that using frequency counts of score trails is well suited for evaluating the effectiveness of software tutors. On the other hand, the wide variability in time trails makes them unusable for predictive purposes when the tutors are used *in-natura*.

Keywords: Intelligent tutors, Educational data mining, k-means clustering

1 INTRODUCTION

Learning curves are popularly used to model the rate of learning. They are typically applied to the number of opportunities availed by a student to practice a skill (e.g., (Cen, Koedinger & Junker, 2006)). An example of these opportunities is the number of hints provided by a tutor while solving a problem. We wanted to model the rate of learning across problems on a concept, i.e., inter-problem or outer loop of a tutor (VanLehn 2006) instead of within a problem, i.e., intra-problem or inner loop of a tutor. We also wanted to apply learning curves to the data collected by a software tutor used *in-natura*, i.e., by real students in unsupervised learning conditions.

For this study, we used problets (problets.org), a suite of 17 software tutors on programming topics. Each problet addresses multiple concepts in one programming topic (e.g., *while* loop, *switch* statement). It presents code-tracing problems, wherein, the student is asked to identify the output of a program, debug a program or predict the state of variables in a program. It provides step-by-step explanation of the correct answer as feedback (Kumar, 2006), in the style of a worked-out example (Sweller & Cooper 1985). This strategy of combining worked-out examples with problem-solving has been shown to benefit learners better than problem-solving alone (Cooper & Sweller, 1987). Problets are adaptive (Kumar, 2006a), and use pretest-practice-post-test protocol to administer problems every time a student uses a problet.

In this study, we attempted to fit a learning curve to the data collected by a problet on *if/if-else* statements. We tried to find patterns in score and time trails and use them to predict when a student had learned a concept or needed further remediation. For the purposes of this study, we define a trail as a sequence of data points, wherein, each data point corresponds to a problem solved using the tutor. For example, a score trail is the sequence of scores earned by students on the problems they solved on a concept.

The tutor covered 9 concepts in Java/C# and 12 concepts in C/C++ (e.g. nested *if-else* statements, back-to-back *if-else* statements, etc.). On each concept, the tutor first presented a pretest problem to prime the student model. If the student solved the problem incorrectly, solved

partially, or skipped solving the problem, the tutor presented practice problems on the concept along with feedback. This feedback comprised of step-by-step explanation of the correct solution. The tutor presented practice problems until the student had mastered the concept, i.e., solved a minimum percentage of problems correctly. Finally, the tutor presented a post-test problem to assess whether the student had learned the concept.

The pretest and post-test problems were the same for all the students. The sequence of practice problems was also the same for all the students. However, each student solved a different number of practice problems due to the adaptive nature of the tutor.

The tutor was used by students in introductory programming courses at high schools, community colleges and baccalaureate institutions as after-class assignment. So, the use was *in-natura* as opposed to *in-ovo* or *in-vivo*: real students in an unsupervised setting as opposed to research subjects in a laboratory setting.

The concepts covered by a proplet can be classified as known, learned, practiced, or attempted as shown in Table 1. A concept is known when the student did not need to use the tutor, i.e. solved the pretest problem correctly. A concept is learned when the student learned by using the tutor. A concept is practiced and not learned when the student failed to solve the post-test problem correctly. If so, the tutor schedules additional practice problems on the concept for the student. Lastly, a concept is attempted when the student could not complete mastering the concept during practice because of the 30-minute limit placed on the duration of the tutoring session.

Table 1: Types of concepts based on learning outcomes in a proplet

Pretest Correct?	Practice	Post-test	Type of Concept
Yes			Known
No	Some/None		Attempted
No	Mastered	Incorrect	Practiced
No	Mastered	Correct	Learned

In the past, we have quantified the learning that occurs with proplets in terms of the number of concepts learned and pre-post improvement in score on the learned concepts (e.g., Kumar, 2016). The objectives of the current study were two-fold:

1. Could we fit a learning curve to the scores data collected by the tutors in an attempt to quantify the rate of learning of each concept?
2. Could we use the time being spent by the student on a problem to predict whether the student knew how to solve the problem or to intervene with affective feedback such as that inspired by growth mindset theory (Dweck,2012)?

2 DATA COLLECTION AND ANALYSIS

We used the data collected by the proplet on `if/if-else` statements over eight semesters: Fall 2012-Spring 2014 and Fall 2015-Spring 2017. During that time, 4,458 students used the tutor. The

tutor recorded the score and the time taken to solve each problem. In each problem, the student studied a program and predicted its output, one line at a time, and in the correct order. So, the answers to the problems were free-form (as opposed to multiple-choice) – the student had to use a correct mental model of the program to solve each problem. For analysis purposes, the scores were normalized to the range 0 → 1.0.

Given that the tutor covered 9-12 concepts per student, after eliminating concepts listed as known in Table 1, we extracted 3,578 student-concepts from the data, i.e., 3,578 concepts that were practiced or learned by students. First, we analyzed the frequency counts of score and time trails for each concept.

3 SCORE TRAILS

Table 2: Score trails with frequency count

Trail No.	Trail	Frequency	Percent age (%)	Trail No.	Trail	Frequency	Percent age (%)
Concept 1 (N=398)				Concept 8 (N=451)			
1	0-1-1	216	54.27	1	0-1-1	216	47.89
2	0-0-1-1-1	65	16.33	2	0-1-1-1	93	20.62
3	0-1-0-1-1	41	10.30	Concept 9 (N=604)			
Concept 2 (N=65)				1	0-1-1	252	41.72
1	0-1-1	27	41.54	2	0-1-1-1	82	13.58
Concept 3 (N=119)				Concept 10 (N=789)			
1	0-1-1	78	65.55	1	0.5-1-1-1	108	13.69
Concept 4 (N=136)				2	0-1-1-1	73	9.25
1	0-1-1	56	41.18	3	0.5-0.5-1-1-1-1	51	6.46
2	0-1-1-1	11	8.09	Concept 11 (N=309)			
Concept 5 (N=56)				1	0-1-1-1	177	57.28
1	0-1-1	34	60.71	2	0-0-1-1-1-1	22	7.12
Concept 6 (N=345)				Concept 12 (N=242)			
1	0-1-1-1	141	40.87	1	0-1-1-1	97	40.8
2	0-0.5-1-1-1-1	35	10.14	2	0-0-1-1-1-1	18	7.44
Concept 7 (N=114)							
1	0-1-1-1	20	17.54				

Table 2 lists the frequency counts of the score trails that occurred the most often for each concept. In each case, we listed all and only the trails subscribed to by more than a handful of students. Each trail begins with the score from the pretest. For example, on concept 1 (practiced by 398 students), 54.27% of the students scored 0 on the pretest problem, 1 on the practice problem and 1 on the post-test problem. As expected, all the score trails start with a score of either 0, i.e., the student did not solve the pretest problem correctly, or 0.5, i.e., the student solved the pretest problem partially correctly, triggering adaptive practice. All the score trails end in 1, corresponding to either the student having learned the concept (scored 1 on the post-test) or having attempted it (scored 1 on the last practice problem). There were score trails that ended with less than 1 corresponding to the student having practiced the concept (scored less than 1 on the post-test) or having attempted it (scored less than 1 on the last practice problem). However, those trails were excluded from the table because the frequency of such trails was low. Since all the problems had ranges of discrete scores, (e.g., 0, 0.5, or 1 on most problems), the learning curve in all the cases was a step function, a

discrete function as compared to the continuous exponential function traditionally used for learning curves.

We observed a 1-0 step down transition, suggesting either an earlier guess or a later slip in solving problems as described in Bayesian Knowledge Tracing (Baker, Corbett & Aleven, 2008) in only one score trail (trail 3 on concept 1). So, in all but one case, the improvement in learning was monotonic. The feedback provided for incorrect solutions seemed to have been sufficient to help solve all subsequent problems correctly.

Given the step pattern of learning, the number of incorrect solutions before transitioning to learned state is a measure of the speed of learning. The speed of learning varied from 1 problem (in most of the learning trails) to 2, observed on concepts 10, 11 and 12. This empirical evidence suggests that at least a small percentage of students found concepts 10, 11 and 12 to be harder than the other concepts.

The score trails on some problems were longer than on the others. This was because the mastery learning criterion used for adaptive practice required students to solve at least 50% of the problems correctly on concepts 1-3 and 60% of the problems on all the other concepts. This resulted in shorter trails for concepts 1-3. The tutor credited concepts 4 and 5 when problems on concept 6 or 7 were solved correctly. Similarly, the tutor credited concepts 8 and 9 for solving concepts 10, 11 or 12 correctly. Therefore, the student could demonstrate mastery of some concepts without solving problems solely dedicated to the concept, resulting in shorter score trails.

On some concepts, viz., 6, 7, 10, 11 and 12, students solved 3-4 consecutive problems correctly at the end including the post-test problem. On these concepts, the mastery criterion used by the adaptive tutor may be requiring students to solve more problems than necessary. In the future, we plan to consider other models such as Bayesian Knowledge Tracing to minimize the number of redundant problems solved by the students.

The rarity of 1-0 transition in the score trails is noteworthy. When students use a tutor *in-natura*, finding trails that defy explanation was to be expected. For example, some of the score trails we found for concept 1 were: 0-0-0-0-1, 0-1-0-0-1-1-1, and 0-0-1-1-0-0-0. However, the frequency counts of these trails were in the single digits, indicating that they were outliers. Our approach of using frequency counts of score trails helps disregard such outliers inherent to *in-natura* use of tutors, and is therefore, well suited for evaluating the effectiveness of software tutors used *in-natura*.

4 TIME TRAILS

Unlike scores on a problem, time taken to solve each problem had much greater variability. So, for each problem, we eliminated outliers and ran k-means clustering for values of $k = 3 \rightarrow 9$. We determined the optimal k value for each problem using elbow method, i.e., the value of k past which, the drop in sum of squares error (SSE) slows. Next, we replaced the time taken by each student on a problem with the centroid value of the cluster to which the student belonged on that problem as per k-means clustering with the optimal value of k selected earlier. Finally, we replaced

the raw time trail of each student with the trail of centroid values and did a frequency count on the centroid time trails. Table 3 lists the most frequent centroid time trails for each concept.

Remarkably, even the most frequent time trail constitutes less than 8% of all the trails on any concept as seen in Table 3. So, there is far less consistency in the time taken by students to solve problems. Thus, it does not have normative value in a tutor used *in-natura*; it cannot be used to determine whether and when intelligent remediation should be provided to the student.

In most time trails, time drops from the first (pretest) problem to the second (first practice) problem. Subsequently though, time was just as likely to increase as to decrease. So, the time spent is more dependent on the problem than on the level of familiarity of a student with the underlying concept, familiarity that is expected to increase with every problem solved. This again calls into question the utility of time trails for evaluating the effectiveness of a tutor used *in-natura*.

Table 3: Time trails with frequency count

Trail Number	Trail	Frequency	Percentage (%)	Trail Number	Trail	Frequency	Percentage (%)
<u>Concept 1 (N=398)</u>				<u>Concept 8 (N=451)</u>			
1	27-18-25	14	3.5	1	47-13-20	16	3.5
<u>Concept 2 (N=65)</u>				2	100-13-20	15	3.3
1	66-10	5	7.7	<u>Concept 9 (N=604)</u>			
<u>Concept 3 (N=119)</u>				1	54-21-29	26	4.3
1	27-14-27	4	3.4	2	139-21-29	24	4.0
<u>Concept 4 (N=136)</u>				<u>Concept 10 (N=789)</u>			
1	39-11-12	6	4.4	1	44-24-25	8	1.0
<u>Concept 5 (N=56)</u>				<u>Concept 11 (N=309)</u>			
1	36-18-16	4	7.1	1	70-29-23-29	7	2.3
<u>Concept 6 (N=345)</u>				<u>Concept 12 (N=242)</u>			
1	22-32-15-37	7	2.0	1	40-18-13-40	5	2.1
<u>Concept 7 (N=114)</u>							
1	30-20-10	4	3.5				

5 DISCUSSION

Typically, learning curve is applied to the number of opportunities available by a student to practice a skill (e.g., (Cen, Koedinger & Junker, 2006)). Such opportunities may be the number of hints provided by the tutor. We on the other hand tried to fit a learning curve to the score earned by the student on successive problems solved on a concept.

We found that the learning curve was a step curve, given that scores on problems were a range of discrete values. We also found that using frequency counts of score trails is well suited for evaluating the effectiveness of software tutors when they are used *in-natura*.

Time spent solving a problem has been of interest to researchers in Intelligent Tutoring Systems. Gowda *et al* (2013) reported finding a relation between response times and shallow learning, concluding that “shallow learners tend to have slower response times than deep learners”. Jarušek & Pelánek (2012) tried to model and predict problem-solving times based on student’s ability. Soh

(2006) found that the total amount of time spent by a student on an intelligent tutor had no correlation with his or her total exam score. We found that the time spent solving a problem had no normative value when the intelligent tutor is used *in-natura*. Time trails are of questionable value for evaluating the effectiveness of such tutors even after time values are normalized.

In the future, we plan to repeat this study for a harder topic such as loops and see if we can reproduce the same results. We plan to replace mastery learning model with Bayesian Knowledge Tracing to reduce the number of redundant problems solved by students. We also plan to consider models that combine score and time trails to see if a combined model can provide further insight into our data. Finally, we plan to apply decision trees to see if we can build a predictive model based on score and time trails

6 ACKNOWLEDGEMENTS

Partial support for this work was provided by the National Science Foundation under grant DUE-1432190.

REFERENCES

- Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008). More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. pp. 406-415.
- Cen. H, Koedinger K and Junker, B. (2006). *Learning Factors Analysis – A General method for Cognitive Model Evaluation and Improvement*. iProc. ITS 2006. pp. 164-175.
- Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79. pp. 347-362.
- Dweck, C. (2012). *Mindset: The New Psychology of Success*. Constable & Robinson Ltd.
- Gowda, S.M., Baker, R.S., Corbett, A.T. and Rossi, L.M. (2013). Towards Automatically Detecting Whether Student Learning is Shallow. *International Journal of Artificial Intelligence in Education*. pp. 23: 50.
- Jarušek P., Pelánek R. (2012). *Modeling and Predicting Students Problem Solving Times*. In: Bieliková M., Friedrich G., Gottlob G., Katzenbeisser S., Turán G. (eds) SOFSEM 2012: Theory and Practice of Computer Science. SOFSEM 2012. Lecture Notes in Computer Science, Vol 7147. Springer, Berlin, Heidelberg. pp. 1-12.
- Kumar, A.N. (2016). *The Effectiveness of Visualization for Learning Expression Evaluation: A Reproducibility Study*. Proc. ITiCSE 2016, Arequipa, Peru. pp. 192-197.
- Kumar, A.N. (2006). Explanation of step-by-step execution as feedback for problems on program analysis, and its generation in model-based problem-solving tutors. *Technology, Instruction, Cognition and Learning (TICL) Journal, Special Issue on Problem Solving Support in Intelligent Tutoring Systems*, Vol 4(1).
- Kumar, A.N. (2006a). A Scalable Solution for Adaptive Problem Sequencing and its Evaluation. Proc. *The Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 06)*. Dublin, Ireland. pp. 161-171.
- Soh, Leen-Kiat. (2006). Incorporating an intelligent tutoring system into CS1, *Proceedings of the 37th SIGCSE technical symposium on Computer science education*, March. pp. 486-490.

- Sweller, J., Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction* 2. pp. 59–89.
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3). pp. 227-265.

A Comparison of Two Designs for Automated Programming Hints

Thomas W. Price¹, Joseph Jay Williams², Samiha Marwan¹

¹North Carolina State University; ²University of Toronto

twprice@ncsu.edu, williams@cs.toronto.edu, samarwan@ncsu.edu

ABSTRACT: A growing body of work has explored how to automatically generate hints for novice programmers. However, little research has explored what trade-offs exist between different ways of providing hints, how users perceive them, and how they can be combined. In this work, we present preliminary data from a study comparing different types of hint support in a block-based programming environment, focusing on code hints and explanatory text hints. We conducted a study in which crowd workers completed two programming tasks with different combinations of hint support. We found that both code hints and text hints are rated as very helpful by users, and showing both together is rated as the most useful. Only code hints improved users' performance during programming on the current task.

Keywords: Automated hints, intelligent programming support, novice programmers

1 INTRODUCTION

There is clear evidence that novices find programming to be a very difficult task to learn (Bennedsen et al., 2007; Watson & Li 2014), and researchers have developed a variety of intelligent support tools to assist them. Programming hints are a popular form of support (e.g. Perelman et al., 2014; Lazar et al., 2017; Yi et al., 2017), since they can be generated automatically, often using student data (Price et al., 2017b; Rivers & Koedinger, 2017), allowing them to scale to new problems and contexts. These automated hints are typically presented as next-step "code hints," which suggest an edit that the student should make to their program to bring it closer to a correct solution, allowing them to proceed when stuck. While small-scale studies suggest that code hints have the potential to resolve student difficulties (Price et al., 2017a), students can find code hints difficult to interpret without explanations (Price et al., 2017c). Others have pointed out that these "bottom-out" hints, which give away part of the correct solution, may not lead to learning (Aleven & Koedinger, 2016; Paaßen et al., 2018). While other types of automated hints have been proposed (Suzuki et al., 2017), little work has explored how different hint types compare and interact in the domain of programming.

In this paper, we present preliminary results from a larger study exploring the design space of programming hints. We evaluated both "code hints" and explanatory "text hints" that explain a domain concept and connect it to the current problem. We report results from a randomized experiment in which Mechanical Turk workers completed two simple programming tasks with different combinations of code and text hints. We find that code hints are rated as most helpful by users, and they can also improve users' performance on the current task. We find that text hints are also regarded as helpful, though less so than code hints, and they offer a complementary benefit to code hints. However, they do not appear to contribute to users' programming performance.

2 METHODS

Design of Hint Support: In this work, we build on an existing system called iSnap (Price et al., 2017a), a block-based, novice programming environment that supports students with hints and feedback. iSnap's hints are generated by the data-driven SourceCheck algorithm (Price et al., 2017b), which uses a database of correct solutions for a given problem to generate hints automatically. Like other data-driven hint-generation systems (e.g. Rivers & Koedinger, 2017; Piech et al., 2015; Paaßen et al., 2018), iSnap's hints are *next-step*, *edit-based* hints, suggesting a specific edit to the student's code which will bring them closer to a correct solution. In this study, we augmented iSnap's "code hints" with explanatory "text hints" that say not only *what* to do but also *why*. Similar to principle-based hints (Dutke et al., 2008) or teaching hints (VanLehn et al., 2005), these text hints explain a relevant programming concept and then connect it to the problem objectives. For example, for an assignment to write a procedure for drawing a polygon, one text hint reads, "The `repeat` block allows you to run the same code a fixed number of times, like drawing each side of a polygon." To add text hints to iSnap for a given problem, we tagged each block in the correct solution to that problem with one or more relevant text hints. Anytime iSnap would show a hint to add that block, iSnap shows the corresponding text hint, either alongside of in place of the code hint, depending on a user's condition (explained below). In this experiment, we generated hints using a comprehensive set of expert-authored solutions, rather than student solutions, as these have been shown to produce higher-quality hints (Price et al., 2018). We tagged each expert-authored solution with text hints to ensure that we supported a variety of solutions.

Population: We recruited 233 total crowd workers through Amazon's Mechanical Turk platform, which has been suggested as an appropriate alternative to university participants (Behrend et al., 2011; Kittur et al., 2008), including CS education research (Lee & Ko, 2015). We studied crowd workers because this allowed us to recruit a larger number of participants, collect fine-grained survey data, and give participants different, uneven levels of support – all of which are difficult in a real classroom setting. We analyzed data from 209 participants (excluding 24 participants due to data collection errors). We only recruited participants who attested to having no programming experience (no courses or workshops), and we paid users \$4-7 to complete the study (varying the amount to increase the speed of recruiting). We did not collect any demographic information from participants.

Procedure: Participants first read through a short tutorial on programming in iSnap for approximately 5 minutes, which covered the user interface of iSnap and explained all programming concepts needed for the later programming tasks (loops, input/output and drawing) using a combination of text and short example animations. Next, participants worked on a programming task (Task 1) for 15 minutes, in which they were to create a program to draw a polygon with any number of sides (chosen by the user). During this programming task, each user was randomly assigned to a condition that determined what type of help iSnap provided for the whole task. iSnap either provided no help, code hints only, text hints only, or code and text hints together, for 4 total conditions. Additionally, for participants who received hints, their condition dictated whether or not they received prompts to self-explain the hints (Chi et al., 1994), but in this preliminary work we analyze these two sub-conditions together. While participants programmed, every two minutes

iSnap interrupted them to take the action dictated by their condition (e.g. showing a code hint), and then asked them for their thoughts on the action (post-help survey). While this timed approach differs from the typical, on-demand way that users request hints in existing systems (Rivers & Koedinger, 2017), previous work shows many users will avoid or abuse help when they can request it on-demand (Aleven et al., 2016; Price et al., 2017c,d). Our timed approach ensures that each user receives hint support frequently and regularly, enabling us to collect more extensive data about perceptions of hints. After this first task, users completed a second 15-minute programming task (Task 2), which was similar to the first task but more challenging. During this task, users received help every 2 minutes, as in Task 1, but the type of help was randomized and independent of their Task 1 condition, allowing us to use it as a form of post-test.

Measures: The work reported here is part of a larger analysis of the experiment, which involved a number of survey measures. However, here we only focus on the post-help survey and users' performance on the two programming tasks. We measured the latter by defining 4 objectives for Task 1 and for Task 2 (e.g. "draw a shape" or "correctly get and use input from the user"), such that each objective was independent, and completing all 4 indicated successful completion of the whole task. We developed an automatic grader to determine which objectives participants completed and verified approximately half of the grades manually. Since users received help in Task 1 based on their condition, it is used as a measure of programming *performance*. In Task 2, the help users received was randomized and independent of their original condition, and it is therefore used as a measure of *learning*, since any differences in performance among the conditions can be attributed to knowledge gained as a result of different programming support on Task 1.

3 RESULTS

We investigated two research questions about the impact of code and text hints on users' outcomes:

RQ1: *How does the type of hint support that users received impact their perception of iSnap's usefulness?* After Task 1, each user rated how useful iSnap's actions were overall (from 0 to 10). Figure 1 shows the distribution of these ratings, organized by what type of help iSnap provided every 2 minutes.

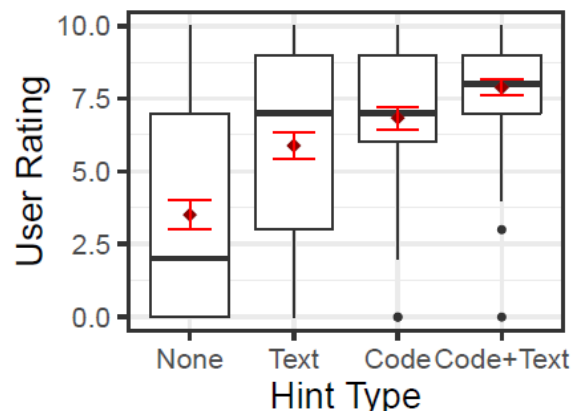


Figure 1: User ratings of how helpful they found iSnap's actions in Task 1, split by condition.

We analyzed users' usefulness ratings using a between-subjects ANOVA. We tested for the main effects of receiving code hints and text hints, as well as the interaction effect between code hints and text hints. We found a significant main effect for receiving code hints ($F(1,205) = 30.3$; $p < 0.001$) and text hints ($F(1,205) = 17.1$; $p < 0.001$), and we did *not* find a significant interaction effect between receiving code and text hints ($F(1,205) = 2.57$; $p = 0.111$). This suggests that users who received code or text hints in Task 1 rated iSnap's actions as significantly more useful than those who did not, and the lack of interaction suggests that there is an additive benefit to receiving *both* code and text hints. Confirming this, we found that the action usefulness ratings from users who received code hints *with* text hints ($N=57$; $M=7.88$; $SD=2.08$) were significantly higher than those who received *only* code hints ($N=43$; $M=6.84$; $SD=2.53$), as indicated by a Mann-Whitney U -test ($U=903.5$; $p = 0.023$).

RQ2: *How do code and text hints impact users' performance on current and future programming tasks?* Figure 2 shows the distribution of objectives completed by users on Tasks 1 and 2, split by the type of support they received *on Task 1*. For Task 1, we conducted an ANOVA on the number of objectives users accomplished, testing the main effects of receiving code hints and text hints and the interaction effect of receiving code hints and text hints. We found a significant main effect for receiving code hints ($F(1,205) = 16.7$; $p < 0.001$) but not text hints ($F(1,205) = 0.052$; $p = 0.821$), and we did not find a significant interaction effect between receiving code and text hints ($F(1,205) = 0.003$; $p = 0.960$). This suggests that only code hints contributed to users' programming performance on Task 1. The number of objectives completed was greater for users who received code hints ($N=100$; $M=2.02$; $SD=1.50$) than those who did not ($N=109$; $M=1.20$; $SD=1.35$), and the difference was significant ($U=7118.5$; $p<0.001$) with a medium effect size (Cohen's $d = 0.575$).

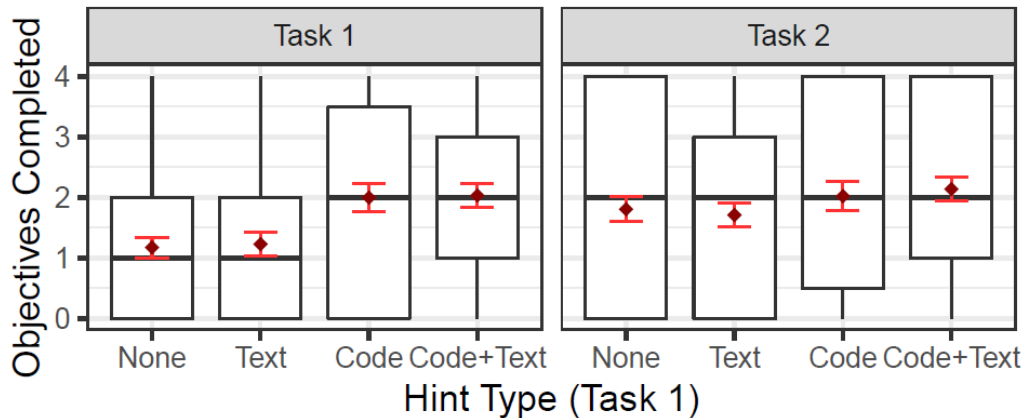


Figure2 : The number of objectives completed by users in Task 1 and Task 2, split by condition on Task 1. Box plots reflect medians and quartiles, and means (diamonds) with standard error bars.

We also investigated users' performance on Task 2 as a function of their original support condition *on Task 1*. We conducted an ANOVA to compare the main effects of having received code hints, text hints and reflective prompts on Task 1. We found no significant main effects for code hints ($F(1,205) = 2.33$; $p = 0.128$), text hints ($F(1,205) = 0.003$; $p = 0.959$), and no interaction between receiving code and text hints ($F(1,205) = 0.253$; $p = 0.615$). While not significant, there was still a small effect of

code hints on the number of Task 2 objectives completed, as shown in Figure 2. The number of Task 2 objectives completed by users who received code hints in Task 1 ($N=100$; $M=2.09$; $SD=1.51$) was greater than those who did not ($N=109$; $M=1.76$; $SD=1.51$), but this difference was not significant ($W=6115.5$; $p = 0.119$) and had a small effect size (Cohen's $d = 0.218$).

4 DISCUSSION AND CONCLUSION

Our results show that learners find both code and text hints to be useful on their own, but the combination of both is perceived as the most useful. This suggests that both types of hints offer different information to the user, and both contribute the perceived utility of the hint. Code hints were also perceived as more useful than text hints, perhaps because they provided an immediately actionable suggestion, which all novices could follow, while text hints required users to interpret and apply the provided domain knowledge. In future work, we will investigate users' survey responses to better understand what benefits they perceived from each type of hint. Despite the perceived utility of both hint types, only code hints improved users' immediate performance, and they had no significant impact on future performance. Text hints did not improve users' performance, either alone or in conjunction with code hints. It is interesting that the perceived utility of text hints did not translate into improved student performance, and we hope to investigate other ways that these hints may have affected user outcomes in future work. It is possible that the impact of text hints would be more apparent over longer assignments. Further, if hints are provided on-demand, rather than every 2 minutes, hints that are perceived as more useful may result in more hint requests.

There are several limitations to this work. Our population consisted of paid crowd workers with no prior programming experience. Their motivations, prior knowledge and priorities may differ from those of other populations of learners where programming hints are used. Additionally, we only studied users during two simple, 15-minute programming tasks, and we have begun further work to investigate if our results generalize to longer or more complex tasks in classrooms. The presence of additional, randomized hints on Task 2 likely added noise to our performance data, making it more difficult to detect the effect of the Task 1 condition. The frequent delivery of hints allowed us to collect rich data about user's perceptions (which we will analyze in future work), but this proactive hint delivery differs from the usual on-demand approach, in which students request hints when needed.

REFERENCES

- Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 26(1), 1–19. <https://doi.org/10.1007/s40593-015-0089-1>.
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3), 800–813. <https://doi.org/10.3758/s13428-011-0081-0>.
- Bennedsen, J., & Caspersen, M. E. (2007). Failure rates in introductory programming. *ACM SIGCSE Bulletin*, 39(2), 32. <https://doi.org/10.1145/1272848.1272879>.

- Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Dutke, S., & Reimer, T. (2008). Evaluation of two types of online help for application software. *Journal of Computer Assisted Learning*, 16(October 2000), 307–315.
<https://doi.org/10.1046/j.1365-2729.2000.00143.x>.
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453–456).
- Lazar, T., Možina, M., & Bratko, I. (2017). Automatic Extraction of AST Patterns for Debugging Student Programs. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 162–174). https://doi.org/10.1007/978-3-319-61425-0_14.
- Lee, M. J., & Ko, A. J. (2015). Comparing the Effectiveness of Online Learning Approaches on CS1 Learning Outcomes. In *Proceedings of the eleventh annual International Conference on International Computing Education Research - ICER '15* (pp. 237–246). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2787622.2787709>.
- Paaßen, B., Hammer, B., Price, T. W., Barnes, T., Gross, S., & Pinkwart, N. (2018). The Continuous Hint Factory -Providing Hints in Vast and Sparsely Populated Edit Distance Spaces. *Journal of Educational Data Mining*, 1–50.
- Perelman, D., Gulwani, S., & Grossman, D. (2014). Test-Driven Synthesis for Automated Feedback for Introductory Computer Science Assignments. In *Proceedings of the Workshop on Data Mining for Educational Assessment and Feedback*.
- Piech, C., Sahami, M., Huang, J., & Guibas, L. (2015). Autonomously Generating Hints by Inferring Problem Solving Policies. In *Proceedings of the ACM Conference on Learning @ Scale* (pp. 1–10).
- Price, T. W., Dong, Y., & Lipovac, D. (2017a). iSnap: Towards Intelligent Tutoring in Novice Programming Environments. In *Proceedings of the ACM Technical Symposium on Computer Science Education*.
- Price, T. W., Liu, Z., Catete, V., & Barnes, T. (2017c). Factors Influencing Students' Help-Seeking Behavior while Programming with Human and Computer Tutors. In *Proceedings of the International Computing Education Research Conference*.
- Price, T. W., Zhi, R., & Barnes, T. (2017b). Evaluation of a Data-driven Feedback Algorithm for Open-ended Programming. In *Proceedings of the International Conference on Educational Data Mining*.
- Price, T. W., Zhi, R., & Barnes, T. (2017d). Hint Generation Under Uncertainty: The Effect of Hint Quality on Help-Seeking Behavior. In *Proceedings of the International Conference on Artificial Intelligence in Education*.
- Price, T. W., Zhi, R., Dong, Y., Lytle, N., & Barnes, T. (2018). The impact of data quantity and source on the quality of data-driven hints for programming. In *Proceedings of the International Conference on Artificial Intelligence in Education*. http://doi.org/10.1007/978-3-319-93843-1_35.
- Rivers, K., & Koedinger, K. R. (2017). Data-Driven Hint Generation in Vast Solution Spaces: a Self-Improving Python Programming Tutor. *International Journal of Artificial Intelligence in Education*, 27(1), 37–64.

- Suzuki, R., Head, A., Soares, G., D'Antoni, L., Glassman, E., & Hartmann, B. (2017). Exploring the Design Space of Automatically Synthesized Hints for Introductory Programming Assignments. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2951–2958). <https://doi.org/10.1145/3027063.3053187>.
- VanLehn, K., Lynch, C., Schulze, K., & Shapiro, J. A. (2005). The Andes physics tutoring system: Five years of evaluations. In *Proceedings of the International Conference on Artificial Intelligence in Education*.
- Watson, C., & Li, F. W. B. (2014). Failure rates in introductory programming revisited. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education* (pp. 39–44).
- Yi, J., Ahmed, U. Z., Karkare, A., Tan, S. H., & Roychoudhury, A. (2017). A Feasibility Study of Using Automated Program Repair for Introductory Programming Assignments. In *Proceedings of the Joint Meeting on Foundations of Software Engineering* (pp. 740–751). <https://doi.org/10.1145/3106237.3106262>.

Using Legacy Data to Build Bayesian Knowledge Tracing Model and Evaluate Its Effectiveness

Vanesa Getseva and Amruth N. Kumar

Ramapo College of New Jersey, Mahwah NJ 07430, USA
{vgetseva, amruth}@ramapo.edu

ABSTRACT: We used legacy data collected by an intelligent tutor on the programming topic of selection that administered pretest-practice-post-test protocol to compute the four parameters of Bayesian Knowledge Tracing model. We calculated the probability of prior knowledge based on the percentage of students who solved the pretest problem correctly, and probability of learning based on the percentage of students who learned the concept using the tutor. We calculated the probabilities of guessing and slipping for each problem presented by the tutor during the practice session. Next, we used 25-fold cross-validation to evaluate whether the resulting BKT model would have reduced the number of practice problems solved and time spent by the students represented in the legacy data. We found that in 69.22% of the cases, students would have saved time with the BKT model. They would have saved a mean of 1.425 minutes and 1.764 problems per student per concept learned using the tutor. These results support the incorporation of Bayesian Knowledge Tracing into the tutor.

Keywords: Intelligent Tutoring System, Student Modeling, Bayesian Knowledge Tracing, Evaluation.

1 INTRODUCTION

Student model is essential for facilitating adaptation in intelligent tutoring systems. Bayesian Knowledge Tracing (Corbett *et al.* 1992) is one of the more popular methods of modeling student's knowledge. The model consists of four parameters per concept. In the past, in order to estimate the four parameters, researchers have used baseline approach (Beck 2007), bounded guess and slip approach, Dirichlet Priors (Beck *et al.* 2007), contextual estimation (Baker *et al.* 2008) and empirical probabilities (Hawkins *et al.* 2014). In this study, we present an empirical approach based on legacy data collected by an intelligent tutoring system that administers pretest-practice-post-test protocol. Our approach differs from earlier attempts in that we calculate guess and slip parameters for each problem, not just each concept. We use the calculated BKT model to evaluate its effectiveness in terms of time and effort saved for the students represented in the legacy data.

Our interest in using Bayesian Knowledge Tracing is to improve adaptation. Currently, the tutor uses a naive scheme to determine whether the student has learned a concept during practice, and therefore, whether the student is ready for post-test on the concept. In this naive scheme, a student is said to have mastered a concept if the student solves a certain percentage (e.g., 60%) of the problems correctly. If the Bayesian Knowledge Tracing model can determine that a student has learned a concept with fewer practice problems, it would in turn reduce the number of unnecessary problems solved and time spent by the student with the tutor.

1.1 THE SOFTWARE TUTOR

For the current study we used legacy data, meaning previously collected data. We used the data collected by a software tutor on selection statements. The tutor covers 12 concepts and uses pretest-practice-post-test protocol during every tutoring session (Kumar 2014).

Pretest is used to prime the student model for adaptation (Kumar 2006a). During the pretest, the student is presented one problem per concept. If the student solves the problem correctly, no more problems are scheduled. If the student solves it incorrectly, step-by-step explanation of the correct solution is provided as feedback (Kumar 2006) in the style of a worked example (Sweller & Cooper 1985) and additional practice problems are scheduled.

During the adaptive practice stage, problems are presented to the student on each of the concepts on which the student solved the pretest problem incorrectly. The tutor keeps presenting practice problems on each concept until the student has demonstrated mastery of the concept. Mastery of a concept is defined as having solved a minimum number of problems and solved a minimum percentage of them correctly. During practice, the tutor provides feedback after each problem.

During the adaptive post-test, the tutor presents a post-test problem on each of the concepts the student has mastered during practice. If the student solves the post-test problem incorrectly, the tutor schedules additional practice problems on the concept. Otherwise, the concept is marked as having been learned.

1.2 BAYESIAN KNOWLEDGE TRACING

Bayesian Knowledge Tracing (Baker *et al.* 2008) has been widely used in intelligent tutors to model student knowledge. It is a two-node hidden Markov model containing mastered and unmastered nodes that uses four parameters: L_i , T , G , S . $P(L_i)$ is the probability a student has mastered a concept at a moment i , $P(L_0)$ being the probability that the concept was mastered by the student before using the system. $P(T)$ is the probability a student will transfer from unmastered to mastered state for a given concept. $P(G)$ is the probability a student guesses, i.e., solves a problem on an unmastered concept correctly. $P(S)$ is the probability a student slips, i.e., solves a problem on a mastered concept incorrectly. We propose to compute the four parameters of the model using legacy data.

2 BUILDING BKT MODEL

2.1 DATA DESCRIPTION

For this study, we considered the data collected by a tutor on selection statements (`if / if-else`) over 8 semesters: Fall 2012 – Spring 2014 (4 semesters), used by 2312 students; and Fall 2015 – Spring 2017 (4 semesters), used by 2146 students. During those 8 semesters, the data was collected from multiple institutions. For this study, we combined the data from all 8 semesters. The data contained C++, Java and C# users. Selection tutor covered 9 concepts generic to C++/Java/C# and 3 additional concepts specific to C++. The tutor used the mastery learning criteria of minimum 1 problem solved (including pretest problem) and minimum 60% of the problems solved correctly for all 9 generic concepts. For the 3 additional C++ concepts, it used the criterion of a minimum of 50%

of the problems solved correctly.

Each problem presented by the tutor contained a complete program: the student was asked to predict the output of the program one line at a time. Even though the number of lines of output varied from one problem to another, the score on every problem was normalized to $0 \rightarrow 1.0$: 0 when the student did not correctly identify any line of output; 1.0 when the student correctly identified all lines of output; and a partial grade in between when the student either failed to identify all the lines of output or identified redundant/non-existent output. In addition to the score, the tutor logged the time spent by the student on each problem.

2.2 COMPUTING THE FOUR PARAMETERS

We calculated $P(L_0)$ and $P(T)$ for each concept covered by the tutor, as follows.

$P(L_0)$: We calculated the probability that a student knows a concept before using the tutor as the percentage of all the users of the tutor who had solved the pretest problem on the concept correctly. $P(L_0)$ of 5 of the 12 concepts were 0.90 or greater. But, they corresponded to small numbers of students, suggesting that with additional data, we may be able to calculate better $P(L_0)$ values for those 5 concepts. Given the high values of $P(L_0)$, indicating that selection was an easier topic for the users of the tutor, we used 0.98 instead of the traditional 0.95 as the mastery criterion for the BKT model.

$P(T)$: We calculated the probability that a student learns a concept as the percentage of students who solved the pretest problem incorrectly, solved practice problems and went on to solve the first post-test problem correctly. These were the students who learned the concept by using the tutor.

We computed $P(G)$ and $P(S)$ for each of the practice problems as follows:

$P(G)$: Probability that the student guessed the correct solution to the problem – as the percentage of students who had solved the prior problem on the concept incorrectly or partially, but solved the current problem correctly.

$P(S)$: Probability that the student slipped and solved the problem incorrectly – as the percentage of students who had solved the prior problem on the concept correctly, but solved the current problem incorrectly or partially. For the first practice problem, we estimated this to be 0.01 since students were never presented a practice problem by the adaptive tutor unless they had solved the pretest problem incorrectly.

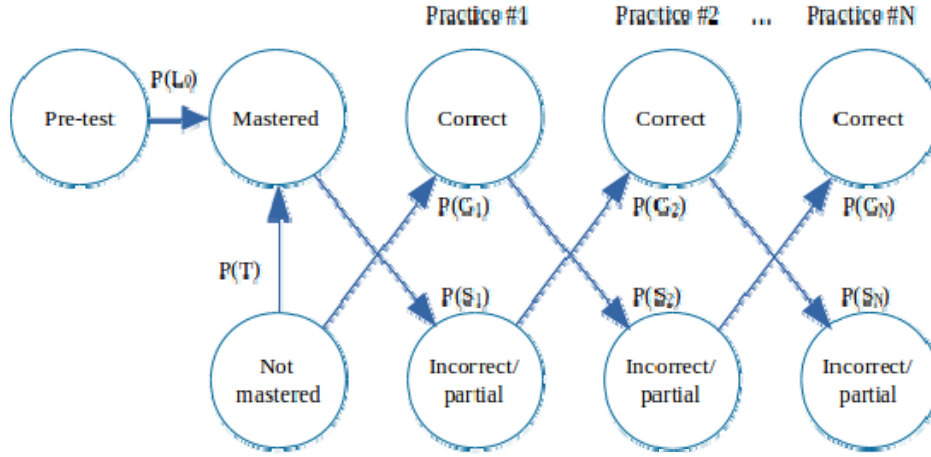


Figure 1: BKT Model with two parameters per concept and two per practice problem

Figure 1 illustrates the BKT model for a concept. As discussed above, BKT uses four main parameters: L_i , T , G , S . Initially, the probability $P(L_i)$ that the student has mastered the current concept at time i is equal to $P(L_0)$. Each problem solved by the student updates the probability that a concept has been mastered using the following equations: (1) and (2) update the probability the concept had been mastered before the current problem, and (3) updates the probability the concept was mastered during the current problem (Baker *et al* 2008):

$$P(L_{i-1}|Correct_i) = \frac{P(L_{i-1}) * (1 - P(S_i))}{P(L_{i-1}) * (1 - P(S_i)) + (1 - P(L_{i-1})) * P(G_i)} \quad (1)$$

$$P(L_{i-1}|Incorrect_i) = \frac{P(L_{i-1}) * P(S_i)}{P(L_{i-1}) * P(S_i) + (1 - P(L_{i-1})) * (1 - P(G_i))} \quad (2)$$

$$P(L_i|Action_i) = P(L_{i-1}|Action_i) + ((1 - P(L_{i-1}|Action_i)) * P(T)) \quad (3)$$

Our approach is similar to empirical probabilities approach (Hawkins *et al.* 2014). But, we used pretest problem to compute $P(L_0)$ instead of the first practice problem. We calculated $P(G)$ and $P(S)$ for each practice problem and used the percentage of students who learned each concept to compute $P(T)$ for the concept. Several attempts have been made to individualize BKT parameters per student with the aim of improving its fit. *Our approach is different in that we have tried to customize performance parameters to the problems solved by the students because no two problems are alike in terms of the provided context or expected answer.*

3 EVALUATING THE BKT MODEL

We used k-fold cross-validation to estimate the performance of our predictive BKT model. Since our data consisted of 3600 records, we chose $k=25$, so that each of the randomly constituted 25 subgroups contained 144 records. Since the tutor covered 12 concepts, each subgroup of 144 records was expected to be a fairly good sample of the overall data. We used each of the 25 groups

to calculate the time and problems saved using the BKT model that was built using the other 24 subgroups. Finally, we computed the mean of the time and practice problems saved per concept across all 25 runs.

Table 1: Results of BKT Evaluation.

Concept	Number of Students who			Mean Time Saved (in Minutes)	Mean Number of Problems Saved	Number of Students
	Saved Time	Made no Difference	Lost Time	Per Student		
1	109	262	27	0.22	0.50	398
2	47	6	10	0.32	0.92	63
3	98	10	10	0.60	1.10	118
4	122	14	0	0.79	1.71	136
5	53	3	0	0.44	1.36	56
6	336	6	0	2.71	2.80	342
7	105	6	0	2.89	3.81	111
8	376	35	21	1.06	1.46	432
9	127	413	64	0.18	0.27	604
10	672	25	92	5.12	3.68	789
11	272	17	20	1.31	1.94	309
12	175	29	38	1.47	1.62	242
	69.22	22.94	7.83	1.425	1.764	
	Total Percentage			Mean		

As a result, we found that in 69.22% of the cases, students would have saved time with the BKT model (Table 1). Students would have saved time/practice problems on some concepts more than others. The concepts on which they would have saved the most time were the harder concepts (6, 7, 10, 11 and 12) – corresponding to nested and multiple `if/if-else` statements. Overall, they could have saved a mean of 1.425 minutes (out of 30 minutes set aside for the tutoring session) and 1.764 problems per student per concept learned using the tutor. In practice, any saved time could be used to learn additional concepts or end the tutoring session earlier.

For this study, we combined data of C++, Java and C# users of the tutor. It is conceivable that the four parameters will be different for the different programming languages. We also combined the data of students from high schools, community colleges and undergraduate institutions. Once again, it is possible that the four parameters will be different for these different levels of students. We combined the records of students with different treatments during practice (self-explanation, optional feedback). We did not consider the relationships among the various concepts, i.e., we treated all 12 concepts as being independent and mutually exclusive. This is a fallible assumption in programming domain. In the future, we plan to consider using a Bayesian network to connect these concepts. Since this study showed that using Bayesian Knowledge Tracing helps improve adaptation by reducing the number of practice problems solved by the learners, we plan to incorporate Bayesian Knowledge Tracing into the tutor to benefit future users of the tutor. Finally, the tutoring suite contains 17 different tutors. We plan to repeat this study for all the tutors using legacy data collected over the last five years.

ACKNOWLEDGMENTS

Partial support for this work was provided by the National Science Foundation under grant DUE-1432190.

REFERENCES

- Baker, RSJ, Corbett, A.T., Aleven, V. (2008). More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, Berlin, Germany. pp. 406-415.
- Beck, J. (2007). Difficulties in inferring student knowledge from observations (and why you should care). *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*. pp. 21-30.
- Beck, J.E., Chang, K-m. (2007). Identifiability: A Fundamental Problem of Student Modeling. *Proceedings of the 11th International Conference on User Modeling (UM 2007)*. pp. 137-146.
- Corbett, A.T., Anderson J.R. (1992). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4. pp. 253-278.
- Hawkins W.J., Heffernan N.T., Baker R.S.J.D. (2014). Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. In: Trausan-Matu S., Boyer K.E., Crosby M., Panourgia K. (eds) *Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science*, vol 8474. Springer. pp. 150-155.
- Kumar, A.N. (2014). A Model for Deploying Software Tutors. *IEEE 6th International Conference on Technology for Education (T4E)*, Amritapuri, India. pp. 3-9.

- Kumar, A.N. (2006a). A Scalable Solution for Adaptive Problem Sequencing and Its Evaluation. Proceedings of The Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 06), Dublin, Ireland. pp. 161-171.
- Kumar, A.N. (2006). Explanation of step-by-step execution as feedback for problems on program analysis, and its generation in model-based problem-solving tutors. Technology, Instruction, Cognition and Learning. (TICL) J. Special Issue on Problem Solving Support in Intelligent Tutoring Systems, 4(1)
- Sweller, J., Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction 2. pp. 59–89.

Sharing and Reusing Data and Analytic Methods with LearnSphere

Kenneth R. Koedinger

Carnegie Mellon University
koedinger@cmu.edu

John Stamper

Carnegie Mellon University
jstamper@cs.cmu.edu

Paulo F. Carvalho

Carnegie Mellon University
pcarvalh@cs.cmu.edu

ABSTRACT: This workshop will explore LearnSphere, an NSF-funded, community-based repository that facilitates sharing of educational data and analytic methods. The workshop organizers will discuss the unique research benefits that LearnSphere affords. In particular, we will focus on Tigris, a workflow tool within LearnSphere that helps researchers share analytic methods and computational models. Authors of accepted workshop papers will integrate their analytic methods or models into LearnSphere’s Tigris in advance of the workshop, and these methods will be made accessible to all workshop attendees. We will learn about these different analytic methods during the workshop and spend hands-on time applying them to a variety of educational datasets available in LearnSphere’s DataShop. Finally, we will discuss the bottlenecks that remain, and brainstorm potential solutions, in openly sharing analytic methods through a central infrastructure like LearnSphere. Our ultimate goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers in order to advance the learning sciences as harnessing and sharing big data has done for other fields.

Keywords: Learning metrics; data storage and sharing; data-informed learning theories; modeling; data-informed efforts; scalability.

1 WORKSHOP BACKGROUND

The use of data to improve student learning has become more effective as student learning activities and student progress through educational technologies are increasingly being tracked and stored. There is a large variety in the kinds, density, and volume of such data and to the analytic and adaptive learning methods that take advantage of it. Data can range from simple (e.g., clicks on menu items or structured symbolic expressions) to complex and harder-to-interpret (e.g., free-form essays, discussion board dialogues, or affect sensor information). Another dimension of variation is the time scale in which observations of student behavior occur: click actions are observed within seconds in fluency-oriented math games or in vocabulary practice, problem-solving steps are observed every 20 seconds or so in modeling tool interfaces (e.g., spreadsheets, graphers, computer algebra) in intelligent tutoring systems for math and science, answers to comprehension-monitoring questions are given and learning resource choices are made every 15 minutes or so in massive open online courses (MOOCs), lesson completion is observed across days in learning management systems, chapter/unit test results are collected after weeks, end-of-course completion and exam scores are collected after many months, degree completion occurs across years, and long-term human goals like landing a job and achieving a good income occur across lifetimes. Different paradigms of data-driven

education research differ both in the types of data they tend to use and in the time scale in which that data is collected. In fact, relative isolation within disciplinary silos is fostered and fed by differences in the types and time scale of data used (cf., Koedinger et al., 2012).

Thus, there is a broad need for an overarching data infrastructure to not only support sharing and use within the student data (e.g., clickstream, MOOC, discourse, affect) but to also support investigations that bridge across them. This will enable the research community to understand how and when long-term learning outcomes emerge as a causal consequence of real-time student interactions within the complex set of instructional options available (cf., Koedinger et al., 2010). Such an infrastructure will support novel, transformative, and multidisciplinary approaches to the use of data to create actionable knowledge to improve learning environments for STEM and other areas in the medium term and will revolutionize learning in the longer term.

LearnSphere transforms scientific discovery and innovation in education through a scalable data infrastructure designed to enable educators, learning scientists, and researchers to easily collaborate over shared data using the latest tools and technologies. LearnSphere.org provides a hub that integrates across existing data silos implemented at different universities, including educational technology “click stream” data in CMU’s DataShop (Stamper et al., 2011), massive online course data in Stanford’s DataStage and analytics in MIT’s MOOCdb (Veeramachaneni et al., 2014), and educational language and discourse data in CMU’s new DiscourseDB (Jo et al., 2016). LearnSphere integrates these DIBBs in two key ways: 1) with a web-based portal that points to these and other learning analytic resources and 2) with a web-based workflow authoring and sharing tool called Tigris. A major goal is to make it easier for researchers, course developers, and instructors to engage in learning analytics and educational data mining without programming skills.

The main goal of this workshop is to provide attendees with hands-on experience using Tigris for learning analytics. We hope that this year we will be able to attract attendees that have been exposed to LearnSphere from these past events, although we will have some tutorial activities included for new attendees as well. This workshop builds off a successful LAK 2018 Tutorial, and workshop at AIED/EDM 2017.

2 ORGANIZATIONAL DETAILS

2.1 Type of event

Workshop

2.2 Proposed Schedule and Duration

Table 1: Proposed Half-Day Schedule.

Time	Item
1:30p	Introductions
2:00p	Tigris workflow tool (Lecture & Demos)
2:30p	Hands-on I: Build custom analysis workflows using existing Tigris components

3:30p	Coffee Break
4:00p	Hands-on II: Breakout sessions (upload your own data; create workflow components)
4:45p	5-minute participant talks about proposed or created workflows
5:15p	Closing/High Level Discussion

2.3 Type of Participation

Mixed participation will be through submission of reviewed abstracts, invited guests, and open registration. For participants who have accepted abstracts or are invited by the workshop committee, we have allocated approximately \$20,000 from our grant funding to cover registration and travel costs.

2.4 Activities

Activities will include presentations from workshop organizers, invited guests, and short presentations from accepted abstract presenters. Hands on sessions will include demos and group work towards implementing analytics.

2.5 Expected Numbers

We expect 15-20 participants based on previous workshops.

2.6 Activities to Recruit Attendees

We will create a website to announce the workshop and method of submitting abstracts. The Learning Analytics, Educational Data Mining, and LearnLab mailing lists will be used to direct potential attendees to the workshop website. In addition, we will invite a number of invited guests. Both accepted submissions and invited guests will have the chance to receive funding to attend.

2.7 Required Equipment

Projector and screen will be required by organizers. Attendees will need to bring laptops and will need adequate internet connectivity.

3 OBJECTIVES AND OUTCOMES

Broadly, this workshop offers those in the Learning Analytics community an exposure to LearnSphere as a community-based infrastructure for educational data and analysis tools. In opening lectures, the organizers will discuss the way LearnSphere connects data silos across universities and its unique capabilities for sharing data, models, analysis workflows, and visualizations while maintaining confidentiality.

More specifically, we propose to focus on attracting, integrating, and discussing researcher contributions to Tigris, the web-based workflow authoring and sharing tool. Workshop submissions in the form of abstracts will involve a brief description of an analysis pipeline relevant to modeling

educational data as well as accompanying code. Prior to the workshop itself, the organizers will coordinate with authors of accepted submissions to integrate their code into Tigris. A significant portion of the workshop will be dedicated to hands-on exploration of custom workflows and workflow modules within Tigris. Authors of accepted submissions will present their analysis pipelines, and everyone attending the workshop will be able to access those analysis pipelines within Tigris to a variety of freely available educational datasets available from LearnSphere. The goal is to generate -- for each workflow component contribution in the workshop -- a publishable workshop paper that describes the outcomes of openly sharing the analysis with the research community.

Finally, workshop attendees will discuss bottlenecks that remain toward our goal of a unified repository. We will also brainstorm possible solutions. Our goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers we can advance the learning sciences as harnessing and sharing big data has done for other fields.

4 REFERENCES

- Jo, Y., Tomar, G., Ferschke, O., Rosé, C. P., & Gašević, D. (2016, April). Pipeline for expediting learning analytics and student support from data in social learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 542-543). ACM.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.
- Stamper, J., Koedinger, K.R., Baker, R., Skogsholm, A., Leber, B., Demi, S., Yu, S., Spencer, D. (2011) Managing the Educational Dataset Lifecycle with DataShop. In Kay, J., Bull, S. and Biswas, G. (eds). *Proceeding of the 15th International Conference on Artificial Intelligence in Education* (AIED2011).
- Veeramachaneni, K., Halawa, S., Dernoncourt, F., O'Reilly, U. M., Taylor, C., & Do, C. (2014). Moocdb: Developing standards and systems to support MOOC data science. *arXiv preprint*. arXiv:1406.2015.

Using Curriculum Pacing in LearnSphere to Visualize Student Learning Trajectories

Nirmal Patel
Playpower Labs
India

nirmal@playpowerlabs.com

Dhaval Prajapati
Playpower Labs
India

Saumya Mehta
Playpower Labs
India

Parth Agrawal
Playpower Labs
India

Derek Lomas
Delft Institute of Technology
Netherlands
j.d.lomas@tudelft.nl

ABSTRACT

We propose to build a Curriculum Pacing workflow component in the LearnSphere environment. Curriculum Pacing is a way to visualize student learning trajectories through curriculum data. It is a visual learning analytic method that allows its users to observe how students interacted with curriculum topics over time, which modules of the curriculum were visited by students over and over, and when in time students interacted with previously seen content. The pacing visualization is useful for data-driven decision making for multiple stakeholders in education. EDM researchers can use pacing plots to build hypotheses about student learning behavior. Instructors and curriculum coordinators can ensure that their students are moving at an expected pace and identify content areas that are being difficult for their students, instructional designers can look at how students are moving through the curriculum and compare it to their expectations, and potentially, data scientists and machine learning engineers can see if there is enough variation in data to drive content recommendation algorithms.

Keywords

Sequence Visualization, Learning Analytics, Curriculum Pacing

1. INTRODUCTION

Learning analytics researchers are increasing their use of temporal student data to understand what patterns of student behavior are correlated with desirable outcomes. Analysis of student learning processes is becoming easier by using tools such as frequent pattern mining [1], process mining [1], and very recently, a new method called Curriculum Pacing [2]. There are many challenges when it comes to understanding student learning trajectories, because of the combinatorial explosion of the possible learning sequences in simple settings. For example, if an intelligent tutor allows for 10 possible student actions, and if students can take up to 100 different actions, this permits 100^{10} different possible student learning trajectories. Although in practice, we find that only a small fraction of these possibilities occur. Even then, approaches such as sequence clustering have to be used to aggregate similar student behavior [2, 3]. In a nutshell, temporal student data are complex and difficult to make sense of.

Data visualization is one of the most widely used ways of making sense of complex datasets. There are many reasons behind visualizing data, but the most prominent reason which is often cited is that the summary statistics of data can easily hide the actual structure of the data [4]. Graph-based visualizations are one of the easiest methods to see structure in temporal data, but these visualizations often become complex and hard to

interpret [3]. When it comes to educational data, interpretability of data is as important as data's ability to predict outcomes because we need to know what makes difference for student outcomes, not just an accurate prediction of them. So, if we are visualizing complex educational datasets, we want to be able to make sense of the data visualizations.

Curriculum Pacing visualization allows us to visualize student trajectories through a curriculum in an interpretable way. We get to see what students are doing over time, and these activities tie to different parts of the curriculum. We can see data of many students in the same visualization, without highly increasing the visual complexity. Instructors can see how students are moving through their curriculum, whether any parts of their course are being difficult for students, and when students are revisiting specific content areas. Using the pacing visualization, instructional designers can see whether student movement through the curriculum matches their expectations. Education administrators can also use these visuals to identify classrooms that are lagging behind others, and offer help. Last but not least, if data scientists are using student trajectory data to drive recommendation algorithms, they can see whether the data have the desired variability and properties to give meaningful recommendations.

2. WORKFLOW METHOD

2.1 Data Inputs

Column Name	Description
Anon Student Id	Anonymous ID of the student
Problem/Step Start Time	The Start time of the problem or step (depends on the type of the data)
Problem Hierarchy	The location in the curriculum hierarchy where this problem occurs

Table 1: Data columns of DataShop student-step or student-problem [5] data used for Curriculum Pacing workflow component.

The Curriculum Pacing visualization workflow component in LearnSphere will take DataShop student-step or student-problem level datasets as input. Only a handful of columns will be used from these datasets to produce the pacing visualization.

Apart from the standard DataShop columns, the workflow component will also take a few input parameters to customize the visualization.

Parameter	Description
Problem Hierarchy Order (optional)	A CSV file with two columns that assigns each Problem Hierarchy value an integer that locates the Problem Hierarchy value in the curriculum. If not provided, Problem Hierarchy column will be sorted alphabetically using the <code>gtools::mixedsort()</code> function in R. In other words, this input defines the ordinal or factor levels of the Problem Hierarchy column from the DataShop student-step or student-problem data.
Time Scale Type	Relative or Absolute. Relative time normalizes Problem/Step Start Time 1, and absolute time preserves the actual timestamps of student interactions.
Time Scale Resolution	Hour, Day, Week, or Month. Student data will be aggregated at the level of the provided resolution.
Minimum Time Unit	An integer or a timestamp in YYYY-MM-DD HH:MM:SS format. If Time Scale Type is Relative, then the component will remove the student data before the given normalize integer time unit. If the Time Scale Type is Absolute, then the component will remove the student data before the given timestamp.
Maximum Time Unit	Similar to the Minimum Time Unit
Plot Type	Usage (Number of Students) – plots student usage over time Usage and Performance (Number of Students and Percent Correct) – plots student usage and performance over time

Table 2: Parameters besides the primary input file for the Curriculum Pacing workflow component.

Using these inputs, the workflow component program will generate the necessary output.

2.2 Workflow Model

Curriculum Pacing is a visual learning analytic method so it operates mainly by transforming the input data into a certain format and producing a data visualization out of them.

The visualization will be produced as a 2D plot with an X and a Y-axis. The X-axis will represent time and the Y-axis will represent the position in the curriculum. Input data of all of the students will be aggregated to produce the output.

The X-axis will be a continuous axis and will represent either relative or absolute time. Relative time will be in the units as defined by the Time Scale Resolution parameter. For example, if the Time Scale Type is 'Relative' and the Time Scale Resolution is 'Week,' then the values 1, 2, 3 etc. on X-axis will represent the 1st week of student usage, 2nd week of student usage etc. Absolute time will be binned by the units as defined by the Time Scale Resolution parameter. For example, if the Time Scale Type is 'Absolute' and the Time Scale Resolution is 'Week,' every Problem/Step Start Time will be changed to the preceding Monday. Similarly, if the Time Scale Resolution parameter is set to 'Month,' every Problem/Step Start Time will be changed to the 1st of the month. The range of the X-axis will be limited from the Minimum Time Unit to the Maximum Time Unit.

The Y axis will be an ordered discrete axis (or an ordinal axis) and will contain Problem Hierarchy. This will represent where the student is in a curriculum at a given point in time (which can be relative or absolute.) By default, Y-axis will be sorted alphabetically using the `gtools::mixedsort()` function in R. If the users desire a different order, they will be able to modify the order of Y-axis values by providing an optional input parameter 'Problem Hierarchy Order.'

If the Plot Type is set to 'Usage (Number of Students),' the plot will be produced as a 2D heatmap. Each cell of the heatmap will be filled with the hue representing the number of students at a given point in time and a position in the curriculum. If the Plot Type is set to 'Usage and Performance (Number of Students and Percent Correct),' the plot will be produced as a 2D scatterplot with the size of the dots representing the number of students and color of the dots representing the average percent correct across all of the problems at a given point in time and a position in the curriculum.

2.3 Workflow Outputs

The workflow will output a single data visualization combining data of all of the students in the input data, in an SVG format. Besides this, a raw data file that produced the data visualization will also be exported. Figure 1 shows two examples of the visualization output, one for each of the possible Plot Type parameter options.

3. DISCUSSION

Looking at the example shown above, we can infer multiple things about the students and the course. First of all, we can see that one big group of students started at beginning of the curriculum, and went through the course material as time went on. This can be inferred from the near 45-degree diagonal band in the plots. We can also see that there is another band that starts halfway of the Y-axis, which shows that data for a group of students starts from the middle of the curriculum. Within this small group, we can also see that a subset of students went ahead to complete the course faster than other students. This can be inferred from the vertical band that branches out from the upper diagonal band. Other important features of pacing plots are vertical and horizontal lines. Vertical lines typically indicate parts of the curriculum that students interact with in a short timespan, and horizontal lines usually show parts of the curriculum that students repeatedly interact with as time goes on. Although, the horizontal lines can also appear if students are following different time schedules while going through the curriculum.

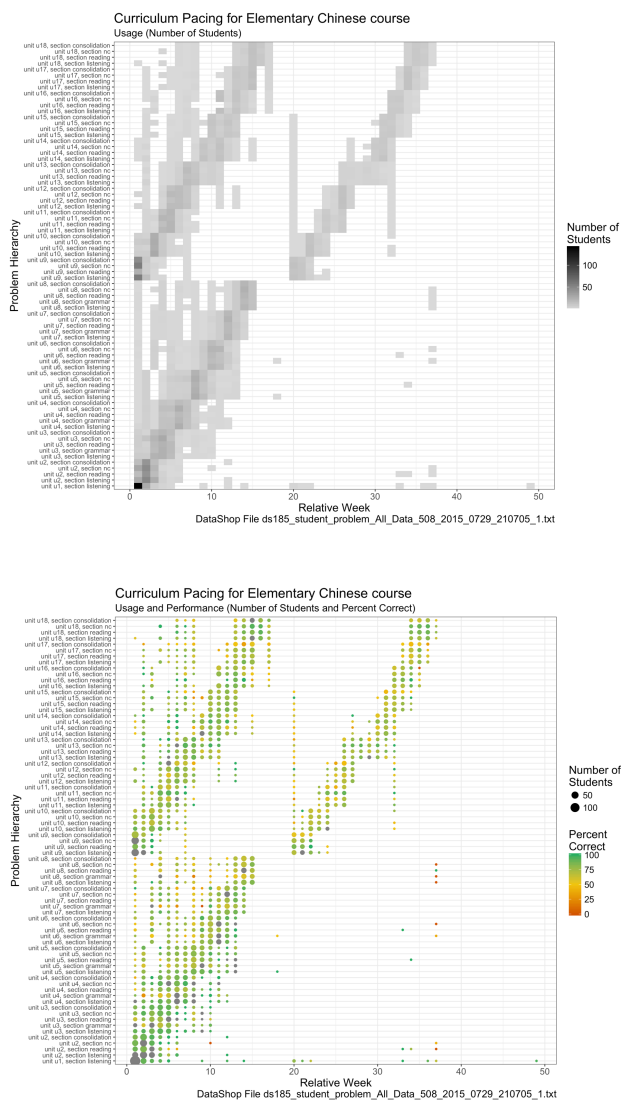


Figure 1: Examples of Curriculum Pacing plots. The plots are using DataShop Elementary Chinese course data from 212 anonymized students. The first plot shows student trajectories through the course units over time, while the second plot shows both the trajectories and average student performance at different time points. Both plots are on a relative week time scale.

The second plot has one marked difference, it shows student performance using a color scale. The performance is measured using average percent correct of all of the students and all of the outcomes (except hints) at a point in curriculum and a point in time. The green color shows 100% correct, yellow 50%, and red 0%. Using these colors as cues, we can find topics that students found difficult (red dots.)

There are multiple goals that can be achieved using curriculum pacing visualization:

- EDM researchers can use pacing plots to build hypotheses about why students might be going through the curriculum in a certain way versus the other.
- An instructor can easily find out whether all of the students are moving at an expected pace or not. To identify students who are not able to follow the schedule, we can also make the visualization interactive so that by hovering over different parts of the visualization, we can know the related students.
- An instructional designer can compare the student learning trajectory to expected trajectory, and find out whether there is a subset of student that is following a different learning trajectory and see how it can be better supported.
- An instructional designer can look for difficult topics, and see if there are frequent visits to previous topics for the difficult topics. If a lot of the students are revisiting similar previous topics, the instructor can find out by talking with them whether reviewing previously seen content was helpful for the students to understand the difficult topic.
- A data scientist can see if there are any topics where students might be applying different learning strategies such as spaced practice, mass practice, revisiting specific topics after a difficult topic, and whether there is enough variation in the data to model successful student learning strategies.

Curriculum Pacing visualization can be used in many different ways and can act as a starting point of further inquiry in student learning.

4. REFERENCES

- [1] Romero, C., Ventura, S., Pechenizkiy, M. and Baker, R.S. eds., 2010. Handbook of educational data mining. CRC press.
- [2] Patel, N., Sharma, A., Sellman, C. and Lomas, D., 2018, June. Curriculum Pacing: A New Approach to Discover Instructional Practices in Classrooms. In International Conference on Intelligent Tutoring Systems (pp. 345-351). Springer, Cham.
- [3] Patel, N., Sellman, C., Lomas, D., 2017, July. Mining frequent learning pathways from a large educational dataset. In Proceedings of the 3rd International Workshop on Graph Educational Data Mining (pp. 27–30).
- [4] Anscombe, F. J., 1973. Graphs in Statistical Analysis. American Statistician 27 (1) (pp. 17–21).
- [5] PSLC DataShop Documentation. <https://pslcdatashop.web.cmu.edu/help?page=export>

Learning Curves Segmented by Mastery

Stephen E. Fancsali, Michael Sandbothe, Steven Ritter

Carnegie Learning, Inc.

{sfancsali, msandbothe, sritter}@carnegielearning.com

ABSTRACT: We describe a modest but important modification to the visualization of learning curves as a means by which to judge the quality of knowledge component models in learning analytic frameworks like those provided by LearnSphere’s DataShop and its Tigris workflow tool. The modification centers on the idea that aggregate learning curves are often less informative than visualizations that provide learning curves for segments of students based upon whether (and when) mastery of the knowledge component in question is achieved by the student. We describe the proposed modified LearnSphere Tigris workflow and provide an example of its output.

Keywords: cognitive modeling, skill modeling, knowledge components, learning curves, mastery learning, visualization

1 INTRODUCTION

Learning technologies like Carnegie Learning’s MATHia, based on its Cognitive Tutor technology (Ritter, Anderson, Koedinger, & Corbett, 2007) and Anderson’s ACT-R cognitive architecture (Anderson & Lebiere, 1998), take as fundamental the atomization of a target domain into fine-grained, discrete knowledge components (KCs) or skills. Substantial literature in educational data science is devoted to the data-driven improvement of cognitive models of domains, comprised of such KCs, based on empirical “learning curve” visualizations of student mastery (or acquisition) of such KCs over time (e.g., in LearnSphere’s DataShop; Koedinger et al., 2011). The most common approach to learning curve analysis plots either aggregate error or correctness rate over opportunities at which students could demonstrate knowledge or mastery of a KC within a learning system. In rough sketch, improvements to cognitive/KC models follow the idea that learning curves should, over time, reflect learning by “smoothly” (i.e., roughly monotonically) increasing (if correctness is plotted) or decreasing (if error rate is plotted). To the extent that learning curves deviate from such a pattern (e.g., “flat” learning curves, curves with a “saw tooth” pattern of increases and decreases, etc.) empirical evidence would suggest the potential need for modifications to the KC model (e.g., that a KC might be split into two KCs in a refined model) to better model student learning of the domain.

The present proposal builds on work first presented by our colleagues several years ago (Murray et al., 2013; Nixon, Fancsali, & Ritter, 2013), which addressed a problem for the standard approach to learning curve analysis. Learning curves, according to the standard approach, are “aggregate” in the sense that each plotted point provides the proportion or percentage correct (or in error) for all students in a dataset at each opportunity. However, in any learning platform with a mastery learning regime (Bloom, 1968), student attrition occurs in the sense that students drop out of the dataset as they master the particular KC under consideration (i.e., the sample size for each plotted opportunity

decreases over time). Murray, et al. (2013) suggest that “segmenting” learning curves into “bins” of students who reach mastery at the same opportunity provides a means by which some learning curves that, in the aggregate, do not “show” learning may in fact reveal learning that is obscured by aggregating over large samples of students with attrition.

Consider the KC “Calculate intercept using general linear form” in Carnegie Learning’s MATHia workspace titled “Graphing Linear Equations Using a Given Method.” A screenshot of problem solving in MATHia involving this KC is provided as Figure 1.¹ In this problem, the student is provided two opportunities to practice this KC, one in which they calculate the x-intercept and one in which they calculate the y-intercept. After calculating these intercepts, students work with a graph to plot the line described by the equation. Learning curve analysis of this KC can, for example, be used to help determine whether or not KCs should be specified that map to the calculation of the x-intercept using general linear form separately from the calculation of the y-intercept using general linear form.

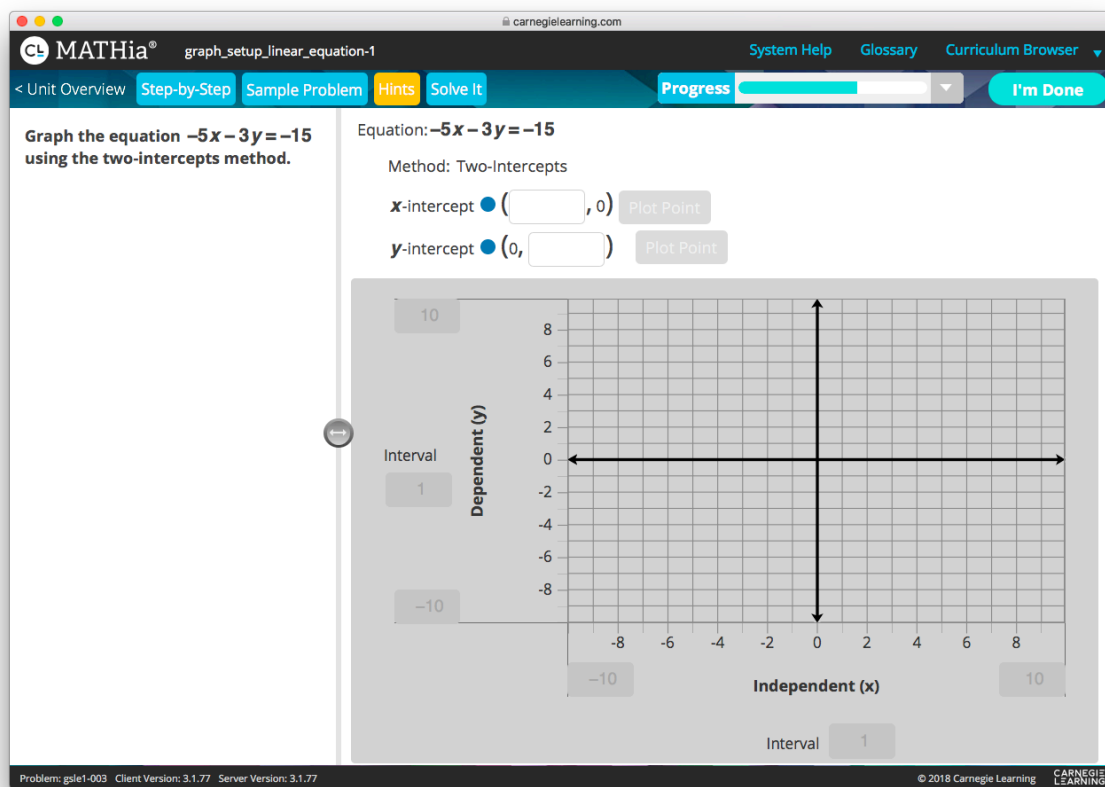


Figure 1: Screenshot of MATHia problem solving in the “Graphing Linear Equations Using a Given Method” workspace. This problem provides practice on the KC “Calculate intercept using general linear form.”

¹ We consider this example and its implications in more detail in a companion piece in another workshop at this conference (Ritter, Fancsali, & Sandbothe, 2019).

The learning curve for this KC is illustrated in Figure 2. Aggregated over 14,646 students, this learning curve, on visual inspection appears to show “no learning” and would likely be categorized as such by the existing learning curve categorization tool available within the Learning Curves visualization component of the Tigris workflow tool. We posit that this KC is a target for a cognitive model improvement, but not because this learning curve is “flat” and purportedly shows no learning. Rather, a majority of students master this KC in a reasonable number of opportunities, but a substantial minority of students struggle to master this KC.

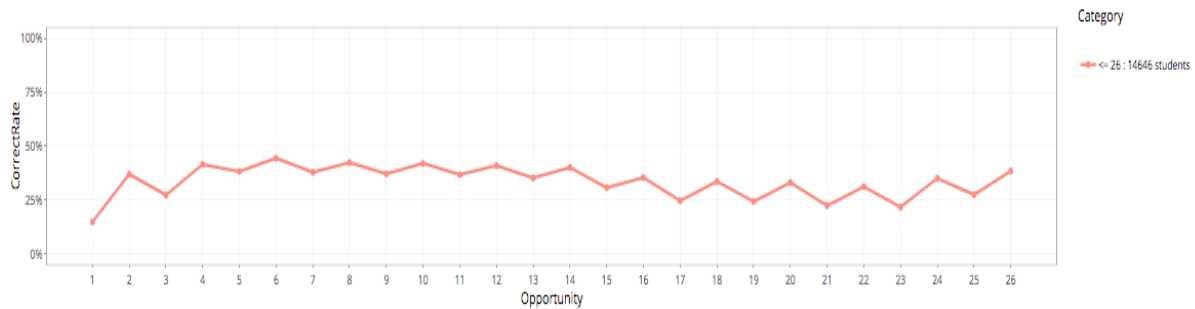


Figure 2: Aggregate learning curve for “Calculate intercept using general linear form” (n = 14,646; with attrition over time/opportunities)

Evidence of this struggle is illustrated by the type of “segmented” learning curves proposed by Murray et al. (2013). The particular shape of segments of these more nuanced learning curves, specifically the saw tooth pattern of the lowest performing three segments, may suggest instructional design improvements and avenues for further research (Figure 3; also see Ritter, Fancsali, & Sandbothe, 2019). Figure 3 also provides an example of the target output for this proposed workflow.

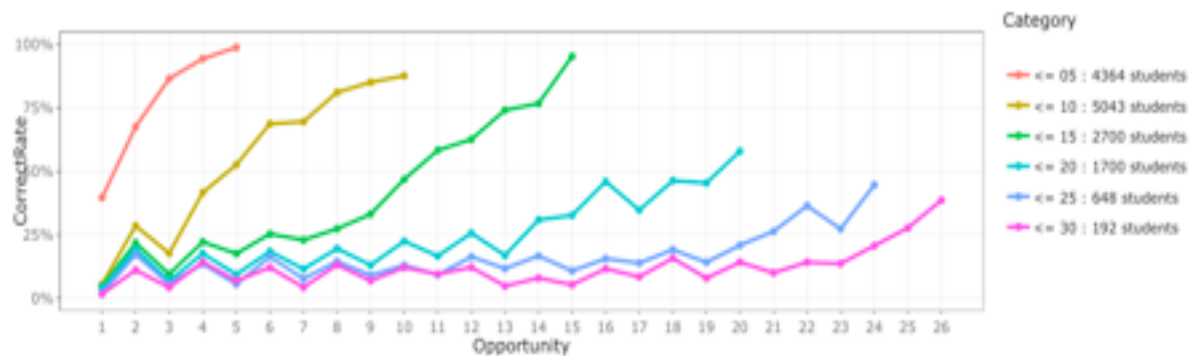


Figure 3: Learning curve for “Calculate intercept using general linear form” segmented by mastery. Each curve represents a cluster of students based on the practice opportunity at which they were judged by MATHia to have reached mastery (e.g., those that master in 1-5 opportunities [Category: ≤ 05 ; n = 4,364], 6-10 opportunities [Category: ≤ 10 ; n = 5,043], etc.)

2 WORKFLOW METHOD

Our proposal could be implemented within the Tigris workflow tool as a modification to the existing Learning Curves Visualization component. Alternatively, it could be implemented as a new

component. As such, the overall workflow is similar to that of generating learning curves in Tigris in its current state.

2.1 Data Inputs

The proposed, modified (or new) Learning Curves Visualization component will take as input a standard PSLC DataShop student-step rollup dataset and rely on the same columns/features in the student step rollup presently used by the Learning Curves Visualization component. The current Learning Curves Visualization component takes as input a set of parameters for learning curve classification. The new/modified component would take optional information about a criterion upon which to segment learning curves based on the opportunity at which students mastered KCs. In the absence of such a criterion, the last practice opportunity could be used to segment learners (whether or not mastery was achieved by the last practice opportunity). Additional parameters for the new/modified component will relate to how learning curve segmentation should be configured (e.g., the number of segments into which students in the dataset should be divided, whether those students who master a KC should be segmented separately from students who fail to master a KC, etc.). In the discussion, we consider how existing parameters for learning curve categorization may be helpful for analyzing segmented learning curves.

There are at least three means by which information about how to adjudicate student mastery of each KC encountered can be made available to the Segmented Learning Curve Visualization component:

- 1) as a feature that is added to the student step rollup indicating the opportunity on which a system judged a student to have mastered a KC,
- 2) as an inference using the Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1995) parameters produced by the BKT Analysis component and the “Predicted Error Rate” in its student-step rollup output (from which a learner’s probability of KC mastery can be inferred), relying on a mastery criterion threshold included among the parameters for the Segmented Learning Curve component, or
- 3) by implementing a new Mastery Determination analysis module in Tigris that would take a student step rollup as input (whether directly from input, from the BKT analysis module, etc.) and infer a mapping from each student-KC pair to the opportunity at which a learning system or statistical model judged student mastery according to some configurable criterion (or a null value if mastery is never judged as being achieved). This would allow for a more flexible and configurable approach to mastery determination (i.e., not relying on an arbitrary assignment as in the first option or on traditional BKT criteria as in the second option).

2.2 Workflow Model

The rendering of segmented learning curves is straight forward given the inputs described in the previous section. Based on the configuration options (e.g., number of segments) and information about the opportunity at which (and whether) students mastered KCs, multiple learning curves on the same plot are rendered in the same manner as currently carried out by the Learning Curve Visualization module. If desired, power law and other fitted curves could also be learned for each of

the learning curves so segmented, though this is an area as yet unexplored by the authors. The rendering of these types of segmented learning curves is implemented in R as part of a learning engineering workbench developed for internal use by Carnegie Learning; this workbench produced images that are Figures 2 and 3. It should be relatively straightforward to implement similar code within the Tigris framework.

2.3 Workflow Outputs

The output of this workflow modification will be nearly identical to the output of the existing DataShop Learning Curves Visualization module, save for differences due to the proposed modification. PNG or similar image files will be rendered for each KC's segmented learning curve(s) with modified legends to represent the "binning" of students, and corresponding XML files representing these learning curves will also be produced, with appropriate properties to represent the additional information required to interpret learning curves segmented by mastery.

3 DISCUSSION

Learning curves segmented by mastery provide visualizations for analysis of whether sub-populations of students perform in importantly different ways as they practice KCs in different contexts. The aggregate learning curve of Figure 2 may be roughly classified as displaying "no learning" or as manifesting a modest "saw tooth" pattern in which correctness rates go up and down at alternating opportunities to practice the KC. The latter pattern is much more apparent in Figure 3 for the set of 2,540 students who struggle and do not master this KC (i.e., the bottom three curves that do not approach 100% correctness) while the majority of students go on to master the KC over time (with generally "smoothly" increasing learning curves).

Especially with larger data sets, considering these sub-populations may provide important insight into how more advanced students may be well-served by a cognitive model which keeps the present KC model intact while struggling students may best benefit from a skill model that "splits" this particular KC into two (e.g., in this case, splitting x-intercept calculations from y-intercept calculations) (see Ritter, Fancsali, & Sandbothe, 2019). Applied to sub-populations via these segmented learning curves, the "classification" affordances presently provided by the DataShop Learning Curve Visualization module may prove even more valuable. In any event, considering sub-populations in datasets with thousands of students with respect to initial knowledge, learning rate, and potential cognitive/instructional model improvements will provide a fruitful area for future research.

REFERENCES

- Anderson, J.R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bloom, B.S. (1968). Learning for mastery. *Evaluation Comment* 1(2). Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs.
- Corbett, A.T. & Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User-Modeling and User-Adapted Interaction* 4, 253-278.

- Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2011). A data repository for the EDM community: The PSLC datashop. In S. Ventura, C. Romero, M. Pechenizkiy, & R.S.J.d. Baker (Eds.). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.
- Murray, R.C. et al. (2013). Revealing the learning in learning curves. In H.C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.). *Artificial intelligence in education (AIED 2013)*. (LNCS Vol. 7926, pp. 473-482). Berlin: Springer.
- Nixon, T., Fancsali, S.E., & Ritter, S. (2013). The complex dynamics of aggregate learning curves. In S.K. D'Mello, R.A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*. (pp. 338-339).
- Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. (2007). Cognitive Tutor: applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249-255.
- Ritter, S., Fancsali, S.E., & Sandbothe, M. (2019). Conceptual change as evidence of learning. In *Companion Proceedings 9th International Conference on Learning Analytics & Knowledge (LAK19)*.

Human-Centered Data Science for Educational Technology Improvement using Crowd Workers

Steven Moore¹, John Stamper²

Carnegie Mellon University, HCII

StevenJamesMoore@gmail.com¹, jstamper@cs.cmu.edu²

Soniya Gadgil

Eberly Center for Teaching Excellence and Educational Innovation

soniyag@andrew.cmu.edu

ABSTRACT: Learning curves (LC) provide a concise way to visualize student learning over time. Analysis of these curves can identify which knowledge components (KC) might be misaligned or at the very least where a problem in the system exists. While beneficial to system and course improvement, this analysis is time consuming and can be taxing when hundreds of KCs are present. Utilizing crowd workers, LCs can be mapped to categories and rank ordered, indicating which need improvement the most. Leveraging the categorization and rankings from these workers, a finer grained grouping can be achieved that indicates which LCs need attention first and foremost. This creates a more efficient analysis, helps to maintain the iterative cycle of system and course improvement, and provides another step towards leveraging crowdsourcing for educational improvement.

Keywords: Crowdsourcing, Visual Analytics, Data Analytics, E-learning

1 INTRODUCTION

The proliferation of data on students interacting with online learning environments has enabled new opportunities for understanding student performance in recent years (Baker & Inventado, 2014). It enables the construction of models on how students progress through the learning process and assists in identifying the gaps in their knowledge. Building on these student models for the purpose of tracking student learning over time has been a key area of focus in the educational technology community for many years as well (Murray, 2003). Cognitive Tutors, such as those from Carnegie Learning, utilize student models and are adaptive to student knowledge by tracking the mastery of skills or knowledge components (KCs) (Fanscali et al., 2013). The models that map KCs are generally created with the help of subject matter experts and cognitive scientists. Unfortunately, these knowledge component models (KCMs) do not always correctly model skills, which can impede student learning. When a KCM for a cognitive tutor is incorrectly modeled, it can cause incorrect problem selection and waste valuable student time on skills they have already mastered.

Learning analytics can address this problem and presents an opportunity for continuous improvement of the models using data driven techniques (Stamper & Koedinger, 2011). At present, DataShop (Stamper et. al, 2010) has user interface affordances that utilize a new framework for learning curve (LC) categorization to assist in identifying areas of improvements in the student models of the educational technology. The analysis of these LCs to provide insights into student

models has been around for many years (Anderson, Conrad, & Corbett, 1989). In addition to using these curves to improve student models, the algorithmic use of fitting learning curves has been used to improve upon cognitive models used in intelligent tutoring systems (Cen et al., 2006). While this categorization can assist in identifying which KCs might be misaligned or incorrect in the KCM, the process is still time consuming.

The use of crowd workers is common with educational technology, but often in a way that leverages the workers or users specific content knowledge (Anderson, 2011; Weld et al., 2012). Recently, crowdsourcing has become increasingly popular for content development in the educational domain (Porcello & Hsi, 2013; Paulin & Haythornthwaite, 2016). We propose a workflow that takes a slightly different approach, utilizing crowd workers in a way that does not necessarily leverage their domain expertise or have them develop content in anyway, while still benefiting from their input. This proposed workflow will leverage crowd workers to help with a time consuming and often tedious part of LC analysis that is necessary for course and educational system improvements. We look to utilize crowd workers in order to both better categorize and to provide a priority-ordered ranking of the learning curves for a given dataset, so that the largest improvements can be made in the quickest time frame.

For this proposed workflow, crowd workers from Amazon’s Mechanical Turk, known as turkers, will be recruited to review a set of learning curves. They will select which category each LC fits under and assign it a unique rank order, based on how much it needs to be improved. This ranking of the LCs will be made available to the workflow user, providing a priority view of which LCs to focus their limited time on. It will extend the categorization currently offered by DataShop, utilizing the LC images from the learning curve visualization component in LearnSphere.

Ultimately this workflow looks to be a first step in getting more towards the human-in-the-loop aspect for LearnSphere and leverage crowd workers for work in the learning sciences. We want to leverage the human judgement ability and classification to build upon the existing classification of LCs by DataShop and to make the analysis portion more efficient. This will ultimately assist in the continuous iterative improvement cycle needed in many educational systems.

2 WORKFLOW METHOD

2.1 Data Inputs

The input into this workflow comes from the learning curve visualization component. This LC component outputs a series of Portable Network Graphic (png) images that correspond to the LCs for each present knowledge component in the initial dataset. These images make up the file output of the LC component, which is the primary input into our proposed workflow. The current output size and file names for the images are appropriate for the workflow’s needs. While the image file sizes are small, in order to keep the bandwidth and latency low, we suggest compressing the resulting file in a ZIP file to be used with our workflow. Our workflow can then unzip the images and use them for the model, however as it stands making use of the currently output file from the LC component is also functional.

These LC images are already anonymized regarding their content area, as the images are all titled “KC” followed by an incremental number, as shown in Figure 1. The png files are also similarly named, which assists with confidentiality as well as mapping the image to the corresponding KC that is used in the LC visualization component and any prior analysis ones. Additionally, the images have

the toggleable option to include their DataShop curve categorization next to the curve's title. While we suggest leaving this off, the proposed workflow could include it depending on what the user ultimately wanted to achieve or how they expect it to bias the crowd workers, if at all.

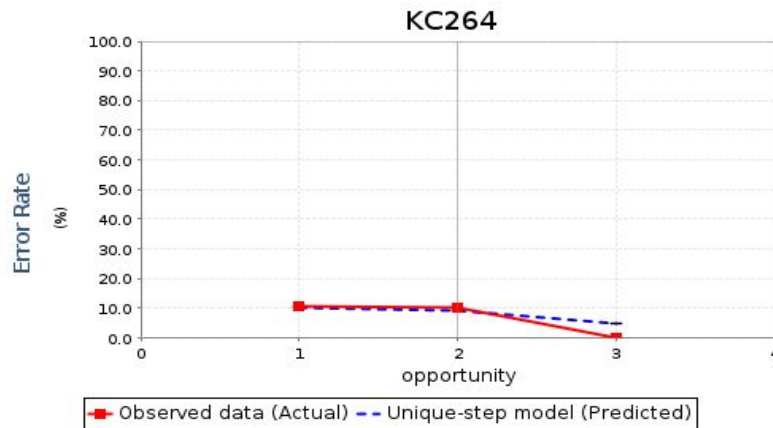


Figure 1: Example learning curve, with the assigned KC title

A second and optional input for the workflow is the XML output from the LC visualization component. This data can be used to provide additional information in the table columns that this workflow outputs. This accompanying data, along with the classification and rankings of the LCs, can provide a high-level view of how well the KCs are mapped via a concise tabular view. It is also trivial to map this XML data back to the corresponding LC, as the KC name and number joins the two. As a first pass and minimally viable workflow, the file name for the images contains enough information to construct the appropriate output for this model without this secondary input.

2.2 Workflow Model

Once the image files and optional corresponding XML data have been input into this model, the images need to be grouped for their presentation to the crowd workers. The workflow will have a configurable input detailing the size of these groups, which corresponds to how many LCs a turker will be reviewing. By default, we suggest a value of ten, as it requires a low amount of time and lends itself to having a commonly quantifiable ranking scale, in this case ranking from 1-10. The second configurable option offered is the grouping of LCs by categories, as labeled in the learning curve visualization component. With this option, enabled by default, the component will select LCs from the same category to present to the crowd workers when possible. For instance, when enabled ten LCs from the too little data category may be selected. If there are not enough for a given category, the component will fill in the rest with LCs from a different category, so that the assigned grouping count is always met. As a first pass for this workflow, each LC will only be reviewed by a single turker, unless LCs are needed to fill in the gaps for groups that do not meet the group size parameter.

With the LC images formed into their given groups, conforming to the two configurable parameters, the next step is to make the assignment, known as a HIT, for Mechanical Turk. Amazon offers a variety of free APIs that can be used to programmatically generate an assignment on the platform using different common programming languages. These APIs will be leveraged, along with a provided HTML template file, to embed the LC images so that the turkers can review them. The first part of the HIT will explain the task at hand, which involves turkers reviewing a series of graphs and

ranking them in terms of which need the most improvement. In this case, needing improvement corresponds to which LCs do not demonstrate learning or a good fit for the given KC. To provide these workers with a frame of reference and background information on these concepts, a brief yet informative exert will be used to explain LCs and their corresponding five categories. An example of such text can be found below (Stamper et al., 2010).

“A learning curve visualizes changes in student performance over time. The line graph displays opportunities across the x-axis, and a measure of student performance along the y-axis. A good learning curve reveals improvement in student performance as opportunity count (i.e., practice with a given knowledge component) increases.”

After the turkers read the HIT instructions and the exert regarding LCs and their categories, using concise language pulled from DataShop, they will be presented with five learning curves. Each of these LCs will distinctively fall into one of the five aforementioned categories, such as the LC for too little data having a single point or the LC for low and flat having five points that all remain in the 10-15% range. To establish a baseline for accuracy, the turkers will first be asked to categorize each of these five curves. In addition to categorizing them, they will be asked to rank the LCs in a unique order of 1-5, where 1 indicates the present LC that needs the least improvement (such as a good one) and 5 indicating an LC that needs the most improvement (such as still high). These LCs are to be ranked in comparison to one another, so all five will be proximally located near one another in the interface. This is done in order to gauge their accuracy of the presented information and interpretation of the LCs. If they incorrectly categorize or fail to rank an LC in an order that is far off, their results will not be included in the output.

Following the accurate completion of this baseline portion, the turkers will be instructed to perform the same task for a set of the grouped LCs that were input from the learning curve visualization component. An example with the default configuration enabled might present ten LCs from the still high category, all located near on another on the same page, and ask the turker to again select which category each curve would fall into and how they would uniquely rank order each curve in terms of needing improvement. Note in Figure 2, showing an example of how an LC might be presented, the values 4 and 9 are greyed out since each ranking can only be used once per grouping. Once all presented LCs are ranked and their perceived category is selected, the turker can submit their HIT for completion.

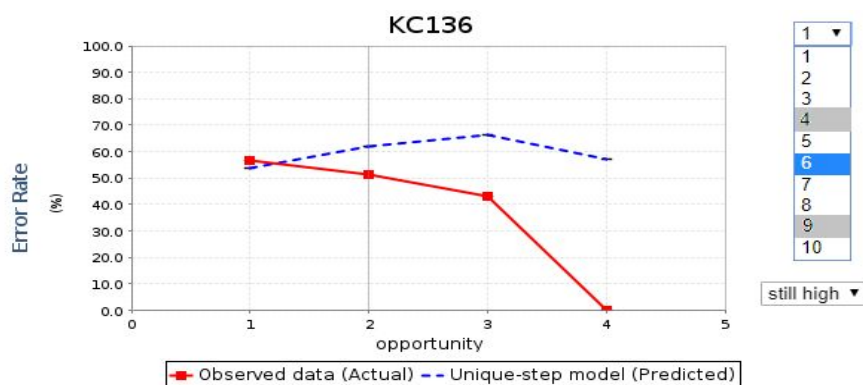


Figure 2: Prototype of how the learning curves can be ranked and categorized by a crowd worker

Currently the price point suggested for this proposed workflow task is 0.70 USD, based on similar image review tasks present on the platform while still offering a living wage amount. Additionally, this task is expected to take no more than five minutes to complete and even faster if they repeat the task for a new set of LCs.

2.3 Workflow Outputs

The primary output will be a text file to display, similarly to the results view for imported data from DataShop or another source. This will display correctly in an HTML friendly format and have the option to be downloaded, so that it can be imported into another environment, such as R Studio or Excel, for further analysis. This output text file will contain the tab separated data in an organized view with each row corresponding to an input learning curve. XML data from the learning curve visualization component, consisting of the DataShop assigned category, number of curve points, and KC name will be present for the columns of the table. It will be trivial to map this output data back to other data frames consisting of more detailed KC and LC information, since these rows can be matched by the common LC name found in both. Additionally, two more columns providing the crowd worker assigned LC category and ranking order will be present in the file. These two columns are the core analysis addition, their usefulness is detailed in the following discussion section.

3 DISCUSSION

One of the key goals of this workflow is to build upon the three-step process of LearnSphere: import, analysis, and visualization. An aspect that commonly gets neglected, but is essential in connecting this iterative process, is refinement. Many educational systems and courses often take initial efforts to construct appropriate content, but they unfortunately fail to iterate on these efforts after evaluation. While issues like a lack of continued funding might affect this lack of effort, the time such efforts take is a large barrier. This workflow looks to mitigate that by using crowd workers to further categorize and rank the LC visualizations so the ones needing the biggest improvement can easily come to light. The idea is the largest improvement and impact can be made back into a course, by addressing these most troublesome and ill-fitting KCs as identified by the crowd workers in their LC review.

While DataShop currently categorizes curves, it can benefit from having a knowledgeable human assist in the categorization process. For larger datasets, there might be hundreds of curves which fall into a given category. This automatic grouping becomes less useful when the user is unable to easily assess which curves might need the most attention, especially from such a large collection that would be difficult to display all at once. Having crowd workers take these categories and rank the LCs in them in order of which appear to need the most attention provides a better way to efficiently select which KCs to work on. Even with fine tuned parameters, the categories assigned by DataShop sometimes do not accurately group or portray KCs that need attention. For instance, the two LCs in Figure 3 are categorized the same, yet it is clear the bottom LC is representative of a KC that would need attention by comparison. It also allows the comparison of human categorized LCs to the categories assigned by another EDM workflow, in this case DataShop.

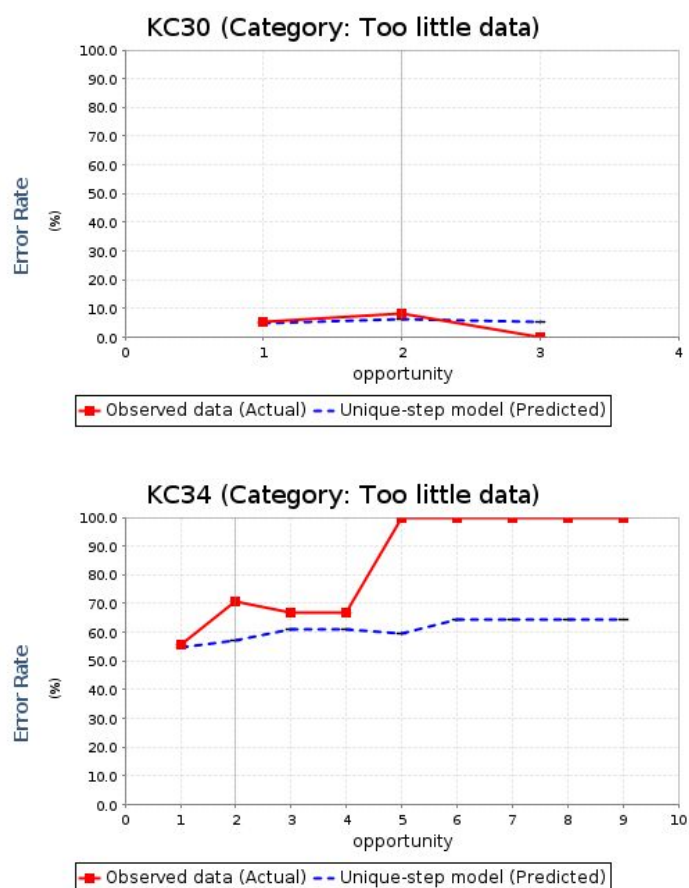


Figure 3: Two learning curves placed in the same category, with the bottom curve demonstrating lower learning than the learning curve at the top

Integrating humans in the loop helps to not only more accurately identify which KCs need improvement, but helps to maintain this iterative process of using the data these systems output to improve them. This is a key aspect of several professions, such as learning engineers and instructional designers, one that is often difficult to maintain and time-consuming. This offers a cheaper and faster alternative, all while removing a more tedious aspect of the process, so that these professionals can leverage their expertise where it counts.

A final goal of this workflow is to see where we can leverage crowdsourcing for work in the learning sciences that is not directly related to content creation or curation. The concept of a LC may sound filled with jargon at first, but it can be boiled down to basic line graph interpretation. Workers can still contribute meaningful input to the process, despite not having an explicit background in a domain related to learning sciences. Other aspects of educational data analysis can benefit from breaking down the task in a similar way, so crowd workers can contribute without needing such expertise. This lack of expertise might also provide a unique lens to look at the problem, categorization, etc. in a way that provides beneficial insights into improvements. This workflow's code can also be leveraged for components at different parts of the workflow, not just following the visualization portion like this component functions. Other instances, especially regarding data preprocessing, may leverage from the review of crowd workers before moving onto the next component.

REFERENCES

- Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science*, 13(4), 467-505.
- Anderson, M. (2011). Crowdsourcing higher education: A design proposal for distributed learning. *MERLOT Journal of Online Learning and Teaching*, 7(4), 576-590. Automated Feedback Generation for Introductory Programming Assignments.
- Balakrishnan, R. (2006, March). *Why aren't we using 3D user interfaces, and will we ever?* Paper presented at the IEEE Symposium on 3D User Interfaces. <http://dx.doi.org/10.1109/vr.2006.148>
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer New York.
- Cen, H. et al. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *ITS '06* (2006), 164–175.
- Fancsali, S. E., Ritter, S., Stamper, J., & Nixon, T. (2013). Toward “hyperpersonalized” Cognitive Tutors. In *AIED 2013 Workshops Proceedings Volume* (Vol. 7, pp. 71-79).
- Gagné, M., Forest, J., Vansteenkiste, M., Crevier-Braud, L., van den Broeck, A., Aspel, A. K., . . . Westbye, C. (2015). The Multidimensional Work Motivation Scale: Validation evidence in seven languages and nine countries. *European Journal of Work and Organizational Psychology*, 24(2), 178-196. <http://dx.doi.org/10.1080/1359432x.2013.877892>
- Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In *Authoring tools for advanced technology learning environments* (pp. 491-544). Springer, Dordrecht.
- Paulin, D., & Haythornthwaite, C. (2016). Crowdsourcing the curriculum: Redefining e-learning practices through peer-generated approaches. *The Information Society*, 32(2), 130-142.
- Porcello, D., & Hsi, S. (2013). Crowdsourcing and curating online education resources. *Science*, 341(6143), 240-241.
- Stamper, J., Koedinger, K.R. (2011) Human-machine Student Model Discovery and Improvement Using DataShop. In Kay, J., Bull, S. and Biswas, G. eds. *Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011)*. pp. 353-360. Berlin Germany:Springer.
- Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A data analysis service for the learning science community. In *International Conference on Intelligent Tutoring Systems* (pp. 455-455). Springer, Berlin, Heidelberg.
- Weld, D. S., Adar, E., Chilton, L., Hoffmann, R., Horvitz, E., Koch, M., ... & Mausam, M. (2012, July). Personalized online education—a crowdsourcing challenge. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 1-31).

Modeling step duration to enhance the Additive Factors Analysis Model

Irene-Angelica Chounta
University of Tartu
chounta@ut.ee

ABSTRACT: In this paper, we explore how we can use step duration as a feature for predicting student performance. In particular, we aim to implement an enhanced version of a standard cognitive student model, the Additive Factors Analysis Model (AFM) using step duration as a quadratic predictive feature. Our work builds on related research that suggests that response time can provide information with respect to correctness and that the relationship between response time and student performance is non-linear. The model we implement here will support extensive testing of the approach using various datasets and it will contribute to gaining insight with respect to the relationship between response time and student.

Keywords: student modeling, step duration, intelligent tutoring systems

1 INTRODUCTION

In this paper, we propose the implementation of an Additive Factors Analysis Model (AFM) (Cen, Koedinger, & Junker, 2006, 2008) enhanced with a quadratic, step duration parameter. The motivation is to use the aspect of time in order to improve the performance of student models. Related research has explored the use of response or reaction time to model students' activity in learning tasks (I.-A. Chounta & Avouris, 2015). Even though research studies have shown that response time can potentially be a good predictor of post-test scores, it does not always predict performance in individual learning steps (Lin, Shen, & Chi, 2016). At the same time, prior studies suggest that the relationship between response time and student performance is non-linear (Carvalho, Gao, Motz, & Koedinger, 2018; Daniel & Broida, 2004). On the one hand, a student needs a minimum amount of time in order to process the problem, retrieve appropriate information, and to construct a correct response. If the student attempts to respond too fast, this can mean that either they did not really process the task as required or that the student attempts to game the system. On the other hand, if the student takes too long to respond, this may indicate lack of background knowledge, failure to retrieve critical information, and inability to address the step (I. A. Chounta & Carvalho, 2018).

In this paper, we propose a new modeling approach for predicting student performance using the student's response time. In particular, we build on the hypothesis that there is no linear relationship between student response time and correctness: a student who takes either too little time or too long to respond to a step (where a step can be either a tutor's question or task), will most likely be unsuccessful for this particular step. Therefore, we argue that modeling a student's response time as a quadratic factor - rather than a linear one - will result in more accurate and better performing student models (I.-A. Chounta & Carvalho, 2019). This rationale is depicted in **Figure 1**.

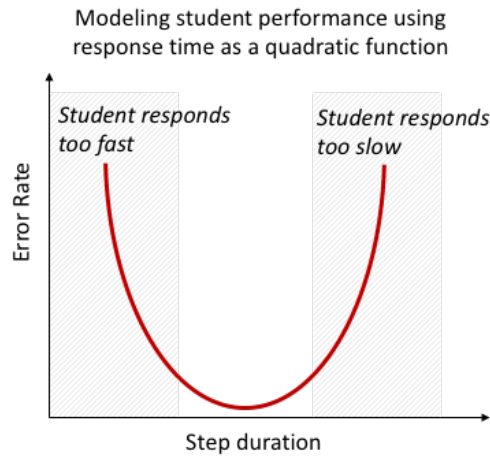


Figure 1. The research hypothesis of this work is that a student who takes either too little time (left grey area) or too long (right grey area) to respond to a step, will most likely be unsuccessful for this particular step resulting to a high error rate in student’s performance. Here we explore whether modeling a student’s response time as a quadratic factor will result in more accurate student models.

The significance of this work is two-fold: first, being able to use response time consistently as a predictive feature will contribute towards improving the performance of student models; second, it will offer insight with respect to the relationship between response time and student performance.

2 WORKFLOW METHOD

2.1 Data Inputs

As data input, our workflow uses standard PSLC DataShop transaction-level datasets (Koedinger et al., 2010). In particular, we use the following fields: *Anon Student Id*, *Duration (sec)*, *Tutor Response Type*, *Attempt At Step*, *Outcome* and *KC (Single-KC, Unique-step, Default)*. Moreover, there is the need for an additional field that will contain information about practice opportunities on the KC-level. That is, how many times a student practiced a specific KC until this given moment.

After importing the data, minor data processing is required in order to discard outliers or to treat specific conditions (like for example, hints). Additionally, the student-step roll up datasets can be used. In this case, we use the following fields: *Anon Student Id*, *First Attempt*, *Step Duration(sec)*, *KC (Single-KC, Unique-step, Default)* and *Opportunity (Single-KC, Unique-step, Default)*. Like before, minor data cleaning and preprocessing is necessary.

2.2 Workflow Model

Our model builds on the AFM and enhances it by adding response time (or else, step duration) as a quadratic feature. For the implementation of the AFM model, we followed Datashop’s proposed approach¹ shown in the regression formula (1):

¹ <https://pslcdatashop.web.cmu.edu/help?page=rSoftware>

$$(1) \text{ AFM} = \text{Outcome} \sim \text{Student} + \text{KC} + \text{KC:Opportunity}$$

where:

- *Outcome* is the result per step – correct or incorrect;
- *Student* stands for the student id of the student who carries out this step;
- *KC* is the skill involved in this step;
- *KC:Opportunity* stands for the number of previous attempts a student had on this particular skill.

In order to take into account students' response time when predicting performance, we enhance the standard AFM by adding step duration as a quadratic component to the original AFM model. This is depicted in the regression formula (2).

$$(2) \text{ AFM-QT} = \text{Outcome} \sim \text{Student} + \text{KC} + \text{KC:Opportunity} + \text{step_duration} + (\text{step_duration})^2$$

where:

- *Outcome* is the result per step – correct or incorrect;
- *Student* stands for the student id of the student who carries out this step;
- *KC* is the skill involved in this step;
- *KC:Opportunity* is the number of previous attempts a student had on this particular skill.
- *step_duration* is the time the student took to carry out this step (in seconds).

The AFM-QT model has been implemented and tested in R². We are currently working on the Tigris implementation. Our goal is to have the model ready before the workshop so that we can test it extensively and together with other participants.

2.3 Workflow Outputs

Our objective is to compare the predictive performance of the AFM-QT model with respect to different data inputs – that is, the transaction-level and the student-step roll up – and with respect to other modeling implementations – that's is, the AFM and potentially the Performance Factors Analysis Model (PFM) (Pavlik Jr, Cen, & Koedinger, 2009) and the Instructional Factors Analysis Model (IFM) (Chi, Koedinger, Gordon, Jordon, & VanLahn, 2011). Thus, we expect – as outcomes – measures of the models' quality and performance that can be used to assess the predictive fit of the model to data. In particular, we would aim for the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Cross-Validation estimate of Accuracy (CV.ACC) and the Root Mean Square Error (RMSE). These metrics have been used in related work for choosing between parametric models with different numbers of parameters (Cen et al., 2006; Pavlik Jr et al., 2009)

² <https://cran.r-project.org/>

Additionally, we aim to retrieve predictions of student performance from the AFM-QT model on the transaction and step levels as well as to use learning curves for data visualization.

3 DISCUSSION

Our overarching goal is to identify an appropriate way to model response time as a predictor of student performance. Taking into account that the relationship between response time and performance in terms of correctness is not linear, we propose to model step duration as a quadratic parameter. To do that, we build on a standard cognitive model (AFM) and we enhance it by adding a quadratic, step duration parameter (AFM-QT).

To further study the effect and potential benefits of this approach, we aim to test and compare the AFM and the AFM-QT over a wide range of datasets. This is a time-consuming process that requires processing and manipulation of extremely big datasets as well as computationally demanding procedures for comparing the performance of different student models.

With this work, we aim:

- to communicate this research line to users of Datashop and LearnSphere and to provide them with tools that will allow them to apply our approach on their data;
- to encourage other researchers to reproduce our study and to pursue further collaboration;
- to support our work by developing a tool that will help us test our research hypothesis on multiple datasets in a cost-efficient and automated way.

For future work, we plan to extend this approach in combination with the Performance Factors Analysis Model (PFM). We envision this is an important step because PFM differentiates between correct and incorrect steps and thus, it allows modeling step duration separately for correct and incorrect outcomes.

REFERENCES

- Carvalho, P. F., Gao, M., Motz, B. A., & Koedinger, K. R. (2018). Analyzing the relative learning benefits of completing required activities and optional readings in online courses. *Methods*, 34, 68.
- Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems* (pp. 164–175). Springer.
- Cen, H., Koedinger, K., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In *International Conference on Intelligent Tutoring Systems* (pp. 796–798). Springer.
- Chi, M., Koedinger, K. R., Gordon, G. J., Jordon, P., & VanLahn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions.
- Chounta, I. A., & Carvalho, P. (2018). Will time tell? Exploring the relationship between step duration and student performance. In *Rethinking Learning in the Digital Age. Making the Learning Sciences Count* (Vol. 2, pp. 993–996). London, United Kingdom: International Society of the Learning Sciences.

- Chounta, I.-A., & Avouris, N. (2015). Towards a time series approach for the classification and evaluation of collaborative activities. *Computing and Informatics*, 34(3), 588–614.
- Chounta, I.-A., & Carvalho, P. F. (2019). Square it up! How to model step duration when predicting student performance. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*.
- Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, 31(3), 207–208.
- Koedinger, K. R., Baker, R. Sj., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*.
- Lin, C., Shen, S., & Chi, M. (2016). Incorporating Student Response Time and Tutor Instructional Interventions into Student Modeling. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 157–161). ACM.
- Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis—A New Alternative to Knowledge Tracing. *Online Submission*.

Python Bootcamp for Learning Analytics Practitioners (Full day Tutorial)

Alfred Essa

McGraw-Hill Education
alfred.essa@mheducation.com

Shirin Mojarad

McGraw-Hill Education
shirin.mojarad@mheducation.com

Neil Zimmerman

McGraw-Hill Education
neil.zimmerman@mheducation.com

ABSTRACT: The hands-on tutorial will provide a rigorous introduction to python for learning analytics practitioners. The intensive tutorial consists of five parts: a) basic and intermediate python; b) statistics and visualization; c) machine learning d) causal inferencing and d) deep learning. The tutorial will be motivated throughout by educational datasets and examples. The aim of the tutorial is to provide a thorough introduction to computation and statistical methodologies in modern learning analytics.

Keywords: python, machine learning, statistics, causal inferencing, deep learning, visualization

TUTORIAL TOPICS

1.1 *Python*. Python is the de facto language for scientific computing and one of the principal languages, along with R, for data science and machine learning. Along with foundational concepts such as data structures, functions, and iteration we will cover intermediate concepts such as comprehensions, collections, generators, map/filter/reduce, and object orientation. Special emphasis will be given to coding in “idiomatic Python”.

1.2 *Exploratory Data Analysis, Statistics*. In this section we will introduce the core python libraries for exploratory data analysis and basic statistics: numpy, pandas, matplotlib and seaborn. We will use the Jupyter Notebook environment for interactive data analysis, annotation, and collaboration. Exploratory data analysis is a foundational step for deriving insights from data. It also serves as a prelude to building formal models and simulations.

1.3 *Machine Learning*. In this section we will introduce participants to basic machine learning concepts and their application using the scikit-learn library. We will show how to predict continuous and categorical outcomes, for example, using linear and logistic regression. This demonstration will show how to create an entire prediction pipeline from scratch, starting from loading in data, cleaning and standardizing it, building the model, and demonstrating its validity through cross-validation. Some discussion of what an educator might do with such a model will be included.

1.4 *Causal Inferencing*. In this section of the tutorial we build on our statistical understanding of correlation to study causality. Randomized control trials (RCTs) are considered the gold standard in efficacy studies because they aim to establish causality of interventions. But RCTs are very often impractical to carry out and have other limitations. Causal inference from Observational Studies (OS) is another form of statistical analysis to evaluate intervention effects. In causal inference, the causal effect of an intervention on a particular outcome is studied using observed data, without the need for randomization in advance. In this tutorial, we will show design of an OS to leverage the large amounts of data available through online learning platforms and student information systems to draw causal claims about their effectiveness.

1.5 *Deep Learning*. In this section we introduce how to build deep learning models. Deep learning is one of the fastest growing areas of machine learning and is particularly well suited for very large datasets. We begin by building a toy deep learning model by scratch in python. This is to understand the five foundational concepts of deep learning: *neurons* as the atomic computational unit of deep learning networks; neurons as organized in stacked *layers* to achieve increasingly abstract data representations; *forward propagation* as the end-to-end computational process for generating predictions; *loss* and *cost functions* as the method for quantifying the error between prediction and ground truth; and *back propagation* as the computational process for systematically reducing the error by adjusting the network’s parameters. After developing a conceptual understanding of deep learning, we apply some standard Python libraries such as Keras, PyTorch, and TensorFlow to build deep learning models.

PREREQUISITES AND APPROACH

Students should be familiar with basic programming concepts, preferably Python or R. The hands-on workshop will be entirely interactive. Students will learn by coding in the Jupyter Notebook environment. All the work will be done on the cloud use the Google Colab Research environment. All instructional materials, including slides and the Jupyter Notebooks, will be available in a public *github* repository.

OBJECTIVES AND OUTCOME

The aim of the tutorial is to introduce the core techniques and methodologies in data science for the learning analytics practitioner. After having attended the tutorial the student will be well versed in each of the principal topics in data science. Because we will make extensive use of educational datasets, attendees will also acquire practical knowledge in how to apply data science methodologies in learning analytics and data mining.

Developing A Learning Analytics Community for Ethical Discourse

Author(s): James Folkestad

Colorado State University

James.Folkestad@ColoState.EDU

George Rehrey

Indiana University

grehrey@indiana.edu

Linda Shepard

Indiana University

lshepard@indiana.edu

Dennis Groth

Indiana University

dgroth@indiana.edu

Mathew Hickey

Colorado State University

Matthew.Hickey@ColoState.edu

ABSTRACT: This half-day interactive workshop responds to an on-going need to thoughtfully and intentionally consider, and sometimes reconsider, the ethical implications of the rapidly advancing field of Learning Analytics (LA). The pioneering work of other scholars will provide the starting point for our conversations, including Drachsler & Greller's (2016) DELICATE checklist Hoel and Chen's (2018) EP4LA Toolkit, and Sclater's (2014) Code of Practice. Case studies and possible dilemmas (Willis, Slade, & Prinsloo, 2016), along with previous institutional efforts (Colorado State University) will also frame our discussions. During the workshop, participants will develop strategies for creating a sustainable and inclusive community to advance principle-based LA practices on their campuses. By completing an Action Plan Worksheet, participants will consider the alignment of institutional goals with LA, the value of including key stakeholders in ethical discourse, and the development of a flexible framework for reviewing emerging LA practices and activities. They will also reflect upon how the development of local communities dedicated to ethical discourse can contribute to, and benefit from, joining a broader international Community of Transformation across higher education.

Keywords: Collaborative, community, ethics, principles, community of transformation, social networks, change management

1 WORKSHOP BACKGROUND

Learning Analytics and the use of student records to inform practices in higher education, while relatively new, is already changing the face of higher education. Institutions are developing applications to identify students at-risk (Arnold & Pistilli, 2012), to assist with course pathways (Heileman, Babbitt, Abdallah, & Dougher, 2014), or to facilitate course selection (Fiorini, et al., 2018). And yet, while these evidence-based actions are continuously

emerging, concerns remain about ethical issues and adverse consequences (Sclater, 2016) of activities at our schools. This is also the case when it comes to using LA to help and support marginalized populations such as underrepresented minority, first generation, and indigenous students. Researchers and institutions want to enhance the student experience, improve student success, do no harm, and act responsibly and ethically. And yet, we may often find ourselves working in isolation.

1.1 Motivation for the Workshop

When it comes to using data and making decision using learning analytics, what is ethical? How do we know if or when we have crossed an ethical line? Should we adhere to universal principles, or should ethical guidelines be shaped to meet the needs of an institution's specific circumstances? These questions, and many more, were a working session topic at the inaugural Learning Analytics Summit (LAS), held at Indiana University Bloomington in April 2018. Although participants were able to reference previous work (Hoel & Chen, 2018; Center for the Analytics of Learning and Teaching, 2017; Drachsler & Greller, 2016; Charles Sturt University, 2015; JISC, 2015; Pardo & Siemens, 2014; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978), many of them had difficulty pointing to a clear process designed to help answer these foundational questions. Given these uncertainties, this workshop will share a process that can help each individual institution establish a customized set of guidelines that acknowledges the current use of LA on their campus, their future plans and goals, and their connections to the broader LA community. Additionally, such a process creates an opportunity to advance the adoption of LA through dialog and outcome driven conversation with key stakeholders.

1.2 Relevance to Conference Attendees/ LAK Research Community

A process of developing LA principles and subsequent policy, particularly around inclusion and success, is of primary relevance to attendees and the LAK research community. Previous efforts have attempted to provide guidance in the form of checklists (Drachsler & Greller, 2016) and a proposed toolkit called the Ethics and Privacy for LA, EP4LA Toolkit (Hoel & Chen, 2018). Nonetheless, it remains challenging to find well-articulated processes that administrators, faculty and staff can follow to develop LA principles, which need to be clearly articulated prior to establishing policies.

In this workshop, participants will discover ways to engage administrators and faculty in an outcome-driven dialog concerning the ethical dimensions of current and future uses of LA. We also propose that local communities connected to a larger network of schools will provide a basis for ongoing cross-institutional conversations about LA ethics, expand our knowledge of the topic, and provide new opportunities for policy development in a field that is continually evolving.

1.3 Contribution to the Research Field

In the context of implementing innovations in learning analytics, MacFayden, et al., (2014) state:

“success, requires a willingness to engage...with people and resources available in the context and linking them in ways that support work towards the vision.”

In this workshop we provide a process that can evolve ethical discourse in our local and external LA communities. We begin with the ethical conversations in progress (Willis, Slade, & Prinsloo, 2016; JISC, 2015) and expand the discourse with the goal of facilitating a transformative community (Kezar & Gehrke, 2015; Wenger-Trayner, 2015), relying on the power of social networks (Williams, et al., 2013). These communities hold promise to keep pace with the quickly evolving field of LA, reducing barriers to adoption, creating forums for knowledge exchanges, and thereby enhancing sustainability (Kezar, 2016).

1.4 Previous Learning Analytics Work

Concerns surrounding the ethical use of big data have been clearly raised, particularly for marginalized populations, in a variety of social contexts (NYT best seller, *Weapons of Math Destruction*, O’Neil, 2016). Educational researchers share these concerns in the context of learning analytics (Pardo & Siemens, 2014). Concerns about risk of unfair treatment for specific populations (Drachsler, et al., 2015) or lack of guidelines contribute to barriers for the rollout of LA tools (Sclater, 2016).

Willis, et al. (2016) consider the oversight processes at three cross-continental institutions discovering that a framework for evaluating ethical practice is lacking. Sclater (2016) offers a solution by presenting a process and extensive taxonomy that identifies issues to consider for an institution developing a code of practice. While the endgame may be to develop a formal code to inform LA practices, institutions may find themselves at different planning and implementation stages. Colorado State University started their process by considering the institutional mission, and developed a set of ethical principles from which policies may evolve.

MacFayden, et. al (2017) propose the ROMA model as an adaptive process to facilitate innovative changes in complex educational systems. This model recognizes and addresses the cultural shifts that need to take place for full adoption of innovations on our campuses. This six-step process model has been successfully applied to LA contexts (Macfayden, et al., 2017) that include: mapping the context, identifying stakeholders, identifying areas for framing change, developing a strategy for change, considering the resources required, and evaluating changes. While all sequences of the ROMA model will be considered, we are particularly mindful of engaging key players in a community. These key players will serve as a local community, where the group engages in a collective task of understanding LA activities and developing ethical guidance for their campus.

1.5 Communities of Transformation

The power of social networks and communities have been highly effective in other faculty work (Williams, et al., 2013) for introducing change and the sustainability of a changed culture. More recently Kezar and Gehrke (2015) expand the concept of a Community, suggesting that a collective of CoPs embedded within a larger cross-institutional Community of Transformation (CoTs) scale the work to empower and develop cultural norms. They state that “internal and external conditions shape and frame change processes” thus the role of the community is no longer just a facilitator of change but an imperative for transformational change.

Organic in nature, CoTs create and foster a new space for innovation while helping to define a vision for a relatively new domain. Within the community, an adherence to a shared philosophy is central to sustaining and engaging the participants in joint activities, as they learn from one another through mostly long-distance interactions. Through those interactions, CoTs continue to develop their shared philosophy, organically support new leadership, and most importantly, generate guiding documents to advance a particular field of practice (Kezar & Gehrke, 2015).

The purpose of this workshop is not to create rules for oversight, that is the decision of each institution base upon their institutional context, but rather to guide leaders and participants in carefully framing a process to consider the directions and implications of this emerging work, to “move from isolated action to cultural norm,” (Williams, et. al, 2013).

1.6 Workshop Organizers

Workshop organizers reside at large research institutions. They include the Vice Provost for Undergraduate Education, the Assistant Vice Provost for the Bloomington Assessment and Research Office, and 2 directors from LA center’s.

2 WORKSHOP DETAILS

This ½ day workshop participants will create a plan for engaging their administrators, faculty and other stakeholders in a conversation about the ethics of learning analytics. Participants will first explore the challenges and successes of similar efforts before developing an action plan for ethical discourse on their campuses. They will also be provided the opportunity to join a CoT dedicated to advancing principle-based LA policy through collaboration and sharing, which is the ultimate goal of this workshop.

2.1 Workshop Activities

The purpose of the workshop is to engage scholars and practitioners in ethical dialog, including the procurement, provisioning and use of student data for LA activities. Attendees will be contacted prior to the workshop and asked to complete a short reading. We will also suggest that they bring any documentation from their campus related to LA ethical principles and/or codes of practice. Through a series of individual activities and small group discussions, participants will be encouraged to complete an Action-Plan Worksheet provided to them during the session. The worksheet will help set the stage for future ethical discourse once they return home, taking into consideration their own institutional culture and context. Participants will also be provided the opportunity to connect to a broader community that will enrich and support future iterations of this work.

2.2 Workshop Outcomes

Participants will: 1) identify existing relevant ethical principles, LA principles, and LA codes of practice, 2) discuss current methods for implementing LA principles and codes of practice, 3) contribute their voices to the ongoing development and advancement of LA principles and subsequent policy, 4) use the ROMA model to create a plan for an engagement strategy for their campus.

2.3 Dissemination of Outcomes

The workshop will be promoted using the #LAK19 on various social media channels and will be announced and disseminated among the members of CSU's and IUB's Center for Learning Analytics and Student Success (CLASS) LA-listserv. The outcomes from the workshop activities will be collected and disseminated using social media (#LAK19), email lists for participants, distributed to IU's and CSU's LA Centers and distributed to the community of practice that is being developed around these issues.

REFERENCES

- Arnold, K. & Pistilli, M. (2012). *Course Signals at Purdue: Using learning analytics to increase student success*. Paper presented at LAK 2012: 2nd International Conference on Learning Analytics and Knowledge. Vancouver, 2012.
- Center for the Analytics of Learning and Teaching. (2017). *Ethical Principles of Learning Analytics*. Colorado State University, 2017. Retrieved from <https://alt.colostate.edu/cotl-ethical-principles-la/>
- Charles Sturt University. (2015). *CSU Learning Analytics: Code of Practice*. Retrieved from http://www.csu.edu.au/data/assets/pdf_file/0007/2160484/2016_CSU_LearningAnalyticsCodePractice.pdf
- Drachsler, H., Cooper, A., Hoel, T., Ferguson, R., Berg, A., Scheffel, M., Kismihok, G., Manderveld, J., & Chen, W. (2015). Ethical and privacy issues in the application of learning analytics. *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 390-391.
- Drachsler, H. & Greller, W. (2016). *Privacy and analytics - it's a DELICATE Issue: A checklist for trusted learning analytics*. Sixth International Conference on Learning Analytics & Knowledge (LAK '16). Edinburgh, 2016.
- Fiorini, S., Sewell, A., Bumbalough, M., Chauhan, P., Shepard, L., Rehrey, G., & Groth, D. (2018). *An application of participatory action research in advising-focused learning analytics*. Learning Analytics and Knowledge Conference 2018. Sydney, 2018.
- Heileman, G., Babbitt, T., Abdallah, C., & Dougher, M. (2014). Efficient Curricula: The Complexity of Degree Plans and the relation to Degree Completion. *Collection of Papers 2014: Higher Learning Commission*.
- Hoel, T. & Chen, W. (2018). *Advancing the Delicate Issue of Ethics and Privacy for Learning Analytics*. Learning Analytics and Knowledge Conference 2018. Sydney, 2018. Retrieved from http://www.hoel.nu/files/Hoel_Chen_LAK-18_companion_proceedings.pdf
- JISC. (2015). "Code of Practice for Learning Analytics." Retrieved from <https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics>
- Kezar, A. (2016). *How colleges change: Understanding, leading and enacting Change*. New York and London: Routledge Taylor & Francis Group.
- Kezar, A. & Gehrke, S. (2015). *Communities of transformation and their work scaling STEM reform*. Pullias Center for Higher Education: Ross School of Education University of Southern California: 1-86.
- Macfayden, L., Dawson, S., Pardo, A., & Gasevic, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy change. *Research and Practice in Assessment*, 9: 17-28.

- Macfadyen, L., Groth, D., Rehrey, G., Shepard, L., Greer, J., Ward, D., Bennett, C., Kaupp, J., Molinaro, M., & Steinwachs, M. (2017). Developing institutional learning analytics “communities of transformation” to support student success. *In Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 498–499.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). “The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.” Bethesda, Md.: The Commission.
- O’Neil, C. (2016). *Weapons of Math Destruction*. New York: Crown Publishing Group.
- Pardo, A. & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438-450.
- Sclater, N. (2014). Code of practice for learning analytics: A literature review of the ethical and legal issues. *JISC*, 2-60.
- Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics*. 3(1), 16-42.
- Wenger-Trayner, E. & Wenger-Trayner, B. (2015). Communities of practice: A brief introduction. *Communities of Practice*, 1-8.
- Williams, A., Verwoord, R., Dalton, H., McKinnon, J., Strickland, K., Pace, J., & Poole, G. (2013). The power of social networks: A model for weaving the scholarship of teaching and learning into institutional culture. *Teaching and Learning Inquiry*, 1(2), 49-62.
- Willis, J., Slade, S. & Prinsloo, P. (2016). Ethical oversight of student data in learning analytics: a typology derived from a cross-continental, cross-institutional perspective. *Educational Technology Research and Development*, 64(5), 881-901.

Achievement Emotions and Attritions in Massive Open Online Courses: Using Machine Learning Models

Hengtao Tang

University of West Georgia

htang@westga.edu

Wanli Xing

Texas Tech University

Wanli.xing@ttu.edu

ABSTRACT: Massive Open Online Courses (MOOCs) have the potentials of opening up access to higher education, but their attrition rates up to ninety percent are now a major concern for the sustainability of MOOCs. To prevent attritions in MOOCs, supporting struggling students from the perspective of emotions may be an attainable option. Prior research on emotions in MOOCs mainly focused on the dichotomy of emotions (e.g., negative and positive), but a more integrative framework was needed. Following the control-value theory of achievement emotions (Pekrun, 2006), this research applied machine learning models to explore achievement emotions and the mechanism of how they influenced attritions in MOOCs. The research identified positive deactivating and negative activating emotions as contributors to dropouts. Design and pedagogical implications are also discussed in the end.

Keywords: Achievement emotions; Machine learning; MOOCs; Attritions.

1 INTRODUCTION

Emotions influenced online learning experience in a complicated way that positive emotion is not always positively related to endurable commitment in online courses and negative emotion may engender a beneficial effect on learning outcomes (Barak, Watt, & Haick, 2016). In fact, emotions in academic settings are beyond the positive/negative dichotomy. The control-value theory of achievement emotions provides an integrative framework to understand emotions relevant to learning (Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, 2017), which include emotional states of positive activation, positive deactivation, negative activation and negative deactivation (Pekrun, 2006). Understanding achievement emotions and their impact on learning commitment holds the potential of revealing how struggling learners gradually dropped a MOOC and thus offered implications on future social-emotional intervention designs.

This work thus explored achievement emotions in MOOCs and quantified their influence on attritions. Using a MOOC dataset, the research first trained and validated a classifier that automatically classified achievement emotion states in the forum posts. Then we used survival modeling techniques to quantify the influence of different achievement emotions on student attrition in MOOCs. The research added to the empirical evidence on facilitating emotional states and supporting learner success and sustained engagement in online courses.

2 THEORETICAL FRAMEWORK

Achievement emotions are emotions directly related to achievement activities and outcomes (Pekrun, 2006). The control-value theory claims that emotions are evoked by subjective estimation of control and value towards learning activities and outcomes (Pekrun, et al., 2017). For learning, subjective control is the individual appraisal of academic agency over the activities and expected accomplishments (e.g., self-concepts, self-efficacy, and outcome expectation), whereas subjective value denotes the self-perceived importance of the tasks and outcomes (e.g., perceived value). So, the control-value theory adopts two dimensions of human affections to differentiate achievement emotions, valence (positive or negative) and activation (activating or deactivating) (Pekrun, Goetz, Titz, & Perry, 2002). In doing so, the theory categorizes achievement emotions into four groups, namely positive activating (e.g., enjoyment, joy, pride), positive deactivating (e.g., relaxation, relief), negative activating (e.g., angry, anxiety), and negative deactivating (e.g., boredom) emotions (Pekrun, 2006; Pekrun et al., 2017).

The impact of achievement emotions on learner achievement and retentions is complex. Specifically, achievement emotions affect cognitive and self-regulative learning strategies and then influence learning performance (e.g., Artino & Jones, 2012). For example, Artino and Jones (2012) find that positive activating emotions are indicators of the efficient use of cognitive and metacognitive strategies. Achievement emotions might also influence learner attritions. For example, Daniels et al. (2009) conclude that positive activating emotions represent positive valence and agency over the goals and thus positively relate to students' retention rates. In contrast, dropout might result from negative deactivating emotions as a result of negative appraisal of control and value over the activities and outcomes (Pekrun et al., 2017).

3 METHODOLOGY

3.1 Research Dataset and Context

This research used a dataset from a creativity MOOC offered on Coursera by a land-grant university in the United States. The course lasted for six modules and each of them addressed a themed topic about creativity. Each module came with an exclusive forum. Additional two general discussion forums were separately dedicated to sharing projects and reflections for instructors and students. In all, 2084 users posted at least once in the course forum, resulting in a total of 13,513 posts.

3.2 Achievement Emotion Detection

First, the research created the training dataset by manually coding 800 posts randomly sampled from the entire dataset. Two researchers independently coded these posts by detecting their emotional states that learners expressed. Cohen's Kappa value was calculated to assess the inter-rater reliability of the coding. The result (0.864) indicated a reliable level of mutual agreement.

Second, the research captured textual features in the coded forum posts to build the machine learning model. The Linguistic Inquiry and Word Count (LIWC) library (Pennebaker et al., 2015) was used to extract language summary features (LSF) and linguistic features (LF). LSF was used to capture the generic diversity of languages in students' expressed emotions. LF was used to calculate the degree to which student used different linguistic dimensions (e.g. tense, grammar) and psychological

constructs (e.g. positive or negative affect). However, LSF and LF are very generic but not specific for the research context. Instead, topics conveyed different types of emotions and information (Buis, 2008). Latent Dirichlet Allocation (LDA) was thus used to derive latent topics and representative words from each topic were collected to become a dictionary (Blei, Ng, & Jordan, 2003). Then the frequency of words matching the dictionary for each post was calculated in the dataset.

Third, to optimize the model performance, the research employed four supervised machine learning algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machines (SVM, polynomial kernel), and Decision Tree. The algorithms were evaluated using 10-fold cross validation to compare the robustness of their predictions. The classical metric and F-measure were calculated to decide which algorithm would be applied to identify the emotional states for the rest of the forum posts.

3.3 Survival Analysis

The research examined the influence of achievement emotions at a certain time point on the tendency of a student to drop the MOOC afterwards. Survival analysis was used because it could provide less biased estimation with due respect for the truncated nature of time series data. The Hazard Ratio was obtained to explain the influence of an independent variable on the probability of student dropping out (Klein & Moeschberger, 2005). Parametric regression of survival analysis was used with Weibull distribution of survival times. All the active students who contributed to the MOOC forum were included. The interval time was defined as participation days. The beginning point of a student participation was the timestamp for the first post, and end point was the timestamp of his/her last post within this course.

3.3.1 Dependent Variable

Dropout: Dropouts were those students had no forum activities in this research. Specifically, this was a binary variable, with true if a student had forum activity in the last week of the course, and otherwise with false if no forum activity was recorded during the last week.

3.3.2 Independent Variables

Expressed Emotions: This independent variable calculated the average frequency of emotions a student had expressed in a week. It was calculated by the total number of posts falling into certain achievement emotions divided by the number of weeks during which the student participated in forum discussions. This independent variable included four variables: positive activating (PA), positive deactivating (PD), negative activating (NA), and negative deactivating (ND) emotions.

4 RESULTS

4.1 Achievement Emotions Detection Results

By comparing the prediction performance of four algorithms using different feature sets (see Table 1), Decision Tree had the most robust and relatively high performance with F-Measure (72.0%) when used all the language summary features, linguistic features, and the LDA topic features. Then the research used Decision Tree algorithm identified achievement emotions for all the forum posts in the entire dataset (see Table 2).

Table 1. Machine learning model performance (F-measure)

	LSF	LF	TF	LSF + LF	LSF + LF + TF
Naïve Bayes	31.1%	17.7%	18.0%	41.5%	36.3%
Logistic Regression	32.4%	39.1%	24.2%	50.4%	61.9%
SVM	68.0%	64.1%	51.5%	61.4%	68.2%
Decision Tree	68.1%	63.5%	63.8%	64.3%	72.0%

Table 2. Descriptive statistics for the achievement emotions in forum posts

	Mean	Median	SD	Min	Max
PA	0.73	0.5	1.12	0.00	18.00
PD	0.56	0.33	0.82	0.00	12.33
NA	0.36	0.00	0.62	0.00	10.33
ND	0.30	0.00	0.59	0.00	7.00

4.2 Survival Analysis Results

Survival analysis was conducted to quantify the influence of achievement emotions on student attrition using the hazard ratio (see Table 3). This model showed that both positive deactivation and negative activation significantly influenced learner survival in this MOOC, but the influence was negative. Specifically, students with positive deactivation emotions were 51.0% ($100\% * (1.51 - 1)$) more likely than the average to drop out from the course and those with negative activation emotions are 76% more likely than the other peers. In contrast, no relationship was found between learner survivals and the other two emotional states (positive activating and negative deactivating).

Table 3. Survival analysis results

	Hazard Ratio	<i>p</i>
PA	1.18	>0.05
PD	1.51	< 0.05
NA	1.76	< 0.01
ND	0.84	>0.05

5 DISCUSSION

The findings of this research indicate that only positive deactivating and negative activating emotions expressed by learners are significantly correlated to learner survivals, but both correlations are negative. In particular, the expressed negative activating emotions are the largest contributors to attritions in this MOOC. For example, if a student expressed anger or disappointment in the forum, this student is much more likely to dropout the course. However, these negative activating learners are the focus in designing effective interventions because they are still making effort (activating) to stay engaged despite the negative emotions. Revisiting the findings of Artino & Jones (2012), elaborated endeavors from both the facilitation and design perspectives are needed to alleviate the negative influence and encourage these struggling learners to invest more effort in the use of adaptive strategies (e.g., cognitive, metacognitive). From the facilitation side, this result reemphasizes the importance of active instructor facilitation for learner retentions. To maintain learner engagement, instructors are recommended to actively mitigate students' negative activating

emotions (e.g., anger, disappointment) in forum posts. In the design sector, designers and instructors might avoid challenging learners with overwhelming contents in both difficulty and amount. In addition, positive deactivating emotions (e.g., relaxation, relief) are negatively related to the course retentions. This might result from various purposes for learners to register for MOOCs. Learners with relatively relaxed emotions might be those who are not intended to complete the course. Surprisingly, positive activating emotions and negative deactivating emotions are not related to the attritions, although they are both correlated with the use of adaptive strategies and the intention to accomplish the learning tasks (Artino & Jones, 2012). Pekrun et al. (2002) even argue negative deactivating emotions have the worst influence on learner achievement. However, both types of emotions are not related to learning achievement or commitment in MOOCs. This does not assume positive activating emotions are not important in MOOCs since the insignificant relationship might result from the relatively loose structure of MOOCs or learners' different enrolling purposes.

This study yields significant implications for the future deployment of MOOCs. Compared with traditional classroom room and online learning, students in MOOCs are often left unattended due to the class size. This study provides a way to detect students' expressed emotions during the course of learning. Tracking and monitoring the emotional status in MOOCs can guide instructors to provide appropriate feedback to students. For example, when the overall emotion in a MOOC forum is too negative activation oriented, then the instructor should be especially careful and provide help to the students since such emotion has the strongest negative influence on students' survival in the course. The discovered functioning mechanism for achievement emotions in MOOCs can also serve as the base to design and develop automatic feedback to support MOOC learners. The performance of built machine learning models in automatically detecting emotional states brings the hope of using computing methods to analyze the conversations in MOOC forums. For instance, the extracted feature sets can be easily used to construct prediction models in other online learning settings. While language summary features and linguistic features can be directly used in other contexts, LDA topic features might be adapted before using in a specific context, which might require some effort.

REFERENCES

- Artino, A. R., & Jones, K. D. (2012). Exploring the complex relations between achievement emotions and self-regulated learning behaviors in online learning. *The Internet and Higher Education*, 15(3), 170-175.
- Barak, M., Wattied, A., & Haick, H. (2016). Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, 94, 49-60.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.
- Buis, L. R. (2008). Emotional and informational support messages in an online hospice support community. *CIN: Computers, Informatics, Nursing*, 26(6), 358-367.
- Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology*, 101(4), 948.
- Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18(4), 315-341.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 99–105.
- Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development*.88(5), 1653–1670.

What Makes a Question Productive?- An Exploratory Analysis toward Demystifying Question Productivity

Yuan Wang¹, Turner Bohlen², Linda Elkins-Tanton^{1,2}, James Tanton²

Arizona State University¹, Beagle Learning²

Elle.wang@asu.edu, turner@beaglelearning.com, lelkinst@asu.edu, james@beaglelearning.com

ABSTRACT: Not all students equally participate in classrooms. However, being able to ask high-quality questions that are productive toward learning is critical toward gaining critical thinking and problem-solving skills. Faced with the challenge of measuring what makes a question productive, the present paper proposes a Question Productivity Index (QPI) including three dimensions: Relevance, Scale, and Articulation, as a means to explore how question productivity can be quantitatively measured and how it can be integrated into the teaching process to increase learners' critical thinking capabilities.

Keywords: Critical thinking, problem-solving, student-generated questions.

1 INTRODUCTION

1.1 Connection between Asking Questions and Critical Thinking

Questioning skills were found to be a critical component of scientific inquiry and active learning (Chin & Brown, 2002). In classrooms, student-generated questions, as opposed to teacher-generated questions, have been considered to be a key process of cultivating critical-thinking capabilities (White & Gunstone, 2014). A UNESCO report addressed that "asking questions" rather than "answering questions" reflects more of the genuine learning process (1983). Student-generated questions can help reveal not only how much content the student mastered but also provoke critical thinking (Watts et al., 1997).

Asking productive questions can help learners reflect and direct their learning (Chin & Brown) and serve as a critical step in the process of knowledge acquisition and processing (Osborne & Wittrock, 1985). However, not all students equally participate in classrooms. It was observed that only a small percentage of learners ask questions (Dillon, 1988), especially in science classes (White and Gunstone, 1992).

1.2 Challenges of Asking Productive Questions

There may be various reasons behind the lack of student-generated questions in classrooms. One primary reason is that asking good and productive questions is not a simple task. It requires significant investment of cognitive load (Pizzini & Shepardson, 1991). Meanwhile, students' different cultural backgrounds and instructors' teaching styles (Good et al., 1987) may also influence whether or not a student decides to raise a question and the quality of the questions. Therefore, it is paramount to understand how to help learners ask productive questions. Yet, it would be fairly challenging to do so if we don't know how productive questions differ from those that are not productive. A better understanding of the productivity of student-generated questions can help enable teachers to

systematically guide learners toward asking productive questions instead of only acknowledging whether a question is good or not. In so doing, learners may be able to better engage in learning activities and improve their critical thinking and problem-solving capabilities.

2 THE QUESTION PRODUCTIVITY INDEX (QPI)

It is often seen in a classroom that a teacher would comment that a student asked a “good” question. But how does a teacher judge this, and how can a student learn to improve? This question is often found harder to answer than expected. Thereby, toward demystifying the productivity of student-generated questions, we designed a Question Productivity Index (QPI), aiming at quantifying productive attributes of student-generated questions. QPI included three dimensions: Relevance, Scale, and Articulation.

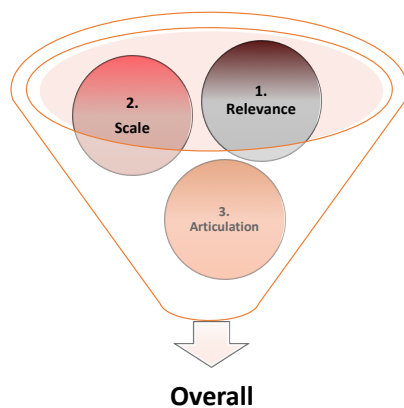


Figure 1: Three Metrics of the QPI

Table 1: Descriptions of QPI Dimensions

Name	Descriptions
1.Relevance	How relevant is the question relevant to the larger learning goal?
2. Scale	The question takes the class one reasonable step from their current knowledge;
3. Articulation	The question is well-posed and uses good grammar.
Overall	An Overall Score on the productivity of the question.

Expert college instructors were recruited to rate a set of student-generated questions from a diverse range of domains including planetary sciences, business, mathematics, etc. Each rater was asked to rate each question on its overall productivity and on the three individual dimensions of the QPI. Thereby, each rater scored each of 109 questions in four ways. Each question was rated by four raters, resulting in 16 scores per question. A QPI rubric were made available to raters to download as a PDF to help with their rating processes. The estimated rating time to complete the rating for 109 questions was around 2 hours.

3 RESULT

3.1 Word Count and QPI Dimensions

Pearson correlations were conducted among the word count of the questions and the QPI ratings. Table 2 below shows that Word Count of the questions was not statistically significantly correlated with Overall, Scale, and Relevance.

On the contrary, Word Count is statistically significantly negatively correlated with Articulation ($r = -.201$, $n = 109$; $p = .036$). This result is plausible in that the longer the question is, its readability may decrease, and thus negatively impacts its articulation.

The “Overall” rating had statistically significantly positive correlations with all three dimensions of QPI, with Articulation ($r = .790$, $n = 109$, $p < .001$), Scale ($r = .455$, $n = 109$, $p < .001$), and Relevance ($r = .869$, $n = 109$, $p < .001$). This result is consistent with the hypothesis that the 3 dimensions of QPI are indicators to the overall productivity score.

Table 2: Pearson Correlation Matrix between Word Count and QPI Dimensions

		Word Count	Overall (Avg.)	Articulation (Avg.)	Scale (Avg.)	Relevance (Avg.)
Word Count	Pearson Correlation	1				
	Sig. (2-tailed)					
Overall (Avg.)	Pearson Correlation	-0.073	1			
	Sig. (2-tailed)	0.454				
Articulation (Avg.)	Pearson Correlation	-.201*	.790**	1		
	Sig. (2-tailed)	0.036	0.000			
Scale (Avg.)	Pearson Correlation	-0.077	.455**	.394**	1	
	Sig. (2-tailed)	0.425	0.000	0.000		
Relevance (Avg.)	Pearson Correlation	-0.030	.869**	.747**	.316**	1
	Sig. (2-tailed)	0.760	0.000	0.000	0.001	

*. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).

3.2 Variance among Raters

To examine the variance among different raters, four clusters of boxplots (See Figure 2.) were graphed. Figure 2 indicates that all four raters tend to agree on their ratings on the “Scale” dimension (the fourth cluster). Four raters presented noticeable differences between their ratings on “Articulation” and “Relevance”.

The strongest differences were shown in the “Relevance” scorings (the third cluster). It is reasonable that the differences on the “Relevance” scorings may be due to the difficulty of gauging how the presented questions relate to the larger class learning goals in disciplines that the raters were not familiar with. After all, raters may not be experts in all the class domains that the questions were extracted from.

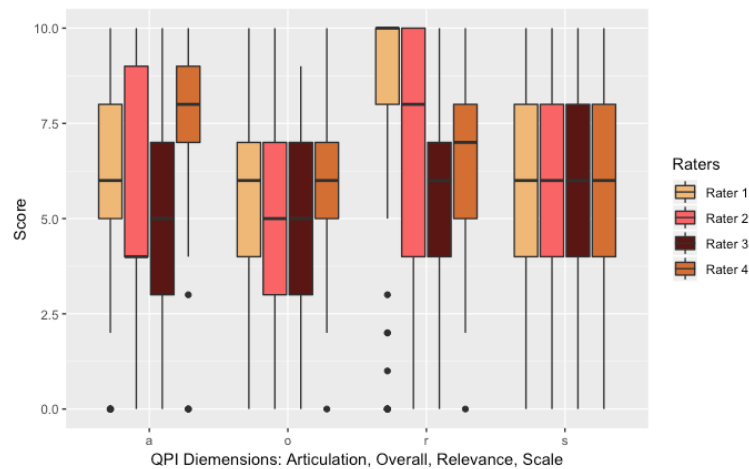


Figure 2: Boxplots of 4 Raters' QPI Scores

4 DISCUSSIONS AND FUTURE DIRECTIONS

In this exploratory analysis, we introduced QPI as a measure to capture productivity of student-generated questions. Results showed that the 3 dimensions, Relevance, Scale, and Articulation, of QPI explained a reasonable amount of variance in the overall productivity scores, which indicates that the QPI design is promising toward measuring question productivity. Meanwhile, Ratings on the “Articulation” and “Relevance” scores exhibited noticeable divergence among raters. Possible reason may include: 1) The definitions of the dimensions may need to be further clarified; 2) the small sample size of raters. Further investigation is needed to explore reasons behind.

In future analyses, we plan to include more linguistic features beyond just word count into the analyses. For instance, NLP measures representing uniqueness of the questions and types of questions are planned to be incorporated into the analyses. Furthermore, we also plan to connect and cross-validate the QPI measure with other existing critical thinking assessment in future iterations.

REFERENCES

- Chin, C., & Brown, D. E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*, 24(5), 521-549.
- Dillon, J. T. (1988). The remedial status of student questioning. *Journal of Curriculum studies*, 20(3), 197-210.
- Gresalfi, M., Martin, T., Hand, V., & Greeno, J. (2009). Constructing competence: An analysis of student participation in the activity systems of mathematics classrooms. *Educational studies in mathematics*, 70(1), 49-70.
- Good, T. L., Slavings, R. L., Harel, K. H., & Emerson, H. (1987). Student passivity: A study of question asking in K-12 classrooms. *Sociology of Education*, 181-199.
- Osborne, R., & Wittrock, M. (1985). The generative learning model and its implications for science education.
- Pizzini, E. L., & Shepardson, D. P. (1991). Student questioning in the presence of the teacher during problem solving in science. *School Science and Mathematics*, 91(8), 348-352.
- Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social-emotional role, and the classroom goal structure. *Journal of educational psychology*, 90(3), 528.
- UNESCO (1983). *Handbook for Science Teachers* (Paris: UNESCO).
- White, R., & Gunstone, R. (2014). *Probing understanding*. Routledge.
- Watts, M., Alsop, S., Gould, G., & Walsh, A. (1997). Prompting teachers' constructive reflection: Pupils' questions as critical incidents. *International Journal of Science Education*, 19(9), 1025-1037.

Identification of the vocabulary needed to foster self-expression in Syrian refugees in Lebanon

Victoria Abou-Khalil

Kyoto University

v.aboukhalil@gmail.com

Mohammad Nehal Hasnine

Kyoto University

Brendan Flanagan

Kyoto University

Hiroaki Ogata

Kyoto University

ABSTRACT: Syrian refugees in Lebanon are seeking migration to a third country with better living conditions. In order to do so, young Syrian refugees are learning English using their mobile phones. Social Emotional Learning (SEL) has shown to be effective when teaching English to people facing environmental stressors. SEL includes fostering the skill of self-expression. Current mobile applications and material available online do not provide the Syrian refugees in Lebanon with the vocabulary to articulate their thoughts due to their general content. In this work we aim to identify the vocabulary needed to improve the self-awareness skill. We propose to collect, analyze and discuss the English vocabulary that Syrian refugees in Lebanon would like to learn. To collect the vocabulary, we asked eight Syrian refugees in Lebanon to use SCROLL, a ubiquitous language learning environment and input the vocabulary they would like to learn during a period of ten days. The obtained words were grouped under different categories. The results inform us on the vocabulary that should be taught to Syrian refugees in Lebanon in order to allow them to express themselves better and nurture their self-awareness skill.

Keywords: Language learning, Social Emotional Learning, Refugees

1 INTRODUCTION

Syrian refugees in Lebanon form a vulnerable population. Their poor socio-economic situation and the low access to education and healthcare is pushing young refugees to seek immigration to a third country (VASYR, 2017). In order to do so, young Syrian refugees are learning English using their mobile phones. Learning English gives them more points and facilitates the approval of their immigration applications. It has been identified that trauma and disorder associated with immigration, family separations, poverty, discrimination, and cultural conflicts negatively impact English Language Learners (ELL) (Niehaus, K., & Adelson, J. L., 2013; 2014). One important implication is the importance of including Social Emotional Learning approaches (SEL) when teaching English to people facing those challenges. Syrian refugees in Lebanon face all of those environmental stressors and would benefit from SEL when learning English. SEL is the process through learners obtain and apply the knowledge, attitudes, and skills necessary to understand and manage emotions, set and achieve positive

goals, feel and show empathy for others, establish and maintain positive relationships, and make responsible decisions (Zins, J. E., & Elias, M. J., 2007). There are five main components of social emotional learning: self-awareness, self-management, social awareness, relationship skills, and responsible decision making (Elias, et al., 2017). Self-awareness refers to being aware of one's feelings, impact on others, and having a growth outlook. This includes learning to articulate one's feelings, mood, or energy level in order to proactively preempt escalating into destructive or disruptive behaviors. Current mobile applications and material available online do not provide the Syrian refugees in Lebanon with the vocabulary to articulate their thoughts as the content is not adapted to them. We propose to collect and analyze what would Syrian refugees in Lebanon like to express. The presence of online dictionaries and online language learning tools allows the users to research and translate words they would like to learn. The collection of all the words forms a vocabulary that can be valuable to inform designers of language learning tools which vocabulary to include when targeting Syrian refugees in Lebanon and refugees in similar life situations. To collect the corpus, we asked eight Syrian refugees in Lebanon to use SCROLL, a ubiquitous language learning environment, to input the vocabulary they would like to learn during a period of ten days. The obtained words were grouped under different categories. The results inform us on the important words that should be included while teaching vocabulary to Syrian refugees in Lebanon in order to allow them to express themselves better and nurture their self-awareness skill.

2 METHODS

2.1 Recruitment of the participants

Eight Syrian refugees were recruited from Syrian refugee communities residing in the Chouf region of Lebanon. We contacted the community leader to help us recruit participants between the ages of 14 and 25. The participants were required to have access to an internet-enabled smartphone and be learning English at the time of the study. We asked the community leader for an equal representation of genders. He was able to gather five men and only three women. The three recruited women did not own a smartphone but could access the smartphone of their brother, father, or husband.

2.2 SCROLL

During this study we use records from the SCROLL System (System for Capturing and Reminding Of Learning Log). Scroll is a digital record of what users have learned in daily life. It allows the learners to log the new words or sentences they learned along with photos, audios, videos and location (Ogata et al., 2011). SCROLL captures what learners are learning as well as its contextual data. The users are then reminded of what they learned in the right place and the right time. Moreover, learners receive personalized quizzes to fortify the learning. Figure 1 is a screenshot from the SCROLL system that shows a log inserted by a learner. The learner appended a picture when creating the log. An English translation of the word قدم (foot) is automatically provided to the learner, and the time is automatically registered. Currently SCROLL has 1705 registered users and contains around 30380 logs (Ogata et al., 2018).

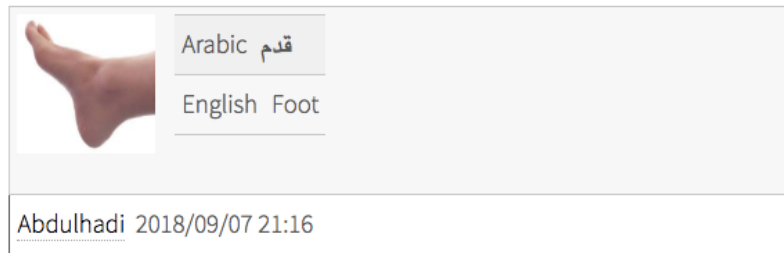


Figure 1: Screenshot of a log added by the participant in the SCROLL system

2.3 Analysis of the data

To have a clear view about which vocabulary the Syrian refugees in Lebanon need in order to express themselves better, the authors sorted the words that the participants added to the system into the following categories: *everyday objects*, *transportation*, *humans*, *everyday activities*, *environment*, *emotions*, *nature*, *abstract concepts*, *immigration*, *conversation*. Knowing the different categories, and the distribution of words in every category would allow us to generate similar vocabulary and teach it to the Syrian refugees in Lebanon. Each category represents the vocabulary that the participants looked up on. Below is the description of the different categories.

Everyday objects: words describing common objects used or present in one's everyday life e.g.: table, pen.

Transportation: words relating to modes of transportation e.g.: bus, airport.

Humans: words describing humans, or body parts and professions.: e.g.: father, hand, baker.

Everyday activities: words or sentences describing common activities e.g.: tooth extraction.

Emotions: adjectives or nouns that describe emotions or feelings: pain, tasty, special.

Abstract concepts: words describing non-material concepts or ideas e.g.: culture, invention.

Immigration: words relating to immigration or the immigration's procedure e.g. court, passport, embassy.

Conversation: words or sentences that are usually used to conduct a conversation e.g.: How are you? Where do you live?.

Environment: words that relate to or describe the participant surroundings e.g.: camp, house, mosque, hospital.

Nature: words that describe nature or animals e.g.: forest, shark.

3 RESULTS AND DISCUSSION

3.1 Vocabulary added to SCROLL by the Syrian refugees in Lebanon

The participants created 282 logs in the SCROLL system. The participants were interested primarily in learning how to describe *everyday objects* (148 words added to the system). The *humans* category was the second most populated one with 41 words. Around one third of the words present in the *humans* category describe family members and relationships. The *nature* category contained 21 words. Through the 17 words in the category *environment*, the participants were interested in learning how to describe their surroundings: *refugee camp*, *house*, *hospital*, etc. 12 words were related to *immigration* and the participants were

interested in learning words that would allow them to describe the processes they are going through to immigrate. The *transportation* category contained nine words and the *abstract concepts* category contained eight. Finally, the two least populated categories are *emotions* and *everyday activities*. A summary of the distribution of the words by category is shown in figure 2.

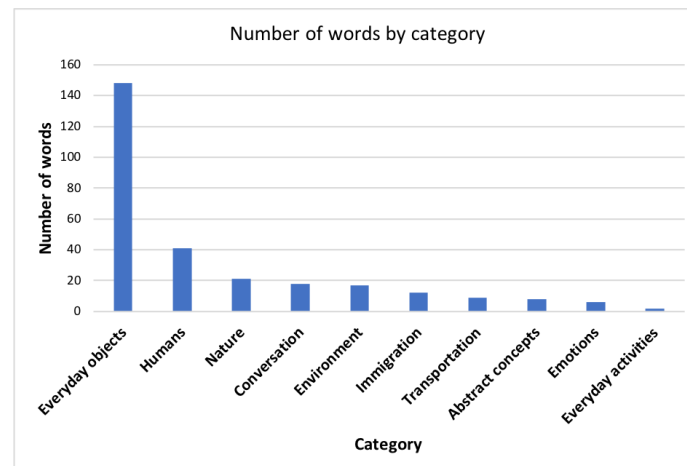


Figure 2: The distribution of the words added by the participants under different categories

3.2 Vocabulary to foster self-expression in Syrian refugees in Lebanon

Most language learners would be interested in learning the vocabulary to talk about their lives and their surroundings, however, some words added by the participants are specific to refugees in Lebanon. The category related to immigration is specific to refugees or immigrants applying for asylum or immigration to another country. The category immigration contains the words: *court, passport, embassy, migration, adviser, judge, denied*, etc. It is important for Syrian refugees in Lebanon to be able to communicate their immigration situation in English and learn the vocabulary related to it as this would help their immigration application. Moreover, the participants were interested in describing their particular environment and added the words: *refugee camp, mosque, church*, that represent their surrounding in Lebanon. Finally, *emotions* and *abstract concepts* were rarely present in the words that the participants wished to learn.

3.3 Opportunities

SEL aims at helping learners foster social-awareness through understanding others. The vocabulary that the Syrian refugees chose to learn give us an insight on their environment, but also on the what is not present in their environment. For example, the *transportation* category included different modes of transportation: *bus, car, plane, boat, bicycle, motorbike*, however, *train* or *metro* were not included. One reason could be that trains and metros are not present in Lebanon. Identifying words that are *missing* from the vocabulary can offer opportunities for teaching about other countries and cultures.

3.4 Limitations of the study

The relatively low number of Syrian refugees in Lebanon that participated in this study gives access to a limited number of words. Even though, the access to this population is difficult, gathering more participants would provide a valuable insight to the diversity of the vocabulary that different learners wish to learn. The authors classified the words into categories. However, those categories might be chosen differently and some words could fit into different categories. Moreover, due to the limited information available about the student intention when they choose to learn a word it is sometimes challenging to know to which category the word should belong e.g.: the word travel could belong to the category *transportation* or to the category *immigration*.

4 CONCLUSION

Syrian refugees in Lebanon form a vulnerable population and are seeking immigration to a third country that provides them better living conditions. In order to do so, knowing English is a valuable asset and young Syrian refugees are learning it using their mobile phones. As a part of an SEL approach to teaching English using mobile phones, we collected vocabulary needed by Syrian refugees in Lebanon. The collection of the vocabulary informs us on the words that Syrian refugees in Lebanon would like to learn to be able to express themselves better in English. We grouped the words into ten different categories: *everyday objects, transportation, humans, everyday activities, environment, emotions, nature, abstract concepts, immigration, and conversation*. We identified vocabulary that should be included in the online content targeting specifically Syrian refugees in Lebanon or refugees in similar situations.

REFERENCES

- Vulnerability Assessment of Syrian refugees in Lebanon (2017): *Vulnerability Assessment of Syrian refugees in Lebanon 2017*. Retrieved from <https://data2.unhcr.org/en/documents/download/61312>.
- Niehaus, K., & Adelson, J. L. (2013). Self-concept and native language background: A study of measurement invariance and cross-group comparisons in third grade. *Journal of Educational Psychology, 105*(1), 226.
- Niehaus, K., & Adelson, J. L. (2014). School support, parental involvement, and academic and social-emotional outcomes for English language learners. *American Educational Research Journal, 51*(4), 810-844.
- Zins, J. E., & Elias, M. J. (2007). Social and emotional learning: Promoting the development of all students. *Journal of Educational and Psychological Consultation, 17*(2-3), 233-255.
- Elias, M. J., Zins, J. E., Weissberg, R. P., Frey, K. S., Greenberg, M. T., Haynes, N. M., & Shriver, T. P. (1997). *Promoting social and emotional learning: Guidelines for educators*. Ascd.
- Ogata, H., Li, M., Hou, B., Uosaki, N., El-Bishouty, M. M., & Yano, Y. (2011). SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research & Practice in Technology Enhanced Learning, 6*(2).

Ogata, H., Uosaki, N., Mouri, K., Hasnine M.N., Abou-Khalil V., & Flanagan B. (2018) SCROLL Dataset in the Context of Ubiquitous Language Learning, Workshop Proceedings of the 26th International Conference on Computer in Education, 418-423, November 26-30, Philippines.

Temporally Rich Features Capture Variable Performance Associated with Elementary Students' Lower Math Self-concept

Shamya Karumbaiah¹, Jaclyn Ocumpaugh¹, Matthew J. Labrum², and Ryan S. Baker¹

¹ University of Pennsylvania, Philadelphia, PA USA, ² Imagine Learning, Provo, UT USA

shamya, ojaclyn@upenn.edu, matthew.labrum@imaginelearning.com, rybaker@upenn.edu

ABSTRACT: A better understanding of the relationship between self-concept in mathematics and fine-grained behavior logs from students' interactions with intelligent tutoring systems (ITSs) could help researchers better understand self-concept, which in turn could lead to improved designs in interventions intended to improve a student's self-concept. Yet, to date, learning analytics researchers have had only limited success in modeling these relationships. This exploratory study uses correlation mining to investigate the potential of temporally-rich features to capture variance in student performance. Results suggest detecting such inconsistencies in students' performance may be key to developing more robust models to infer self-concept, as well as to understanding how differences in it emerge among elementary students.

Keywords: self-efficacy, math identity, math performance, time-series, non-cognitive skill

1 INTRODUCTION

Self-concept has been shown to predict student achievement (Spinath et al., 2006) and appears to be conceptually related to other non-cognitive and motivational constructs including expectancy and value (Eccles and Wigfield, 1995), intrinsic interest (Gottfried, 1985) and intrinsic values (Pintrich et al., 1993). As summarized in Marsh et al. (2005), domain-specific self-concept (e.g., mathematics self-concept) shows developmental patterns of decline from early childhood to adolescence and then increases during early adulthood. The rich environment provided by many intelligent tutoring systems (ITSs), which can provide fine-grained assessment of behavior, affect, and cognition, might prove fruitful for better understanding the developmental changes in this construct. Indeed, Bernacki et al. (2015) found learners' self-efficacy (self-beliefs related to a specific task) varied reliably over an algebra unit in Cognitive Tutor in their investigation of the stability of self-efficacy and its relationship to problem-solving performance.

However, there has been limited research into how self-efficacy and self-concept relate to the types of behavior seen in intelligent tutors and other online learning systems. In part, this may be because self-concept survey instruments were deliberately devised to diverge from straight-forward measures of performance like grades or test scores (e.g., Gottfried, 1985). This may explain the limited success in correlating survey measures of self-concept to behaviors in ITSs. For example, Slater et al. (2018) used correlation mining to examine 185 features of student interactions (aggregated at a monthly and yearly level) with an ITS and found only 18 that showed a significant relationship with self-concept. That said, there were interesting patterns in these results, suggesting that future work should focus on engineering features that better captured variance in the students'

performance. Building on the promise of these findings, we present an exploratory paper where we investigate time-series features—which capture more complex and fine-grained characteristics of students' interactions with an ITS over long periods of time. As our data show, these temporally-rich features seem better able to capture the kind of variation needed to characterize math self-concept.

2 METHOD

2.1 Reasoning Mind Foundations

Reasoning Mind *Foundations* (Figure 1) is an intelligent tutoring system for math learning used by over 100,000 elementary school students in the United States including rural, urban, and suburban schools. Many of these students are from traditionally underrepresented populations. In this blended environment, students learn through self-paced problem solving, mathematical games, and interactive explanations. There are three main types of problems in this ITS based on the increasing levels of difficulty: 1) A-level problems on fundamental skills; 2) B-level (optional) problems on a combination of skills; and 3) C-level (optional) problems on higher order thinking skills.



Figure 1: *Foundations'* home screen.

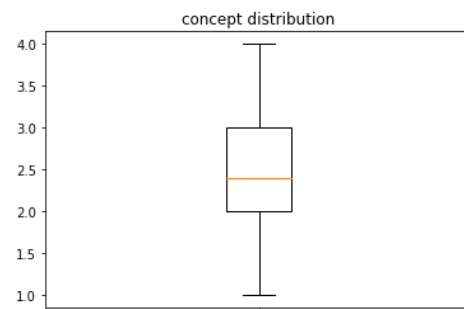


Figure 2: Distribution of average math self-concept scores of *Foundations* students.

2.2 Data

This study examines data collected in the 2017--18 school year from 2nd--5th grade students in Texas classrooms who interacted with *Foundations* as part of their regular mathematics instruction.

Math Self-concept Data - Surveys adapted from Marsh et al. (2005) (e.g., *Math just isn't my thing. Some topics in math are just so hard that I know from the start I'll never understand them.*) were administered at the end of the academic year 2017--2018 to collect 1566 students' self-reports on math self-concept using five items, each on a four-point Likert scale (Cronbach $\alpha = 0.74$; mean = 2.42 (SD = 0.81); Figure 2).

Times-series Feature Extraction - Student performance on A-, B-, and C-level problems in *Foundations* was aggregated to engineer day-level sequences (time series) of the number of correct responses for each student (Figure 3). The average length of the times series is 70 days (SD = 35 days) spread out through the course of two semesters, with considerable differences in daily averages and standard deviations for the number of correct responses in different levels (A-level =

2

4.4 (4.79), B-level = 1.08 (3.42), C-level = 0.61 (2.51)). Time-series features were extracted with a Python package called *tsfresh* (Christ et al., 2018) which, in addition to providing high-level features describing meta-information of the time series, also calculates a comprehensive set of feature mappings that characterizes them. Across the three time series of the problem level performances, a total of 2382 features were extracted.

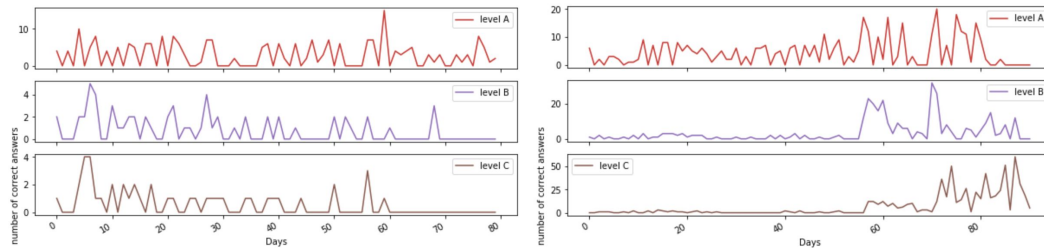


Figure 3: Example time series of the number of correct answers of level A, B, and C problems of a student with high post self-concept (left) and a student with low post self-concept (right). Note the difference in y-axis scales in the subplots.

3 RESULTS

A correlation analysis of the extracted time-series features with math self-concept yielded 113 significant correlations after using the Benjamini and Hochberg (1995) post-hoc procedure to control for false discovery (at $\alpha = 0.05$). All of the significant features were negatively associated with self-concept. Table 1 presents problem-type level aggregate correlations based on ten of the most frequent types of time-series features from the full correlation list¹. As the descriptions in the table show, most of these features capture the greater variances in the correctness of math problems, a result comparable with Slater et al.'s (2018) findings that inconsistent performance is associated with lower math self-concept. For instance, the *change quantile* category, which captures the point-by-point change (mean and variance) in an ordered list of a student's daily math performance, contributes 31 of the 133 significant correlations found.

Other notable patterns emerge from these results. First, feature types (as described in Table 1) differ with respect to the problem levels that are most relevant (i.e., A-, B-, and C-level problems). All three levels emerge as significant for some feature categories, notably the *change quantile* category where A- and B-level problems each contribute 10 significant features and C-level problems contribute 11 more. In contrast, other feature categories are only significant for features derived from C-level problems; this includes *symmetry looking* and *CWT coefficients* (14 features each), as well as *approximate entropy* (5 features), *FFT coefficients* (3 features), and *Ratio beyond R Sigma* (3 features)---a greater description of these features is given in Table 1. Overall, the majority of the significant correlations (78/113) involved features extracted from C-level problem performance, which may be related to the fact that students have more autonomy in their decision to complete C-level problems (i.e., they are usually optional). There are also differences, within certain feature categories, in how those features are operationalized. For example, *change quantile* features

¹ A full list of all significant correlations after controlling for false discovery is available at [link redacted for review]

involving A- and B-level problems rely on earlier quantile ranges of the time series (i.e., 0.1--0.6 and 0.2--0.8), while those involving C-level problems make greater use of later quantiles of the time series (i.e., 0.0--1.0 and 0.4--0.8). That is, for the easier problem sets (A- and B-level problems), self-concept is negatively associated with variance that occurs in the lower ranges of the number of correct responses, while with the more challenging (and optional) C-level problems, the variance is more likely to be relevant in the higher ranges of the number of correct responses.

Table 1: Features categories of the time series (TS) with significant correlation with self-concept.

Category	Counts			mean R	Description ²
	A	B	C		
Change quantiles	10	10	11	-0.16	Average, absolute value of consecutive changes inside different quantile ranges of TS. Higher change corresponds to a higher inconsistency in student performance.
Symmetry looking	0	0	14	-0.14	Boolean value specifying if the distribution symmetric: Is $ \text{mean}(\text{TS}) - \text{median}(\text{TS}) < r * (\text{max}(\text{TS}) - \text{min}(\text{TS}))$? If symmetric, the values in TS occur regularly.
CWT coefficients	0	0	14	-0.15	Coefficients of the continuous wavelet transform for the Ricker wavelet; these give information about the amplitude of TS and how that amplitude varies over time.
Quantile	3	3	4	-0.16	q th quantile value of TS.
Descriptive statistics	1	1	5	-0.15	Mean, median, maximum, mean absolute change, and if $\text{SD}(\text{TS}) > r * (\text{max}(\text{TS}) - \text{min}(\text{TS}))$?
Number of peaks	0	2	5	-0.17	Number of peaks in TS subsequences, where peaks signify a sudden increase in the number of correct responses.
Approximate entropy	0	0	5	-0.18	Amount of irregularity and unpredictability of fluctuations in different TS ranges. Higher entropy corresponds to lower regularity in student performance.
Sum of recurring data points	1	1	1	-0.15	Sum of values in TS that are present more than once (e.g., if a student had 12 correct responses on two or more different days, 12 would be counted towards this feature).
FFT coefficients	0	0	3	-0.15	Fourier coefficients of the one-dimensional discrete Fourier Transform; these give information about the underlying periods in the TS (e.g., weekly or monthly periods of performance).
Ratio beyond R sigma	0	0	3	-0.15	Ratio of student performance values that are farther than $R * \text{SD}(\text{TS})$ away from the $\text{mean}(\text{TS})$, so that larger values show greater variance.

4 DISCUSSION

Learning analytics research has had limited success in modelling asynchronous survey measures of non-cognitive constructs (e.g., self-concept) from student behaviors in an ITS, compared to other

² The detailed feature description is at https://tsfresh.readthedocs.io/en/v0.11.1/text/list_of_features.html

constructs. In contrast to previous research in this area, our study explores time-series features which capture finer-grained and longer-term temporal variations in student performance than often captured in the feature engineering process. We have demonstrated that these time-series features look promising as indicators of elementary students' math self-concept. The primary takeaway of our analysis is the negative relationship between inconsistencies in student performance and their domain-specific self-concept. Thus, an immediate implication of this work is to further examine the rate at which the different problems are being introduced to students in the ITS. C-level problems, which are more difficult and optional, show the most promise for predicting self-concept. Particularly, there seems to be a delay in attempting harder (C-level) math problems (see Figure 3) among students with lower self-concept. This calls into attention the intentionality of the students' behaviors, suggesting that one way to test for self-concept is to give students optional opportunities for more challenging practice. Our future work will use these findings to better design new features, to develop a predictive model of math self-concept for students using *Foundations*, and eventually to design interventions for students with low self-concept.

REFERENCES

- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2):122.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- Bernacki, M. L., Nokes-Malach, T. J., and Aleven, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, 10(1):99–117.
- Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing*, 307:72–77.
- Eccles, J. S. and Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, 21(3):215–225.
- Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology*, 77(6):631.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., and Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2):397–416.
- Pintrich, P. R., Smith, D. A., Garcia, T., and McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3):801–813.
- Slater, S., Ocumpaugh, J., Baker, R., Li, J., and Labrum, M. (2018). Identifying changes in math identity through adaptive learning systems use. In *Proceedings of the 26th International Conference on Computers in Education*.
- Spinath, B., Spinath, F. M., Harlaar, N., and Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, 34(4):363–374.

Strength of Noncognitive Predictors Varies By Course Discipline

James Cunningham

Arizona State University

jim.cunningham@asu.edu

Phil Arcuria

Arizona State University

phil.arcuria@asu.edu

Yuan Wang

Arizona State University

elle.wang@asu.edu

William Morgan

Arizona State University

wsmorgan@asu.edu

ABSTRACT: The present workshop paper employs both logistic and linear regression to measure how noncognitive variables affect grade outcomes differently across ten different course disciplines in an online higher education context. Models were trained on noncognitive data from over 16,000 students predicting more than 97,000 grade outcomes. Both logistic and linear models demonstrate considerable variability in the prediction power of noncognitive factors depending on the course discipline targeted. Six of the noncognitive factors considered stood out as being very predictive across several course disciplines. Time management was the most predictive of the noncognitive variables considered.

Keywords: Learning analytics, noncognitive factors, higher education, predictive modeling

1 INTRODUCTION

In higher education, the effects of noncognitive factors on student success is an active area of research; however, a majority of the focus in recent years has been on the broad effects of noncognitive qualities on outcomes in higher education such as admission, retention, and graduation rates (Akos & Kretchmar, 2017; Bowman, Miller, Woosley, Maxwell, & Kolze, 2018; Niessen, Meijer, & Tendeiro, 2017). In addition, there has been much attention in the past decade looking into using noncognitive factors in the admissions process of universities in addition or instead of the normal mix of admission considerations such as high school grade point average and standardized test scores (Komarraju, Ramsey, & Rinella, 2013; Sedlacek, 2011; Sternberg, Bonney, Gabora, & Merrifield, 2012). Another trend in the use of noncognitive data has been to predict retention and persistence in college level studies (Robertson & Taylor, 2009; Maddi, Matthew, Kelly, Villarreal, & White, 2012). There are also a few studies that seek to correlate noncognitive skills with academic performance measured as general overall GPA or grade outcomes in specific courses. Most of these projects have smaller sample sizes although there are some meta-analyses that examine trends across these studies (Poropat, 2009;

Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Richardson, Abraham, & Bond, 2012). The examination of which noncognitive factors are the important in differing types of courses has not been the focus of research up to this point. There are several reasons why addressing this gap could be beneficial. One reason is that it seems reasonable that characteristics or personality traits that make one more successful in one area of study such as language arts could be less helpful in an analytical area of study such as chemistry. At the same time, identifying noncognitive factors that are important across many course disciplines could also be an important part of understanding the role of these variables in the overall success of students in higher education.

The goal of this study is to examine the impacts of these variables at the course discipline level to look for differences in the effects of noncognitive factors in various course contexts and to look for factors that could have broader impacts across the curriculum.

2 THE DATA

The source of noncognitive data for this study was online personality assessment given to online undergraduate students in an orientation course at a large southwestern university in the U.S. between July 2016 and June 2018. The purpose of using this test in the orientation course was to provide students with fun, non-threatening content that could be used to teach these students how to use the learning management system (LMS). When students received the results from the test, they were asked to submit reflections how the scores did or did not describe what they thought of themselves, discuss their results in the discussion board, and follow other assignments that led them through the basic tasks of using the LMS.

The personality test used in the orientation course consisted of just under 200 questions and ordering exercises designed to measure a variety of 21st century skills, motivators, behaviors, and social-emotional indicators (The Indigo Project, 2018a). The Indigo Project is part of a certain genre of personality tests marketed to K-12 and higher education institutions as a tool to understand strengths and weaknesses of learners and as means to individualize instruction (The Indigo Project, 2018b). Companies marketing these assessments tout high reliability coefficients, but offer scant evidence of validity (TTI Success Insights, 2018; Price, 2015). Despite our concerns about the validity of these tests, we were curious if perhaps the assessment was capturing actual noncognitive data that could be predictive. Each assessment generated scores on 127 noncognitive markers. These markers were divided into seven groups based on the theoretical background from whence the markers were derived (see Table 1). The four major theoretical bases for these markers were: DISC theory based on the work of William Moulton Marston and Walter Vernon Clarke (Marston, 1928), an undisclosed motivational theory for motivator markers, the “system of axiology” developed by Robert S. Hartman (Hartman, 1961), and 21st century skills (the origin of the theory behind the selection and description of the 21st century skills was not disclosed by either The Indigo Project or TTI Success Insights).

Table 1

Seven categories of noncognitive Indigo markers.

Prefix	Number of Markers.	Theoretical Basis
B.	8	DISC
M.	6	Motivational
S25.	25	21 st Century Skills
H.	12	Hartman
BX.	14	DISC
HX.	62	Hartman

Noncognitive scores were obtained for 16,141 students and matched to 97,511 course grade outcomes. The number of courses taken by the students in our sample ranged from 1 to 38 courses ($M = 6.04$; $SD = 4.70$).

Three controls were used in both the logistic and linear models to account for factors that might confound grade outcomes: course difficulty, faculty difficulty, and course load. Course difficulty was controlled with a course difficulty index that used the average grades of a particular course over the previous two years. This index ranged from 1.56 to 4.29 ($M = 3.06$; $SD = 0.31$). The faculty difficulty index averaged the grades given by a particular faculty member over the previous two years. This index ranged from 1.54 to 4.33 ($M = 3.03$; $SD = 0.37$). Course load reflected the number of units attempted by the student in the same semester that the course grade was recorded and ranged from 1 to 27 ($M = 10.35$; $SD = 3.82$).

Outcome variables for the prediction models were based on course grades achieved by students who had taken the Indigo personality assessment. These outcomes were of two types: binary course outcomes based on mastery (grades of B- or higher) for the logistic models and grades used as a continuous measure on a scale of 0 – 4.33 for the linear models.

To specify course discipline groupings, courses were divided by course code prefixes. The top 10 course disciplines that contained the most students who had taken the Indigo Assessment were: astronomy (ASB), biology (BIO), communications (COM), criminal justice (CRJ), computer science (CSE), English (ENG), history (HST), mathematics (MAT), psychology (PSY), and sociology (SOC).

3 METHODOLOGY

In what follows, we describe two methods used to predict grade outcomes grouped by course discipline using the noncognitive variables. The first method is a logistic regression model using the regularization method, elastic net. The second method is multiple linear regression using as the predictors noncognitive variables from the logistic regressions with the highest coefficients to create

parsimonious linear models. These linear models measure the most predictive noncognitive variables with grades as a continuous outcome. The results of the linear models were then compared and contrasted to the results of the logistic regression models.

3.1 Questions for Analysis

We present our results focusing on providing answers to the following questions:

- Q1: In which course disciplines are noncognitive variables most predictive?
- Q2: Which noncognitive variables are most predictive across course disciplines?
- Q3: How much variability in predictive power of noncognitive factors exists between course disciplines?

3.2 Logistic Regression Models

The outcome variable for the prediction models was obtained from the final grade data of the online courses taken by the students who completed the Indigo Assessment. For the logistic regression models, grades were expressed at the binary outcome "did or did not achieve mastery in the course." In terms of grade points, the cutoff in the logistic regressions was 3.0 ("B-") on a scale of 0.0 to 4.3. To build these models, we chose the machine learning regularization and variable selection method, elastic net. Elastic net addresses two issues in our data, a large number of noncognitive variables (high dimensionality) and correlations between the factors (multicollinearity). Elastic net uses the ridge penalty to add a small amount of bias to the models to reduce the problematic variance that occurs when multicollinearity is present (Marquardt & Snee, 1975). This is combined with lasso regression that performs variable selection (Tibshirani, 1996). Thus, elastic net can be expressed as

$$\lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|)$$

where:

- λ represents the overall penalty imposed on the fitted coefficients
 - α represents elastic-net penalty and controls the mix of L_1 and L_2 penalties. It has values that range between 1 (ridge) and 0 (lasso).
 - $|\beta_j|$ represents the L_1 penalty (lasso)
 - β_j^2 represents the L_2 penalty (ridge)
- (Zou & Hastie, 2005).

K-fold cross validation was used to estimate out-of-sample fit for final model selection by partitioning each sample into k equal size subsamples and then systematically rotating through each subsample as the test data and the remaining subsamples as training data (Kohavi, 1995). Sample data was split into 80% training and 20% testing sample sets. All outcomes reported for the logistic regression models reflect results achieved on test datasets.

3.3 Linear Regression Models

The multiple linear regression models used the 10 most predictive noncognitive variables from the logistic regression models based on the size of the coefficients from the outputs of those models. These ten variables were then used to build parsimonious multiple linear regression models to predict all grades in each course discipline using course grade outcomes as a continuous variable.

4 RESULTS

Figure 1 and Table 1 show the predictive power of noncognitive variables relative to different course disciplines. The logistic models were most effective in predicting mastery in biology, criminal justice, and psychology while the linear models most accurately predicted the specific grades of biology and English thus answering Q1. The weakest predictions in mastery were in computer science and mathematics and the weakest predictions in specific grades were in mathematics and psychology. Figure 2 shows a scatterplot demonstrating variability of the relative strengths of predictions in logistic versus linear regressions by course discipline in answer to Q3.

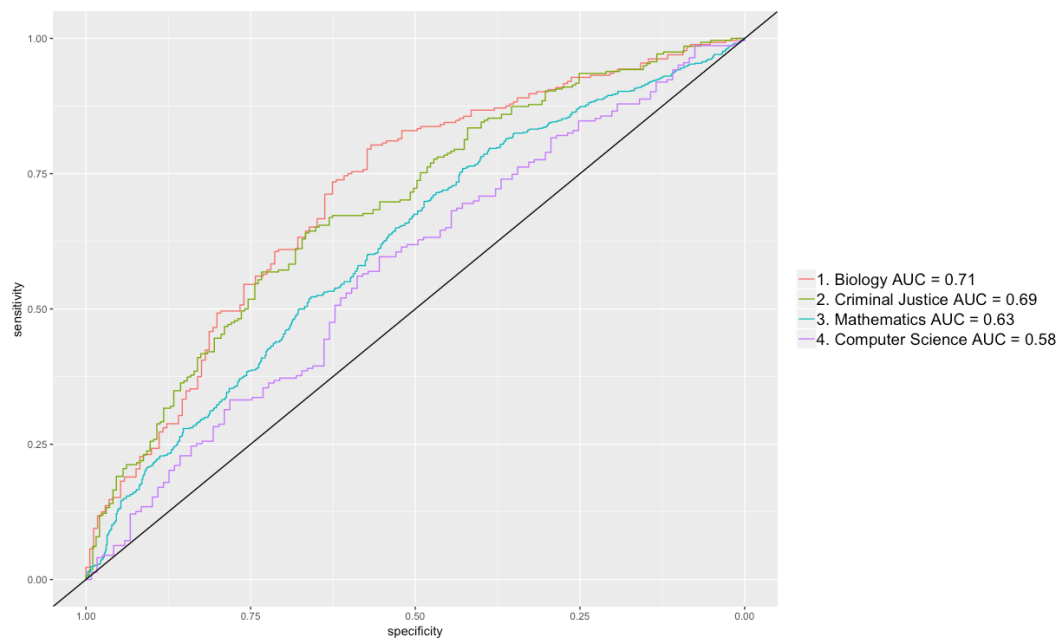
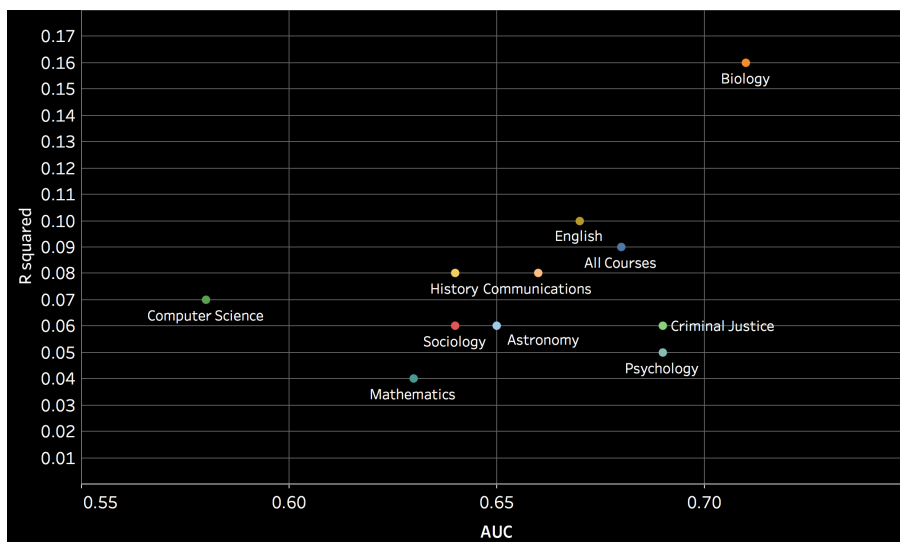


Figure 1: The receiver operating characteristic (ROC) curves for the two strongest predicting models and the two weakest models demonstrates the variability in prediction power of noncognitive variables across different course disciplines.

Table 1: Comparison of the prediction accuracy of noncognitive logistic and linear models relative to grade outcomes achieved by students who took the Indigo personality assessment.

Course Discipline	AUC	R ²
All Courses	0.68	0.09
Astronomy	0.65	0.06
Biology	0.71	0.16
Communication	0.66	0.08
Criminal Justice	0.69	0.06
Computer Science	0.58	0.07
English	0.67	0.10
History	0.64	0.08
Mathematics	0.63	0.04
Psychology	0.69	0.05
Sociology	0.64	0.06

**Figure 2: Comparing relative strengths of noncognitive predictions in logistic and linear regressions.**

In answer to Q2, 6 of 127 noncognitive variables stood out as most predictive across all the course disciplines: Time Priority Management, Resiliency Skill, Compliance Adapted, Conflict Management, Leadership, and Negotiation. These occurred three or more times in the top ten most predictive factors in the logistic regressions (Figure 2). Time Priority Management was especially strong occurring in the top ten list for 7 out of 10 course disciplines.

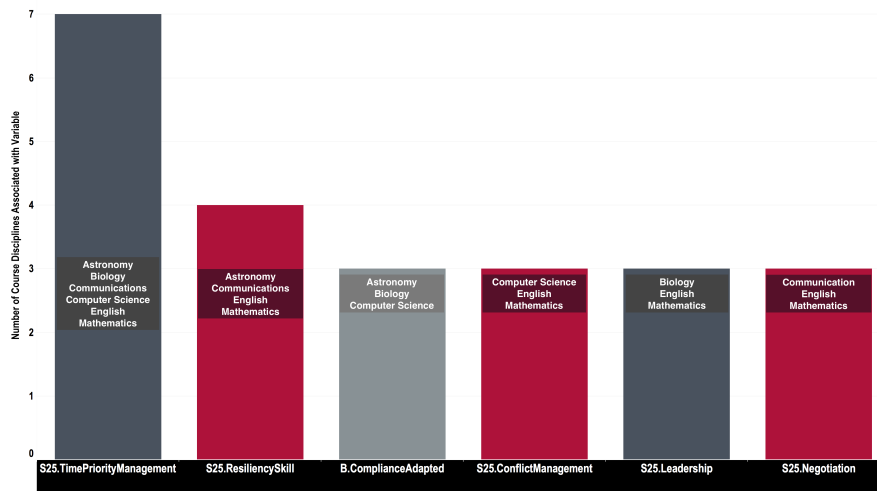


Figure 3: Noncognitive factors that occurred in the top ten most predictive factors in logistic regressions for at least three course disciplines (grey denotes coefficients in the positive direction, red denotes coefficients that are negative, Compliance Adapted had two positive and one negative coefficient).

5 DISCUSSION & FUTURE WORK

Our results show that noncognitive factors do not predict grade outcomes uniformly across course disciplines in higher education. In our study using the Indigo personality assessment, noncognitive factors were especially strong in predicting mastery and grade outcomes in biology courses. However, noncognitive models were much less successful in the prediction of outcomes in courses such as computer science and mathematics. Biology was a standout among the other course disciplines in both the logistic and linear regressions. However, this may have been partly due to the course difficulty index in those courses being especially predictive as well.

The mix of noncognitive factors making up the ten most predictive variables in the logistic models was diverse with 67 of the 127 variables occurring at least once in the top ten most predictive variables for the 10 course disciplines. However, among the noncognitive variables, scores for Time Priority Management were consistently highly predictive occurring in seven of the top ten lists. Time Priority Management was also strongly predictive in the linear models with some of the largest coefficients of the various noncognitive variables. This is consistent with other studies that have found time management to be among the most important noncognitive predictors (MacCann, Forgarty, & Roberts, 2012; Bowman, Miller, Woosley, Maxwell, & Kolze, 2018). Resiliency Skill was also strongly predictive occurring in the top ten list for four of the ten courses. Surprisingly, coefficients for this factor were consistently negative. This was puzzling and deserves further investigation.

6 CONCLUSION

As the impact of noncognitive factors on student success in higher education gets more attention, it is important to realize that impact of noncognitive factors varies considerably between differing course disciplines. Although we found that the noncognitive data of the Indigo is correlated with course outcomes at both the mastery and specific grade level, there was substantial variation between courses. This is may be an important point to consider when planning interventions to increase student success based on noncognitive variables. Careful design may be required to get the desired results.

REFERENCES

- Akos, P., & Kretchmar, J. (2017). Investigating Grit at a Non-Cognitive Predictor of College Success. *The Review of Higher Education*, 40(2), 163-186. doi:10.1353/rhe.2017.0000
- Bowman, N. A., Miller, A., Woosley, S., Maxwell, N. P., & Kolze, M. (2018). Understanding the link between noncognitive attributes and college retention. *Research in Higher Education*. doi:10.1007/s11162-018-9508-0
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*.
- MacCann, C., Fogarty, G. J., & Roberts, R. D. (2012). Strategies for success in education: Time management is more important for part-time than full-time community college students. *Learning and Individual Differences*, 22, 618-623. doi:10.1016/j.lindif.2011.09.015
- Marston, W. M. (1928). *Emotions of normal people*. New York, NY: Kegan Paul, Trench, Trubner &.
- Niessen, A. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, 106, 183-189.
- Price, L. A. (2015). *DISC instrument validation study* (pp. 1-16, Tech.). Boardman, OH: PeopleKeys: Institute for Motivational Living.
- The Indigo Project. (2018a). *Indigo: Assessment validity, reliability, measurement variables, competitor landscape, customers, and partners* [Pamphlet]. The Indigo Project.
- The Indigo Project. (2018b). Indigo assessments. Retrieved from <http://www.indigoproject.org/indigo-assessments/>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1), 267-288.
- TTI Success Insights. (2018). A global leader of reliability and accuracy. Retrieved from <https://www.ttisuccessinsights.com/research/>

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2), 301-320. doi:10.1111/j.1467-9868.2005.00503.x

Learning Analytics to Share and Reuse Authentic Learning Experiences in a Seamless Learning Environment

Mohammad Nehal Hasnine^{1*}, Hiroaki Ogata¹, Gokhan Akcapinar¹², Kousuke Mouri³ and Noriko Uosaki⁴

¹ Kyoto University

² Hacettepe University

³ Tokyo University of Agriculture and Technology

⁴ Osaka University

* hasnine.mohammadnehal.5z@kyoto-u.ac.jp

ABSTRACT: Authentic learning experiences are considered to be a rich source for learning foreign vocabulary. Prevalent learning theories support the idea of learning from others' authentic experiences. This study aims at developing a learning analytics solution to deliver the right authentic learning contents created by one learner to others in a seamless learning environment. Therefore, a conceptual framework is proposed to close the loops in the missing components of the current learning analytics framework. Data is captured and recorded centrally via a context-aware ubiquitous learning system which is a key component of a learning analytics framework. k-Nearest Neighbor (kNN) based profiling is used to measure the similarity of learners' profiles. Authentic learning contents are shared and reused through re-logging function. This paper also discusses how two previously developed tools, namely learning log navigator and a three-layer architecture for mapping learners' knowledge-level, are adapted to enhance the performance of the conceptual framework.

Keywords: Authentic learning experiences, informal learning, learning analytics, seamless learning, share and reuse, ubiquitous logs, vocabulary learning.

1 INTRODUCTION

Authentic learning is referred to real life learning. According to Steve Revington, this kind of learning style encourages students to create a tangible and useful product to be shared with their world (Revington, 2016). In language learning, the idea of using authentic learning materials to teach foreign vocabulary has a long history. Authentic learning materials considered to be a rich source of target language input (Duda & Tyne, 2010). In Glimore's viewpoint, authentic materials and authenticity in foreign language learning opposes contrived materials of traditional textbooks which typically display a meager and frequently distorted sample of the target language while authentic materials offer a much richer source of input for learners (Gilmore, 2007). Tomlinson's viewpoint (Tomlinson, 2008) is equally severe. He claimed that various English language teaching materials, particularly global course-books currently make a significant contribution to the failure of many learners of English to acquire even basic competence in English and to the failure of most of them to develop the ability to use it successfully. Therefore, exposure to authentic language learning contents is crucial for language development, particularly for foreign vocabulary. However, the debate over the role of authenticity, as well as what it means to be authentic, has become increasingly sophisticated and complex over the

years and now embraces research from a wide variety of fields including discourse and conversational analysis, pragmatics, cross-cultural studies, sociolinguistics, ethnology, second language acquisition, cognitive and social psychology, learner autonomy, information and communication technology, motivation research and materials development (Gilmore, 2007).

While authentic learning experiences are crucial in foreign vocabulary learning, how ubiquitous and sensing technologies facilitate in it? During the twentieth century, massive development of sensing technologies made it possible to attain contextual information, such as people, date, precise time, location, theme etc. regarding the learners' usage of various ubiquitous technologies, for example, lifelong camera, multi-touch interface, Wi-Fi, RFID, GPS, wearable smart tracker, and Bluetooth. Using such functions, learners authentic learning experiences can be tracked and recorded. For instance, an international student, upon experiencing a culturally authentic content, records it in the system with its context information (memo), picture/video/voice-data, together with its textual information. Ubiquitous functionalities automatically track the learning location, time, and place etc. By this, a vast amount of rich educational big data on authentic learning experiences can be captured. Now the questions arise, how this vast amount of educational big data to be dealt to improve next-generation education? Also, can learning analytics provide solutions to sharing and reusing those captured authentic learning experiences (i.e. logs) among a community of language learners having similar learning interest in the right way at the right time and place?

The present study aimed at discovering a ubiquitous dataset to innovate a learning analytics solution for sharing and reusing authentic learning contents in a seamless learning environment. The contributions of this paper are- to begin with, an authentic learning experience is defined. We defined an authentic experience is comprised of the word, it's representative picture/video/voice-data, contextual information (i.e. memo), and translation data together with the time (when) and location (where) information. These parameters are must for a content to be treated as an authentic learning experience. These authentic learning experiences are collected using a context-aware ubiquitous language learning system (Ogata et al., 2011) that supports both formal and informal learning seamlessly. After that, a conceptual framework is proposed for putting the missing components together to close the loops (i.e. learning analytics cycle) in their learning analytics framework (Flanagan & Ogata, 2018). The conceptual framework is to impliment Kolb's experimental learning theory(Kolb, 2014) using learning analytics. Finally, an extended objective of this work is to establish a personalized learning path to optimize vocabulary learning. Moreover, this study also aimed to increase foreign language learners' motivation and engagement with location-based learning system.

2 LEARN FROM OTHERS EXPERIENCES: KOLB'S VIEWPOINT AND LEARNING ANALYTICS CYCLES

Kolb's experimental learning theory is a renown learning theory which is widely recognized and accepted not just for language learning but to for learning-focused curriculum development and instructional design (Kolb, 2014). In light of the increasingly competitive and complex learning environments, Kolb's experimental learning theory has been used to carry out many studies over the last two decades. Kolb's experimental learning theory comprises of four phases, each of which involves using different processes to acquire and use information and skills. The four phases are

Concrete Experience (CE), Reflective Observation (RO), Abstract Conceptualization (AC), and Active Experimentation (AE). In CE stage of learning, learners actively experience an activity in real-life or in the classroom. The RO happens when the learner consciously reflects back on that experience. In the AC stage, learners attempt to conceptualize a theory or model of what is observed. Finally, in the AE stage, learners try to plan how to test a model or theory or plan for a forthcoming experience. The assumption of Kolb's learning theory, later simplified by Knutson (Knutson, 2003a) as- we seldom learn from experience unless we assess the experience, assigning our own meaning in terms of our own goals, aims, ambitions, and expectations. From these processes come the insights, the discoveries, and understanding. The pieces fall into place, and the experience takes on added meaning in relation to other experiences.

In relation to Kolb's theory, Doug Clow's idea is a great example for learning analytics practice on a base of established learning theory. Doug Clow (Clow, 2012) has shown how learning analytics cycles overlap with Kolb's theory. In his viewpoint, the learning analytics cycle begins with learners (Phase 1 in Fig.3) of formal and informal learning. The next step to it is, to generate and capture of data (Phase 2 in Fig.3) about or by the learners. For example, demographic, login, clickstream, location etc. about a potential learner. Some of it can be generated automatically while some require a large multidisciplinary team to expend significant effort. The third step is the processing of this data into metrics or analytics (Phase 3 in Fig.3), which provide some insight (Phase 4 in Fig.3) into the learning process. Phase 4 includes visualizations, dashboards, personalized feedback tools where the comparisons of outcome can be measured. We yield this conclusion that, in order to implement a learning theory at an individual level, learning analytics cycles can facilitate by providing information on learners' activities, conception, and actions which, in future leads to propose rich feedback or intervention mechanism.

This paper aimed at closing the loops in the current learning analytics framework (Flanagan & Ogata, 2018). Precisely speaking, in the current setup, elements of the first two phases learners (which is the Phase 1) and dataset (Phase 2) exist. However, the analytics (Phase 3) and interventions (Phase 4) do not exist. Therefore, with this study, we aimed at establishing the relationship between the phases to close the loops.

3 PREVIOUS WORKS TO SUPPORT THE CONCEPTUAL FRAMEWORK

3.1 Location-based sharing

Learning Log Navigator (hereafter LLN), an analytics tool is developed as a function of the system to analyze authentic learning experiences. LLN aims to automatically provide appropriate learning experiences in accordance with the individual learner (Mouri, Ogata, & Liu, 2014). In the LLN system, sharing authentic learning contents happens in three conditions (shown in Fig.1). LLN's analytics first records concrete experiences in the system, and then using location data, it guides learners to the authentic learning environment. Next, when learners reflect on themselves based on reflective observation, the system provides experiences that others learners have learned in the authentic learning environment. Finally, it supports conceptualizing from concrete experiences by recommendation and analysis of learning logs. LLN system recommends authentic learning contents

that were created in the nearest location of a learner's current location. This calculation is based on the latitude and longitude information of a learner. The number of contents to be recommended, which is 10, is controlled by the system. If the number of the recommended task is too much; learners will be confused because it is difficult for learners to select the most accurate one (Mouri, Ogata, & Liu, 2014). If a learner wishes to browse more contents, he/she needs to change his/her learning location (that is, to extend the limit of distance). Learners can learn tasks using the mobile device through these whole flow (Mouri, Ogata, & Liu, 2014).

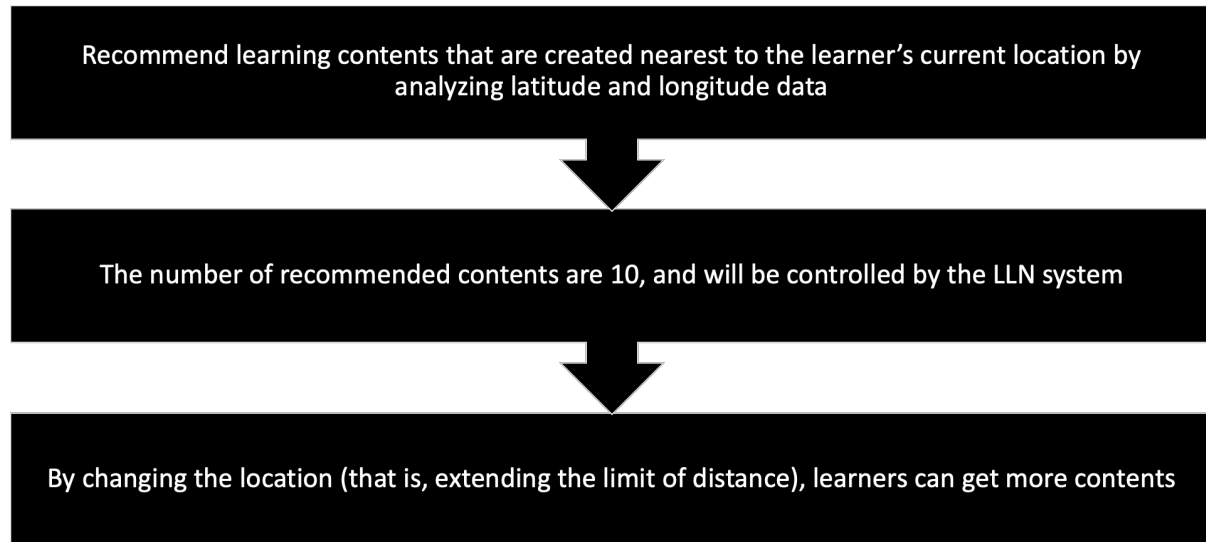


Figure 1: Authentic learning experiences sharing using LLN

3.2 A three-layer architecture

To support LLN, a three-layer architecture (visualized in Fig.2) is developed that identifies learners and knowledge or knowledge and location by using network graph. This architecture visualization can be widened by linking one's own learning logs to the knowledge learned by doing tasks (Mouri, Ogata, Uosaki, & Liu, 2014). The architecture is defined as a three-layer architecture where the upper layer contains each author in order to confirm position of own or other learners, the intermediate layer contains the knowledge that learners learned, and the lowest layer contains data such as location and time (Mouri, Ogata, Uosaki, et al., 2014). In order to realize spatiotemporal visualization of our learning logs, nodes on the intermediate layer are linked to the nodes on the lowest layer (Mouri, Ogata, Uosaki, et al., 2014).

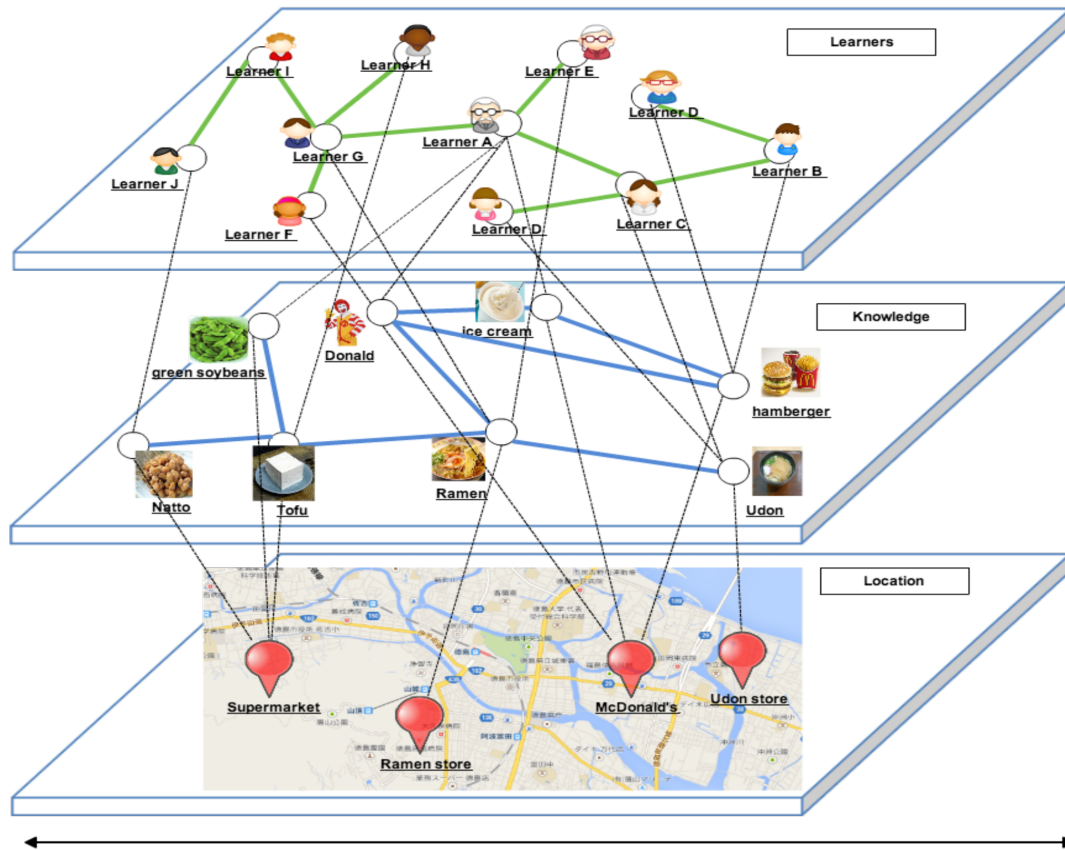


Figure 2: Authentic learning experiences sharing using LLN (Mouri, Ogata, Uosaki, et al., 2014)

4 LEARNING ANALYTICS FOR SHARING AND REUSING CONTENTS

4.1 A Conceptual Framework: Design

Doug Clow's paper has shown that learning analytics cycles overlap with the four phases of Kolb's experimental learning theory. Which mean, learning analytics cycles can be utilized to implement this theory. Based on that, a conceptual framework is proposed based on learning analytics cycle for sharing and reusing authentic learning experiences in a seamless learning environment. The system is a key research component of our learning analytics framework where learners get the opportunity to learn in various learning environments regardless of place or time. The framework is designed in the way that it provides an interface between integrated production and research systems to allow user authentication, information, and learning analytics results to be seamlessly transferred between systems (Flanagan & Ogata, 2018). As stated earlier, at present, Phase 1(users) and Phase 2 (dataset) exit without having any analytic connection. Therefore, with the proposed conceptual framework, this study aimed at closing the loop for missing parts of the learning analytics cycle. Fig.3 shows the conceptual framework that is proposed to support this study. This study is carried out precisely for Phase 3 (analytics) and Phase 4 (interventions).

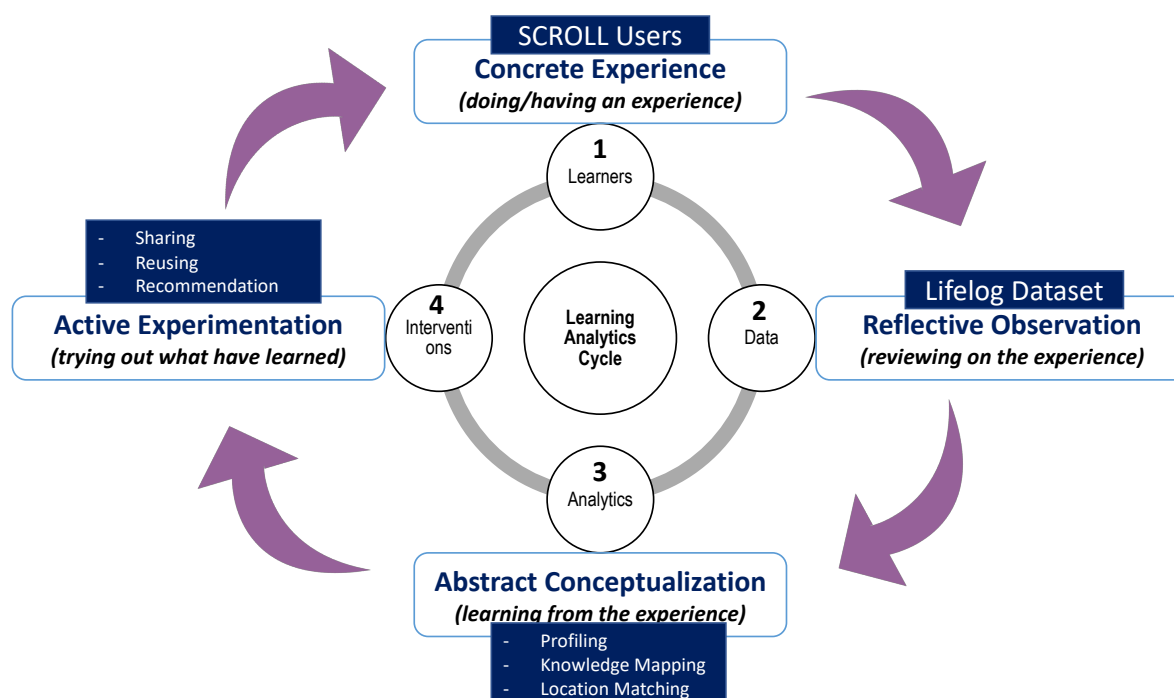


Figure 3: The proposed conceptual framework for closing the loop

4.2 Implement the Framework: Phase-by-Phase

4.2.1 System-used for capturing authentic learning experiences

This present study defines an authentic experience as it comprised of the word, it's representative picture/video/voice-data, contextual information (i.e. memo), and translation data together with the time (when) and location (where) information. These parameters are must for a content to be treated as an authentic learning experience. Foreign language learners' authentic learning experiences are captured as concrete experiences using a context-aware ubiquitous language learning system that offers seamless learning of multiple foreign languages. The system is based on LORE (Log-Organize-Recall-Evaluate) model by which intends to automatically extract meaningful knowledge from past learning experiences so that that information can serve as the guide for future learning (Ogata et al., 2011). The ubiquitous functionalities of the system are capable of recording learners' concrete experiences (such as the geolocation information, vocabulary knowledge, quiz, learning context, contextual image information etc.) as ubiquitous learning logs into its learning record store (LRS)(Hasnine et al., 2018). Learners' activities in the system are recorded precisely in the LRS as xAPI (Experience API) statements. The system also captures various educational big data through its five key features namely, authentic learning logs capturing, lifelogging, share and reuse logs, automatic quiz generation based on past learning logs, and an e-Book reader. Most of the logs are created for either Japanese or English language's vocabulary learning. A learning log-tracking dashboard, shown in Fig.3, is developed where learners can track their formal (eBook-based learning activities) and informal (real-life) learning activities. A time-map is also developed for chronological tracking of

learning contents with learning location and time. This client-server application runs on different platforms including Android mobile phones, PC and general mobile phones (Hasnine, 2018).

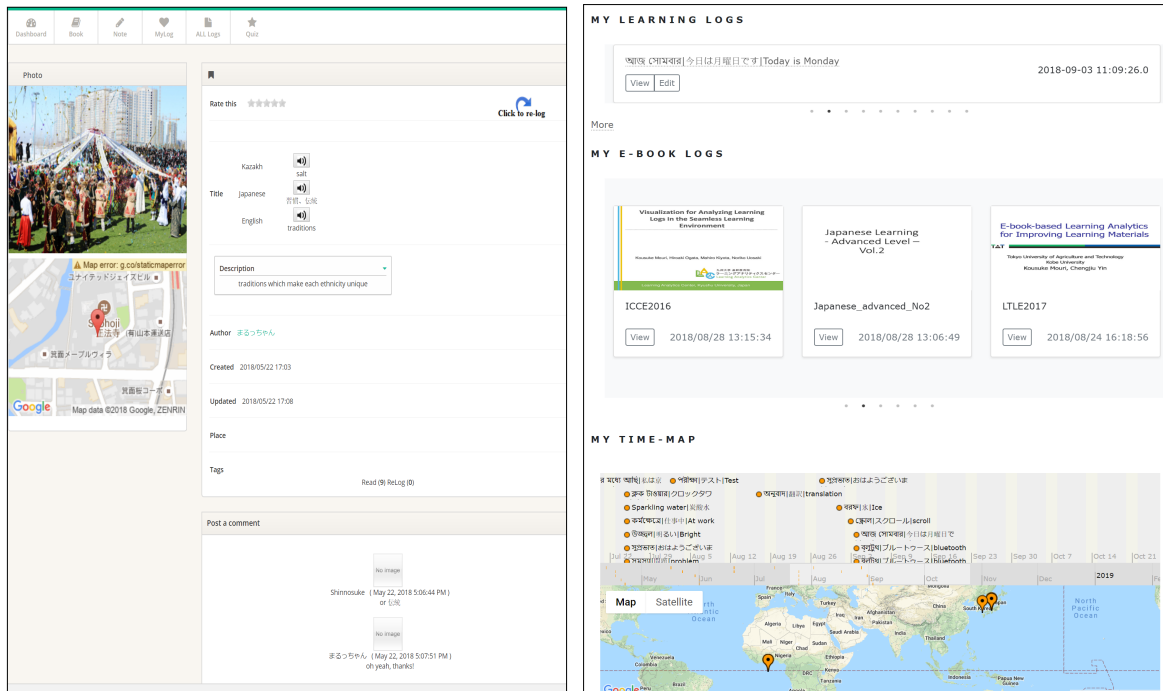


Figure 4: An authentic learning log (on left) and seamless activity tracking function (on right)

4.2.2 Phase 2 (Dataset)

Data is collected primarily from learners of foreign languages, tourists, and international students looking for jobs in Japan. As of now, over 30000 learning logs and over 400000 quiz logs are captured using the system from over 1700 learners from over thirty-nationalities (Ogata et al., 2018). The dataset contains vocabulary Information (words that a learner wishes to learn in a specific context), learner profile Information (such as, name, age, gender, education etc.), cultural Information (information about nationality, social interaction level etc.), study place-time-location (Geo-locational information, place-details, and study time etc.), past Knowledge: Vocabulary that learners have previously acquired (i.e. learning history), and contextual Image Information(unique image features (color, shape, object etc.) that may describe the learning context and/or the word itself. Note that, not all of those logs are counted to be authentic experiences. Logs that meet the definition of authentic learning experiences are counted as authentic learning experiences. It can be reported that, the majority of learners in the system have registered themselves as Chinese, English and Japanese languages as their default languages.

4.2.3 Phase 3 (Analytics)

kNN (K-Nearest Neighbor) is a renowned machine learning algorithm that can find clusters of similar learners based on common properties, and make predictions using the average features of top-k nearest neighbors. kNN is an intuitive and easy-to-implement algorithm. This algorithm is used to develop partner-matching, Facebook's friend-matching and friend-suggestion, Amazon's interest-based book recommendation etc. Aiming to find learners from one's neighborhood having similar demography, we looked at a learner's neighborhood and measured the similarity in profiles. This technique is adapted to improve the matching accuracy and efficiency. In order to run the kNN

algorithm, a metrics with certain parameters are required. Hence, a metrics is formed for profile matching using kNN algorithm. Table 1 briefly summarized the metrics.

Table 1: Data for KNN-based Profiling

Value	Description
User id	Universally Unique Identifier (UUID) of a user
Age	Age of a user
Nationality	Nationality of a user
Target language(s)	Language(s) that a learner registered as language(s) of interest
Past knowledge	Vocabulary that learners have previously acquired (i.e. the learning history)
Knowledge level	Current knowledge level in a target language. For instance, JLPT3 refers to a learner's Japanese language level as intermediate
Learning location	Latitude and longitude data of a learner's learning locations
Time	Time of each session
Learning context	The context that a learner created an authentic learning experience
Image	Image that is uploaded by the learner in the process of capturing a log
Video	Video that is uploaded by the learner in the process of capturing a log

Note that, a log containing either image and/or video clip is treated as an authentic learning experience. We plan to integrate this analysis with our previous developments. As stated earlier, previously we developed two tools namely, LLN system and the three-layer architecture. These tools are used for determining the right person to deliver the right content by analyzing the current learning location and level of knowledge.

4.2.4 Phase 4 (Intervention): Analytics Dashboard and Re-log Function

Re-log function (located on the top-right corner in Fig.4(left)) is developed for reusing and sharing of an authentic learning experience. This function enables a learner to reuse an authentic learning material created by other learner in the system. A prototype dashboard is underway as an intervention which is the first step to an analytics dashboard. In this dashboard, learners will be able to interact with his/her peers of similar interest in foreign language learning.

5 CONCLUSION

The assumption of Kolb's learning theory is that we seldom learn from experience unless we assess the experience, assigning our own meaning in terms of our own goals, aims, ambitions, and expectations. From these processes come the insights, the discoveries, and understanding. The pieces fall into place, and the experience takes on added meaning in relation to other experiences (Knutson, 2003b). From this viewpoint of Kolb's theory, this research initiated to contribute to the learning analytics research community by proposing an analytic method for sharing and reusing one learner's authentic learning experiences among others when learning foreign vocabulary in a seamless learning environment. A conceptual framework is proposed to connect the missing components of our learning analytics framework. With this conceptual framework, this study aimed at connecting learning

analytics phases, namely learners (Phase 1), dataset (Phase 2), analytics (Phase 3) and interventions (Phase 4). For profiling, kNN-based profile matching algorithm used. In order to enhance the performance of the model, two previously developed analytics tools namely, Learning Log Navigator (LLN) was developed that can analyze learners' authentic learning experiences based on learning location, and a three-layer architecture to map a learner with his/her knowledge and learning location, are adapted.

For future works, an experiment is designed to evaluate the proposed framework. The experiment is aimed to find answers to the research questions- First, the right way to deliver the right content at the right time and place; Second, find out the right learner to whom an authentic learning experience can be shared; Third, establish personalized learning path to optimize vocabulary learning. Moreover, the experiment will analyze whether learners' engagement with the system and motivation is increased. Two groups of subjects are planned to recruit for this experiment. One of which is tourists visiting in the city. And, another group is the combination of international students studying Japanese and Japanese students learning the Spanish language. The experiment is designed to guide international students to experience real-life Japanese learning activities by the word they learned either in class or out of the class.

ACKNOWLEDGMENTS

This work is supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304 and Start-up Grant-in-Aid Number 18H05745.

REFERENCES

- Clow, D. (2012). The learning analytics cycle: closing the loop effectively. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 134–138). ACM.
- Duda, R., & Tyne, H. (2010). Authenticity and Autonomy in Language Learning. *Bulletin Suisse de Linguistique Appliquée*, 92, 86–106.
- Flanagan, B., & Ogata, H. (2018). Learning analytics infrastructure for seamless learning. *Companion Proceedings 8th International Conference on Learning Analytics & Knowledge (LAK18)*, 1-6.
- Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2), 97–118.
- Hasnine, M. N. (2018). A Distributional Semantics Model for Image Recommendation using Learning Analytics. In *Early Career Workshop Proceedings of the 26th International Conference on Computer in Education* (pp. 10–12). Manila, Philippines.
- Hasnine, M. N., Mouri, K., Flanagan, B., Akcapinar, G., Uosaki, N., & Ogata, H. (2018). Image Recommendation for Informal Vocabulary Learning in a Context-aware Learning Environment. In *Proceedings of the 26th International Conference on Computer in Education* (pp. 669–674). Manila, Philippines.
- Knutson, S. (2003a). Experiential learning in second-language classrooms. *TESL Canada Journal*, 20(2), 52–64.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Mouri, K., Ogata, H., & Liu, S. (2014). Learning Log Navigator: Supporting Authentic Learning Using Ubiquitous Learning Logs. *Santiago, Chile*, 39.

- Mouri, K., Ogata, H., Uosaki, N., & Liu, S. (2014). Visualization for analyzing ubiquitous learning logs. In Proceedings of the 22nd International Conference on Computers in Education (ICCE 2014) (pp. 461–470).
- Ogata, H., Li, M., Hou, B., Uosaki, N., EL-BISHOUTY, M. M., & Yano, Y. (2011). SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research & Practice in Technology Enhanced Learning*, 6(2).
- Ogata, H., Uosaki, N., Mouri, K., Hasnine, M. N., Abou-Khalil, V., & Flanagan, B. (2018). SCROLL Dataset in the Context of Ubiquitous Language Learning. In Workshop Proceedings of the 26th International Conference on Computer in Education (pp. 418-423), Manila, Philippines.
- Revington, S. (2016). Authentic Learning - What is it? Retrieved January 24, 2019, from <http://authenticlearning.weebly.com/>
- Tomlinson, B. (2008). Language acquisition and language learning materials. *English Language Learning Materials: A Critical Review*, 3–13.

Analytics of the relationship between quiz scores and reading behaviors in face-to-face courses

Tsubasa Minematsu, Atsushi Shimada, and Rin-ichiro Taniguchi

Kyushu University, Japan

minematsu@limu.ait.kyushu-u.ac.jp

ABSTRACT: In this paper, we visualize the relationship between quiz scores and learning behaviors of students from clickstream data in face-to-face courses. Analysis of the learning behaviors can help to understand the reasons why students get quiz scores. In addition, the reasons are useful for teachers to support students. To represent the learning behaviors, we define an action score based on student's reading behaviors in digital textbooks and actions done by students during learning. In this paper, we analyzed 1,914,680 clickstream data collected by learning management systems in Kyushu University. In our analysis, we focused on students who got lower quiz scores than the average scores. Our investigation showed the existence of some students who got lower quiz scores and larger action scores. This implies that these students were active for learning in the course, however, could not get good quiz scores. We also investigated high performance students. The learning scores of these students were independent of their quiz scores.

Keywords: Learning analytics, Clickstream, Visualization, Face-to-face course.

1 INTRODUCTION

Learning supports for students have taken a key role for improving education. Quiz scores from examinations are useful to select strategies of the learning support because the quiz scores reflect how well the students understand contents of courses. Teachers may decide how to support the students based on the quiz scores. However, the support based on only the quiz scores does not take care of motivations of the students. For example, a student 1 gets the same quiz score as a student 2, and their quiz scores are lower than those of other students. However, the student 1 studies harder than the other students do. In this case, the appropriate support for the student 1 may be different from the student 2.

Analysis of student learning behaviors is essential to understand various kinds of information of the students such as motivations and learning path (Davis, Chen, Hauff, & Houben, 2016), and it helps teachers to support students. To understand the learning behaviors, recent works in learning analytics analyze clickstream data and learning activity logs collected from e-learning systems such as massive open online courses (MOOCs) (You, 2016) and M2B systems in Kyushu University (Ogata, et al., 2015). Especially, Kyushu University collects clickstream data in face-to-face courses. In the face-to-face courses, teachers can directly manage the lectures, and the students can be strongly affected from teachers' instruction. Therefore, the educational data collected in the face-to-face courses also are affected by teachers unlike educational data collected from online learning systems such as MOOCs. Analysis of the educational data in the face-to-face courses has been starting such as cross analysis over different courses (Shimada & Konomi, 2017), change detection for learning behaviors (Shimada,

Taniguchi, Okubo, Konomi, & Ogata, 2018), the visualization (Okubo, Shimada, Taniguchi, & Konomi, 2017) and student performance predictions (Okubo, Takayoshi, Shimada, & Hiroaki, 2017).

In this paper, we visualize the relationship between the quiz scores and the learning behaviors from clickstream data in face-to-face courses in order to investigate existence of earnest students who get low quiz scores. This visualization may help teachers to design approaches of learning supports for students. For example, we expect that teachers distinguish such earnest students from the other students, and the teachers apply appropriate supports to them. For this purpose, the learning behaviors are represented based on the major pages frequently read by students and the number of actions such as creating new highlights and taking notes except actions of page accesses. We define an action scores for measuring the learning behaviors, and then we plot the quiz scores and the action scores to investigate the relationship.

2 VISUALIZATION METHOD

2.1 Data preparation

We choose event logs in Kyushu University from dataset provided by LAK19 data challenge organizers because few event logs in Kyoto University and Asia University are available during each lecture time. In our visualization of this paper, we use only event logs from students who attended to all lectures in a course. These event logs were collected from first year students of arts and sciences in information science courses that were face-to-face courses. In these courses, the students read a digital textbook on information science using e-learning systems. The event logs contain information of actions done by students and page numbers when the actions occurred. In addition, we can use quiz scores of the students who took an exam after the course ends. Please refer to the details of LAK19 data challenge dataset (LAK19 Data Challenge, 2018) (Flanagan & Ogata, 2017) (Ogata, et al., 2015).

We use features based on pages read by students at each time for our visualization. However, the event logs do not contain information of pages students read at all time because an event is stored when a student does an action. Therefore, we interpolate the information of pages during no collection of event logs. For example, there are no events between 14:30 and 14:35 in a student, and then an event occurred in page 36 at 14:36. In this case, we consider the student read page 36 between 14:30 and 14:35. We applied the interpolation method to event logs every 10 seconds to obtained page numbers read by students.

2.2 Feature extraction : Action score

In this paper, we focus on the learning behaviors that are relevant to lectures. The learning behaviors are represented based on whether or not students read digital textbooks based on teachers' instructions and the number of actions done by students. The former means that students do not perform no relevant actions such as sleeping when the students read the same page as a teacher. The latter means that students are active when the students do more actions than the other students do.

We define an action scores for measuring the learning behaviors. Let s_i is an action score of student i . The action score s_i is computed as follows:

$$s_i = p_i(1 + a_i) \quad (1)$$

, where p_i is a page score and a_i is an action score of student i . The page score represents student's reading behaviors in digital textbooks based on actions of page transitions. The activity score means learning behaviors based on active actions except page transitions. An example of the action is the creation and deletion of digital textbook highlights.

A page score p_i is computed based on the differences between pages read by student i and major pages frequently read by the students. We assume that teachers' instruction is one of factors to affect students' page transitions. Under this assumption, a student should read the similar pages that the other students read. In our experiments, we found pages read by the students frequently at each time. We consider that the information of the major pages implies the baseline of students' reading behaviors. To understand students' reading behaviors, we compute a histogram of pages read by students in each lecture. Let h_L and $h_L(t, r)$ be a two-dimension histogram in lecture L and the number of students reading page r at time t . The page score p_i is computed as follows:

$$p_i = \frac{1}{\sum_L \sum_t \sum_r h_L(t, r)} \sum_L \sum_t h_L(t, r_i(t)) \quad (2)$$

, where $r_i(t)$ is a page number read by student i at time t . The time range of t is from time at which the lecture L started to time at which the lecture L ended. The page score becomes larger when the student read the same page as the other students.

An activity score of a student is computed from the number of actions done by the student except ones for page accesses such as "NEXT" and "PAGE JUMP". In fact, we use actions of "ADD BOOKMARK", "ADD MARKER", "ADD MEMO", "CHANGE MEMO", "DELETE BOOKMARK", "DELETE MARKER", and "DELETE_MEMO" for computing the activity scores. Instead of using the number of the actions directly, we quantize them based on 25, 50, and 75 percentile points in this paper. This procedure is similar to computation of active learner points (Okubo, Takayoshi, Shimada, & Hiroaki, 2017). We summarize the quantization method in Table 1.

2.3 Visualization

We visualize the relationship between the quiz scores and the action scores using scatter plots. In our preliminary experiments, we found that the page scores were biased depending on courses as shown in Figure 1. Figure 1 illustrates a page score distribution for each course. The distribution color corresponds to the course ID. Action scores may be biased due to this bias. In order to remove the bias between courses, we subtract maximum values of the action scores from all of the action scores in each course, and then we use $\exp(s_i)$ as action scores. The quiz scores are normalized between 0 and 1 in order to adjust its scale to a scale of the action scores.

Table 1: Quantization of the number of actions.

Activity score	0	1/3	2/3	1
#actions	Lower 25 percentile point	Lower 50 percentile point	Lower 75 percentile point	otherwise

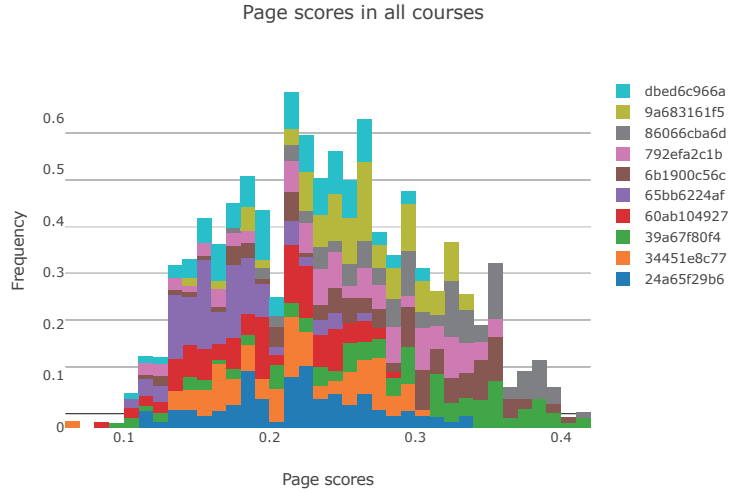


Figure 1: Distribution of page scores. Different colors in the distribution mean different course IDs.

3 RESULT

In Figure 2, we show a result of our visualization method applied to event logs in Kyushu University. Figure 2 contains two histograms of the quiz scores and the action scores. Two dash horizontal lines in Figure 2 mean quiz scores $\mu - \sigma$ and $\mu - 2\sigma$, where μ and σ are a mean and a standard deviation of the quiz scores. Three vertical dash lines in Figure 2 are arranged at equal intervals between a maximum value and a minimum value of the action scores.

We investigated students whose quiz scores are less than $\mu - \sigma$ and $\mu - 2\sigma$. In this investigation, we found some students who have larger action scores and lower quiz scores. These students are distributed in the bottom-right of Figure 2. We also investigated students whose quiz scores were more than μ . We understood some students could obtain higher quiz scores even if their action scores were smaller.

We confirmed the difference of event logs between two students whose quiz scores were lower than $\mu - \sigma$. Figure 3 shows a normalized histogram of pages read by students at each time in course dbed6c966a. The two lines in Figure 3 mean pages read by the two students at each time, and the green and yellow line correspond to the student with the largest action scores and the smallest action score. According to Figure 3, their reading behaviors were different between the two students even if the quiz scores were lower than the mean of quiz scores. This confirmation shows one possibility that we can distinguish the characteristics of students thanks to our visualization even if some students have the same quiz score.

4 DISCUSSION

Our visualization method may help teachers to decide how to support students after lectures. We expect that a teacher may provide different supports for the students in the bottom-right and the bottom-left of Figure 2. For example, teachers can provide supplementary teaching materials for understanding details of contents in the textbook to students in the bottom-right, or teachers may give a short summary to students in the bottom-left. We consider these students in the bottom-right

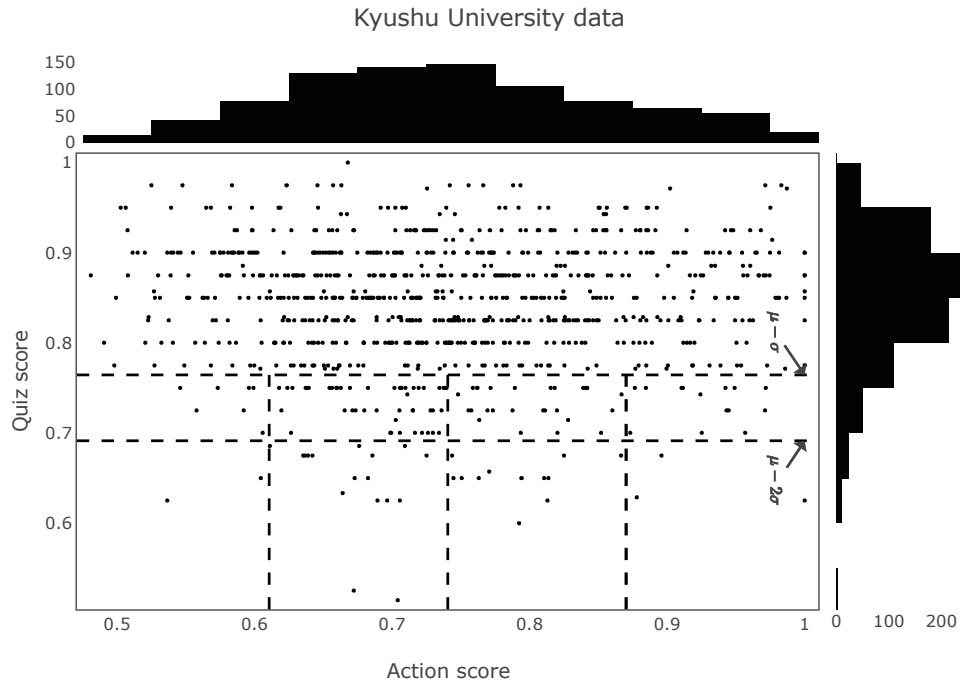


Figure 3: Distribution of quiz scores and action scores.

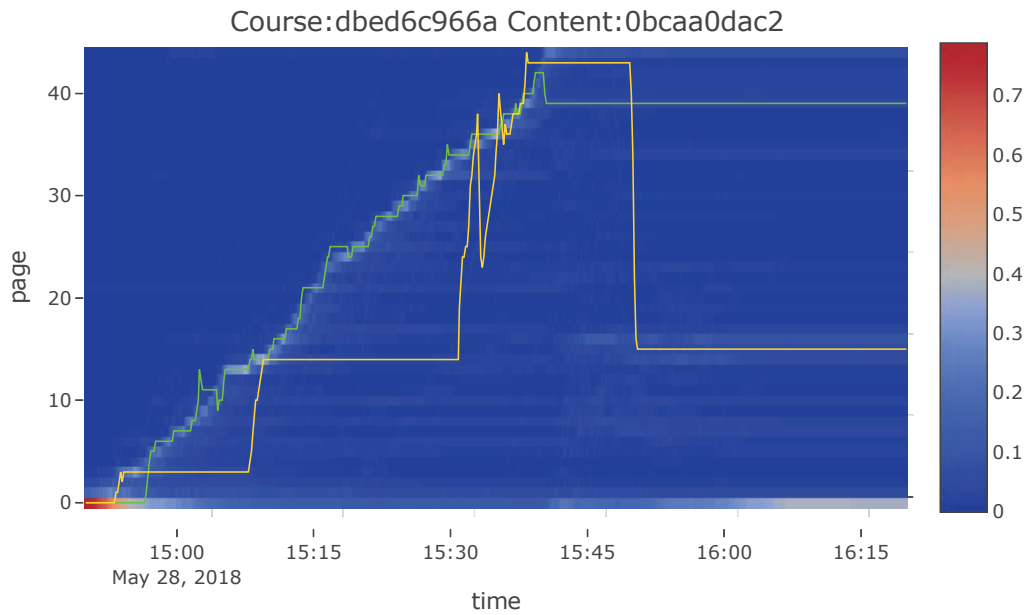


Figure 2: Normalized histogram of students reading a textbook.

of Figure 2 as more earnest students than students in the bottom-left of Figure 2 because they have larger action scores. In this paper, an earnest student means that the student's reading behavior is similar to majority of the other students' reading behavior. In addition, the student proactively do actions such as the creation and deletion of digital textbook highlights.

In our result, there were some students with higher quiz scores but smaller action scores. We consider that this result occurs due to these high performance students who read the textbook at their own pace. According to Figure 2, there is no relation between students' quiz scores and action scores. Therefore, we will not be able to predict the quiz scores using only our action scores. However,

combination of students' quiz scores and action scores is useful for understanding characteristics of the students as mentioned above. In addition, our visualization method based on this combination will help teachers to choose how to support students.

5 CONCLUSION

We analyzed learning behaviors of students in face-to-face courses of Kyushu University. In this paper, we visualize distributions of quiz scores and action scores we defined as reading behaviors. The result of our visualization method implied existences of students whose learning behaviors did not contribute to obtain higher quiz scores. We believe that our visualization can help teachers to support such the students. In this paper, we focused on visualization for the relationship between the quiz scores and the action scores. We will extract patterns of event logs related to higher quiz scores based on our visualization in future.

REFERENCES

- Davis, D., Chen, G., Hauff, C., & Houben, G. J. (2016, 6). Gauging MOOC Learners' Adherence to the Designed Learning Path. The 9th International Conference on Educational Data Mining, (pp. 54-61). Raleigh, North Carolina, USA.
- Flanagan, B., & Ogata, H. (2017). Integration of learning analytics research and production systems while protecting privacy. Proceedings of the 25th International Conference on Computers in Education (ICCE2017), (pp. 333-338). Christchurch, New Zealand.
- LAK19 Data Challenge. (2018, 11). Retrieved from LAK19 Data Challenge: <https://sites.google.com/view/lak19datachallenge>
- Ogata, H., Yin, C., Terai, M., Okubo, F., Shimada, A., Kentaro, K., & Yamada, M. (2015, 11). E-book-based learning analytics in University education. The 23rd International Conference on Computers in Education, ICCE 2015, (pp. 401-406). Hangzhou, China.
- Okubo, F., Shimada, A., Taniguchi, Y., & Konomi, S. (2017, 10). A Visualization System for Predicting Learning Activities Using State Transition Graphs. The 4th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2017,, (pp. 173-180). Vilamoura.
- Okubo, F., Takayoshi, Y., Shimada, A., & Hiroaki, O. (2017). A neural network approach for students' performance prediction. The Seventh International Learning Analytics & Knowledge Conference, (pp. 598-599). Vancouver, BC, Canada. doi:10.1145/3027385.3029479
- Shimada, A., & Konomi, S. (2017, 12). Cross analytics of student and course activities from e-book operation logs. The 25th International Conference on Computers in Education, ICCE 2017, (pp. 433-438). Christchurch, New Zealand.
- Shimada, A., Taniguchi, Y., Okubo, F., Konomi, S., & Ogata, H. (2018, 3). Online change detection for monitoring individual student behavior via clickstream data on E-book system. The 8th International Conference on Learning Analytics and Knowledge, (pp. 446-450). Australia.
- You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. In Internet and Higher Education (Vol. 29, pp. 23-30). doi:<https://doi.org/10.1016/j.iheduc.2015.11.003>

How Students Flip Pages during Lectures? -Comparison between Power Users and Normal Users-

Takuro Owatari, Atsushi Shimada, Tsubasa Minematsu, Rin-ichiro Taniguchi
Kyushu University, Japan
oowatari@limu.ait.kyushu-u.ac.jp

ABSTRACT: In this paper, we tackle the learning behavior analytics of students using e-Book event stream data collected by BookRoll system. There are many types of operations recorded in the logs. In the case of KU dataset, the event stream data was recorded in the face-to-face style lectures over 8 weeks. Our analytics especially focuses on the learning logs recorded during 90-min lecture time. In our research, the difference of three features related to student's e-book text browsing activities is compared between two groups. Our experiments suggested interesting results that the power users tended to flip the pages earlier than normal users.

Keywords: learning behavior, e-Book event stream, educational big data, browsing pattern

1 INTRODUCTION

Much attention has been paid to learning analytics, which is defined as the measurement, collection, analysis, and reporting of data about learners and their context, for the purpose of understanding and optimizing learning and environments in which it occurs (<https://solaresearch.org/>). Traditionally, many studies have focused on clickstream data collected from Massive Open Online Courses (MOOCs), and analyzed the data for prediction of course completion (Crossley 2016), change detection of students' behavior (Park 2017) and so on. Recent years, event stream data from e-Book systems have been also utilized to understand students' learning activities. For example, the data was analyzed for pattern mining of preview and review activities (Oi 2015), understanding learning behavior of students (Yin 2015), browsing pattern mining (Shimada 2016), and performance prediction (Okubo 2016).

In this paper, we tackle the learning behavior analytics of students using e-Book event stream data collected by BookRoll system (Ogata 2015, Flanagan 2017, Ogata 2017). There are many types of operations recorded in the logs. For example, OPEN means that the student opened the e-Book file, whereas NEXT means that the student clicked the next button to move to the subsequent page. Students can use learning tools such as BOOKMARK on pages and HIGHLIGHT on keywords and sentences. These operation logs are also collected in the dataset. In the case of KU dataset, the event stream data was recorded in the face-to-face style lectures over 8 weeks. In the field of learning analytics, learning behavior of students during the lectures has an important role. According to the previous study, the usage of e-Book function has relationship with self-regulated ability (Yamada 2017). Therefore, our analytics especially focuses on the learning logs of three major functions: BOOKMARK, HIGHLIGHT and MEMO, and would like to figure out representative patterns in terms of following aspects.

- How students flip pages during lectures?

- What is the difference in learning behaviors between power users (those who frequently utilize BOOKMARK and HIGHLIGHT operations) and normal users (those who just following the pages)?

In the following sections, we introduce our analytics strategy and primal results.

2 METHODS

2.1 Overview

Our analytics strategy consists of four steps as follows.

1. Browsing page of each student: For each student, the page which the student browsed is estimated every 30 seconds.
2. Browsing heat map: For each lecture and for each page, the number of students who browsed the page is estimated every 30 seconds.
3. Feature extraction: Three kinds of features are calculated from the browsing heat map.
4. Comparison: The extracted features are compared between two groups. The first group contains students who frequently utilize HIGHLIGHT operation, and the second group contains the other students.

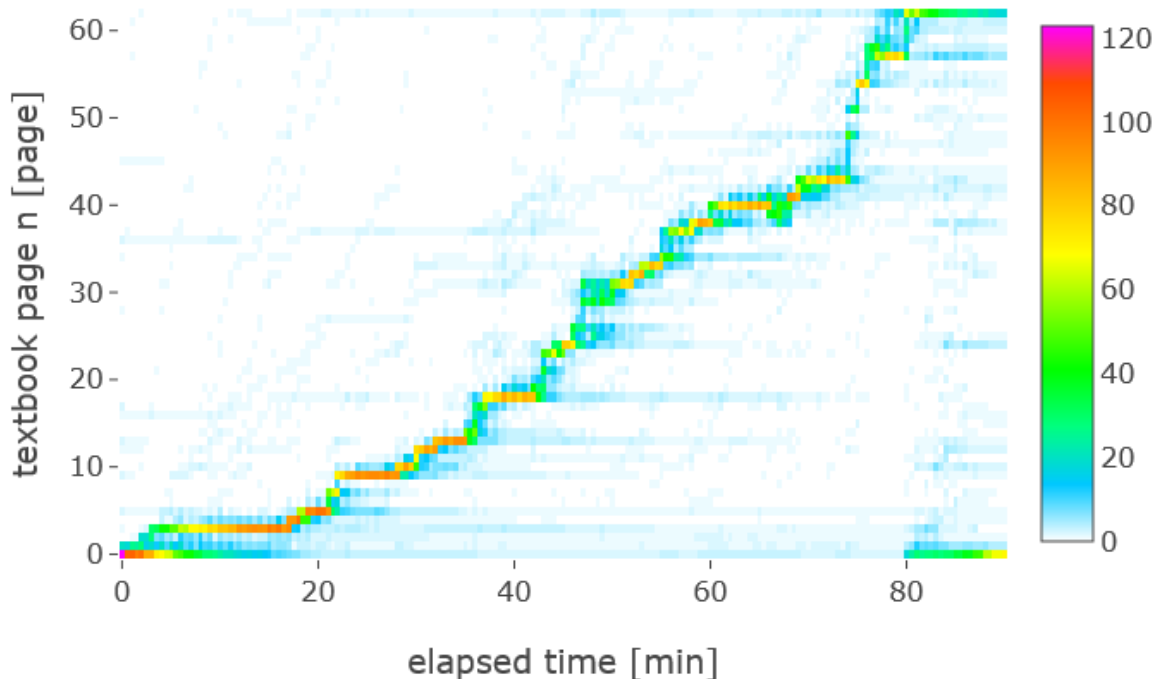


Figure 1: Browsing heat map

2.2 Details

Let t and n be a time slot of 30 second and the page in the lecture material, respectively (see Figure 2 for calculation examples). 30 second is based on assumption that a teacher takes at least about 30 seconds to explain one page of teaching materials. First, our method estimates the page n which is

browsed by each student at t by indicating the longest browsing time during the time slot t . Considering the cross page browsing, we picked up top 2 pages (at most) as the browsing pages of each student. Second, $V_{n,t}$, the number of students who browsed page n at time t (defined as browsers in the following), is indicated for each n and each t . All $V_{n,t}$ of one lecture are represented as Browsing heat map (see Figure1 as an example) on which each color is the value of $V_{n,t}$.

Next, to characterize learning activities in each page n , let P_n be the time at which the number of browsers was the largest on page n . Besides, we calculate two additional features I_n and D_n for each page n : the time when the number of browsers was the most increased/decreased, respectively (see Figure 3). More specifically, three features P_n , I_n and D_n for page n are calculated as follows.

$$P_n = \underset{t}{\operatorname{argmax}} V_{n,t}$$

$$I_n = \underset{t}{\operatorname{argmax}} (V_{n,t} - V_{n,t-1})$$

$$D_n = \underset{t}{\operatorname{argmin}} (V_{n,t} - V_{n,t-1})$$

Finally, learning activities are compared between two groups

G1: Students in 75 percentiles excluding outliers, those who frequently left highlight on the lecture materials.

G2: Students who did not leave any highlight on the lecture materials.

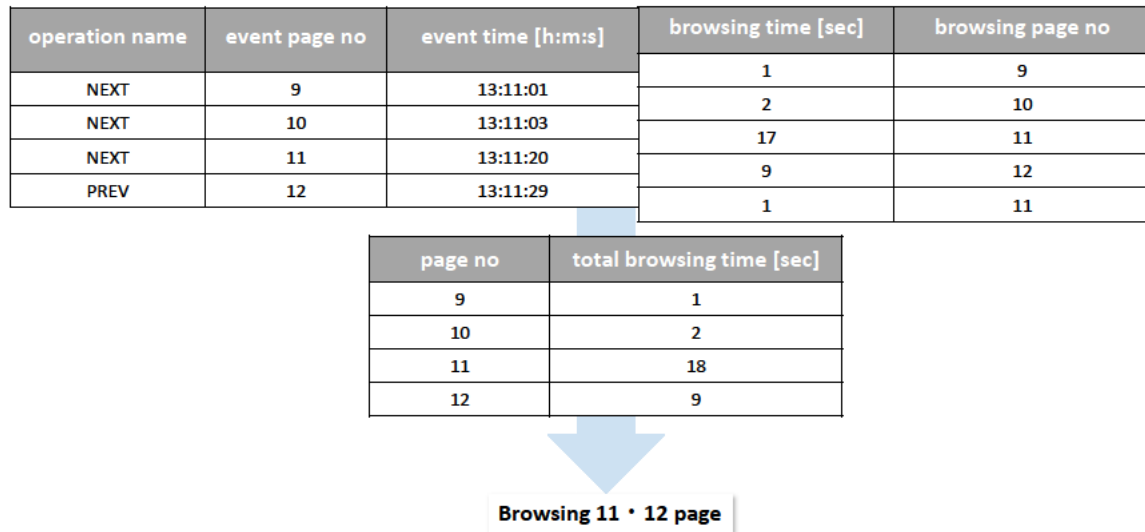


Figure 2 : Estimation of the browsing page

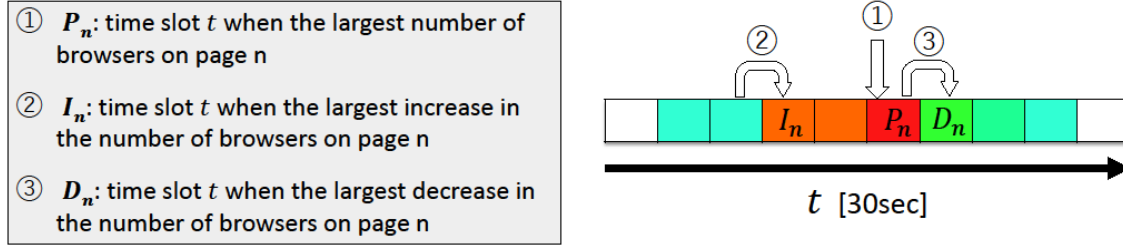


Figure 3: P_n , I_n and D_n on page n in browsing heat map.

Above three features are calculated for each group, and compared to figure out the characteristics of the learning activities.

3 EXPERIMENTS

We conducted experiments to compare the difference of aforementioned three features between two groups. Totally 1328 students (the number of unique users) attended the 10 course lectures over 8 weeks. In our experiments, we analyzed each lecture individually so that cumulative total number of students was 9519 due to independent analyses over 8 week lectures. And we extracted students who satisfied the condition of G1 and G2, resulting in 851 students and 5558 students in G1 and G2, respectively. In addition, the total number of pages used for the analytics was 3020. Remind that the P_n , I_n and D_n represent the timestamp when each feature was observed for page n . Therefore, we extracted these three timestamps of every page for G1 and G2, individually. Then, we compared these timestamps between G1 and G2. More specifically, we classified the comparison result into three categories:

Precede: In the case that the timestamp of P_n , I_n or D_n of G1 is earlier than G2.

Delay: In the case that the timestamp of P_n , I_n or D_n of G1 is later than G2

Equal: In the case that the timestamp of P_n , I_n or D_n is the same between G1 and G2.

Finally, we investigated the percentage of above three categories for P_n , I_n and D_n .

Table 1: Evaluation results

	P_n	I_n	D_n
Precede(%):Ave(Std)	0.507 (0.199)	0.283 (0.185)	0.346 (0.180)
Delay(%):Ave(Std)	0.171 (0.120)	0.157 (0.106)	0.159 (0.115)
Equal(%):Ave(Std)	0.322 (0.165)	0.560 (0.224)	0.495 (0.213)

Table 1 shows the average and standard deviation scores of each category and feature. For example, in almost half of the cases, majority of students in G1 browsed the same page earlier than those in G2 (see the cell at Precede row and P_n column whose value is 0.507). As a whole, we found out that students in G1 tended to flip pages earlier than those in G2 (see Table 1). Before conducting the investigation, we expected that the page flip timing of G1 would be later than G2 because it would take some amount of time to put highlight on keywords. However, the comparison result suggested

the opposite tendencies. In fact, the percentages of precede in G1 were larger than those in G2. It is expected that students in G1 (i.e., power users of e-Book) diligently listened to the teacher's explanations, and left highlight as quickly as possible in order to keep up with the lecture speed.

4 CONCLUSION AND FUTURE WORK

In this paper, we tackled the learning behavior analytics using e-Book event stream data. Through the analytics of e-Book event stream data, we extracted three kinds of features to measure learning behaviors of power users of e-Book system. Our experiments suggested interesting results that the power users tended to flip the pages earlier than normal users. In our future work, we will conduct statistical analyses like as significance test between two groups. On the other hand, the comparison results differed from lecture courses. We are going to continue course-by-course analytics in future. Besides, we will investigate the lecture-wise characteristics and the learning behaviors of power users in details, for example, by analyzing out-lecture activities, tendency throughout the lectures, analytics across universities, and so on.

REFERENCES

- S. Crossley, L. Paquette, M. Dascalu, D. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. LAK'16, 6-14.
- J. Park, K. Denaro, F. Rodriguez, P. Smyth, M. Warschauer. (2017). Detecting Changes in Student Behavior from Clickstream Data. LAK'17.
- M. Oi, F. Okubo, A. Shimada, C. Yin, and H. Ogata. (2015). Analysis of preview and review patterns in undergraduates' e book logs. the 23rd International Conference on Computers in Education, 166–171.
- C. Yin, F. Okubo, A. Shimada, S. Hirokawa, H. Ogata, and M. Oi. (2015). Identifying and analyzing the learning behaviors of students using e - books. the 23rd International Conference on Computers in Education, 118–120.
- A. Shimada, F. Okubo, H. Ogata (2016). Browsing-Pattern Mining from e-Book Logs with Non-negative Matrix Factorization. the 9th International Conference on Educational Data Mining, 636–637.
- F. Okubo, T. Yamashita, A. Shimada, H. Ogata (2017). A Neural Network Approach for Students' Performance Prediction. LAK'17
- H. Ogata, C. Yin, M. Oi, F. Okubo, A. Shimada, K. Kojima, M. Yamada (2015). E-Book-based learning analytics in university education. Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015) pp.401-406.
- B. Flanagan, H. Ogata (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. Proceedings of the 25th International Conference on Computers in Education (ICCE2017), pp.333-338.
- H. Ogata, M. Oi, K. Mohri, F. Okubo, A. Shimada, M. Yamada, J. Wang, S. Hirokawa (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In Smart Sensors at the IoT Frontier, pp.327-350, Springer.
- Yamada, M., Shimada, A., Okubo, F., Oi, M., Kojima, K., & Ogata, H. (2017). Learning analytics of the relationships among self-regulated learning, learning behaviors, and learning performance, Research and Practice in Technology Enhanced Learning, 12, 13. doi: 10.1186/s41039-017-0053-9

Extracting E-book Reading Patterns using Stochastic Block Model

Kanishka Khandelwal

NEC Corporation

k-khandelwal@ay.jp.nec.com

Hiroshi Tamano

NEC Corporation

h-tamano@bx.jp.nec.com

ABSTRACT: To understand the log data of digital textbooks and get insights from it, extracting browsing patterns is basic and meaningful. The state of the art browsing pattern extraction method based on Non-negative Matrix Factorization (NMF) is comprehensive and useful for learning analytics researchers. However, in our survey, we find it's not easy enough for teachers who are non-experts of data analysis to understand and utilize its results. In this paper, we propose a method for typical teachers to easily understand the extracted browsing patterns and utilize them in their classes. We use Stochastic Block Model (SBM) to identify co-clusters of students and reference materials and showcase these relationships in easy-to-interpret format using bipartite graph. In the experiment, we show that SBM extracts useful post class reading patterns and the bipartite graph is easy enough to be understood by non-experts of data analysis. Moreover, we present the results of comparative survey conducted among K-12 teachers to substantiate our method's precedence over state of the art NMF technique in terms of ease of practical use and interpretability. Lastly, we discuss how our results can be used to improve the course/lecture and design the quiz for next class.

Keywords: Stochastic Block Model, Bipartite Graph, E-book Log, Learning Analytics, Visualization

1 INTRODUCTION

Browsing logs of digital textbooks is one of the major log in learning analytics. Analyzing this log data is expected to make learning evidence-based and much more efficient. For example, e-book log was collected in (Ogata, H. et al., 2015) and three major directions of analysis have been taken so far. First is to find the relationship between student's score and his browsing log. It was found that the academic performance is related to preview but not to review learning behavior (Oi, M. et al., 2015; Shimada, A. et al., 2015). Second direction is to extract the reading patterns. Four groups of reading patterns characterized by 'reading forward' and 'reading with backtrack' type behavior were analyzed (Yin, C. et al., 2015) and activity level (i.e., HIGH, LOW, MEDIUM, NONE) transition patterns were extracted using Markov Chains (Akçapinar, G. et al., 2018). Third is to visualize the student reading behaviors. Shimada, A. et al., (2017) proposed a real-time visualization technique that enables teachers to adjust the lecture speed for students in classes.

To understand the log data of digital textbooks, extracting reading patterns such as materials read, order of their access and time spent on each material, for each student is basic and meaningful. It gives insights for teachers to improve their classes and course materials while for learning-analytics

researchers to come up with further analysis. Shimada, A. et al., (2016) proposed Non-negative Matrix Factorization (NMF) based method to identify groups of students with similar slide accesses (referred hereafter as browsing patterns). NMF is applied to the browsing matrix V where i -th row represents i -th page of e-book and j -th column represents j -th student and V_{ij} is 1 if j -th student views i -th page longer than threshold value, otherwise 0. NMF decomposes matrix V into two matrix: W and H . W represents page and latent pattern matrix and H represents latent pattern and student matrix. Further, consensus clustering is performed on matrix H to group students with similar latent patterns. Learning analytics researchers are able to understand the page reading patterns for students from these two matrices. However, when we consider the case where non-experts of data analysis, i.e. typical K-12 school teachers, see these matrices, it seems the results are not easy enough to be interpreted and utilized in their classes. We believe, an easier and intuitive presentation of students' browsing patterns is needed to increase the utility in classrooms.

To this end, we propose a method that extracts such browsing patterns from log data of digital textbooks and showcases the results in an easy-to-interpret format for non-experts of data analysis. Our method uses Stochastic Block Model (SBM) (Wang, Y. et al., 1987; Nowicki, K. et al., 2001) to find reference material and student co-clusters from the browsing matrix. Further, to increase the interpretability, obtained relationships are depicted using a bipartite graph. In our experiment, we apply this method to extract interpretable browsing patterns from post-lecture reading log of university students. In Section 4, we discuss our method's applicability in improving courses and designing quizzes. Lastly, we present the results of survey conducted among K-12 teachers to evaluate the interpretability and utility of our method in comparison to state of the art NMF based method and to validate our proposed real-life applications.

2 METHOD

In this section, we provide the methodology to analyze the browsing behavior of students with respect to time spent on the reference materials while accessing the e-book system. In the first subsection, we describe the pre-processing of log data in order to apply SBM technique to obtain the co-clusters. Next, we briefly describe the Stochastic Block Model and realize its application in this domain. Finally, we aim to provide the results in a manner that is easy to interpret and even people with no expertise in analytics such as K-12 teachers can understand and use the results to improve their courses/lectures.

2.1 Data and Pre-processing

Our target data is logs obtained from digital textbooks that records or can be used to find the time students spent on reference materials (e.g. slides, lectures, book pages, etc.). We convert this log data into a browsing matrix B of binary values where the rows represent different students and columns represent reference materials. Each element in the matrix B thus means whether the student spent enough time ($>$ threshold time) reading the reference material or not.

2.2 Co-clustering using Stochastic Block Model

Co-clustering or biclustering is a popular data mining technique used to concurrently identify clusters in rows (of samples) and clusters in columns (of features) from a data matrix (Tanay, A. et

al., 2005; Charrad, M. et al., 2011). Data matrix can then be rearranged in the form of a block structured matrix where each block (also called co-cluster) contains the samples and features that share a relationship. For example, (Kemp, C. et al., 2006) applied co-clustering using Infinite Relational Model (IRM) on binarized animal-feature matrix to identify groups of animals that have common habitat, anatomical or behavioral features.

Indeed, many biclustering algorithms have been proposed to identify different kinds of co-clusters from the data (Madeira, S. et al., 2004). For the ease of tuning and interpretability of modelling task, in this work, we employ a simple probabilistic latent variable model, Stochastic Block Model (SBM) (Wang, Y. et al., 1987; Nowicki, K. et al., 2001), for the co-clustering task on matrix B. We aim to find clusters of students that share similar reading pattern and identify at the same time what they read.

SBM assumes that K student clusters and L slide clusters exist. Rows and columns in the browsing matrix are assigned clusters by categorical distributions with parameters (that sum to 1) sampled from Dirichlet distribution. The probability of student from cluster k reading a slide in cluster l is given by $\theta_{k,l}$ that comes from a beta prior. We perform variational inference to identify the cluster assignments for each row and column of the browsing matrix. The generative model is given below:

$$\begin{aligned}\pi_1 | \alpha_1 &\sim \text{Dirichlet}(\alpha_1) \\ \pi_2 | \alpha_2 &\sim \text{Dirichlet}(\alpha_2) \\ z_{1,i} = k | \pi_1 &\sim \text{Categorical}(\pi_1) \\ z_{2,j} = l | \pi_2 &\sim \text{Categorical}(\pi_2) \\ \theta_{k,l} | a, b &\sim \text{Beta}(a, b) \\ x_{i,j} | \{\theta_{k,l}\}, z_{1,i}, z_{2,j} &\sim \text{Bernoulli}(\theta_{z_{1,i}, z_{2,j}})\end{aligned}$$

2.3 Visualization

Analyzing student log data can help us identify patterns in the student reading behavior and the insights obtained can be used to profit the stakeholders. However, complexity of these analyses and the results obtained may hinder its application in practical settings. It is thus important to present the results in an easy-to-interpret format. We, thus, propose to visualize the results of such student reading behavior analyses in a bipartite graph $G = (U, V, E)$. In this scenario, vertices of the graph in the first set U can represent groups of students while that in the second set represents clusters of reference material. An edge in E tells us that the student cluster has spent appropriate time on the content cluster.

3 EXPERIMENTS

In this section, we apply the SBM technique to analyze the post-lecture study pattern of students. ‘Bookroll’ (Ogata, H. et al., 2017) is a digital teaching material delivery system which was used for several courses in different universities. The anonymized data (Flanagan, B. et al., 2017) consists of event logs of students’ interaction with this system used for accessing the course related materials (see Table 1). Moreover, additional information such as the timings of the lectures, total number of slides in each lecture material and final quiz score of the enrolled students are provided for every courses.

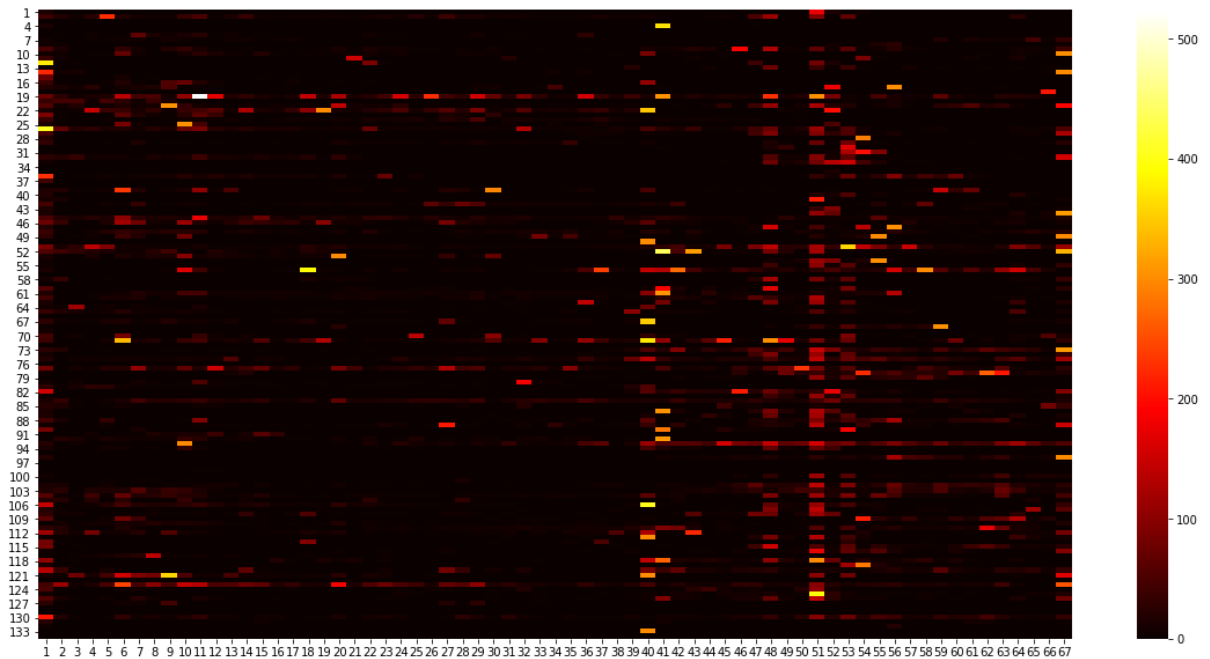
Table 1: Sample Event Log data

userid	contentsid	operationname	pageno	marker	memo_length	devicecode	eventtime
xxxxa7	xxxxx0b	OPEN	1		0	pc	4/9/2018 12:58:04 PM
xxxxa7	xxxxx0b	NEXT	1		0	pc	4/9/2018 1:00:08 PM
xxxxa7	xxxxx0b	ADD_MARKER	2	difficult	10	pc	4/9/2018 1:00:09 PM
xxxxa7	xxxxx0b	PAGE_JUMP	2		0	pc	4/9/2018 1:00:49 PM

3.1 Preparing the data

We present our analysis on the study pattern of students enrolled for the course ‘24a65f29b6’ from KU dataset. The course was run from 09/04/2018 to 04/06/2018 and 134 students enrolled in it. 8 lectures were given each with a separate reference material and a quiz was conducted at the end of the course with an average score of 78.5. To prepare the data for our analysis, we extract the post lecture event log data for the material ‘e18eedce0b’ which was used in the first lecture. Throughout the course duration, students accessed this material for reviewing the course contents. Fig. 1 shows the heatmap of the total time (in seconds) spent by 134 students on each slide (67 in total) in this material across several sessions. To increase the visibility of the heatmap, we threshold the time spent on any slide in one session to a maximum of 300 seconds. While it is evident that some slides were accessed by most students, it is difficult to make any further inferences from the figure.

Next, we convert the heatmap to a binary matrix B with 134 rows for students and 67 columns for slides wherein each entry represents whether the student spent enough time on the slide or not. Specifically, we set the entry b_{ij} to 1 if the time spent by student i on slide j is more than that of 80 percent of the class and 0 otherwise. We apply the SBM technique on this binary matrix to identify the co-clusters of students and content slides.

**Figure 1: Heatmap of time spent in reviewing slides of lecture 1 by students**

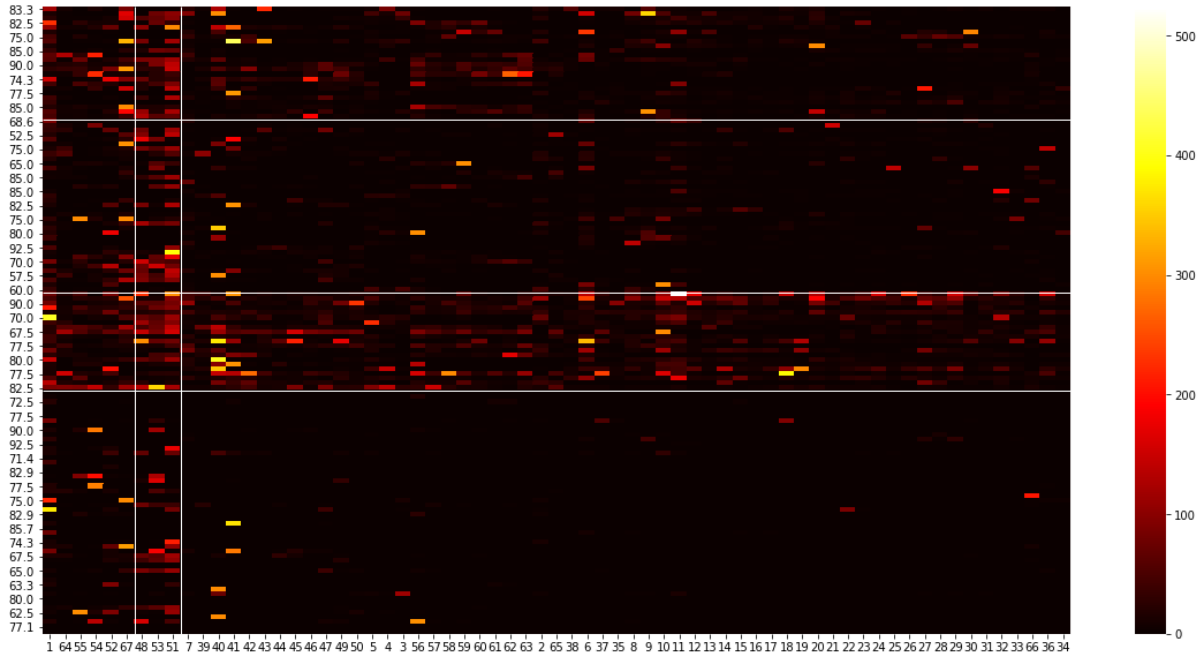


Figure 2: Heatmap of time spent on lectures with the block structure obtained from SBM

3.2 Experimental settings and Results

SBM model has six hyper-parameters; K and L are the maximum number of student and slide clusters, α_1 and α_2 are the Dirichlet parameters, while a and b are the Beta parameters. We perform grid search over the hyper-parameter set $\{K, L, a, b, \alpha_1, \alpha_2\}$ to find optimal values of $\{4, 3, 0.1, 0.1, (1,1,1,1), (8,8,8,8)\}$ that generate consistent block structure in the heat maps.

Fig. 2 shows the co-clustering of students and slides along with the scores of students along the vertical axis. A clear block structure is evident from the fig. 2 as compared to fig. 1. Students are clustered into 4 groups with each group exhibiting a distinct post class study behavior across the 3 detected groups of slides.

3.3 Interpretation and Visualization

Fig. 3 pictorially depicts the information about the study behavior of students obtained from the block structure of fig. 2 in a format that can be easily understood and interpreted by K-12 teachers. The set of nodes in the left part of bipartite graph are the student clusters with sizes 24, 37, 21 and 52. Similarly, the slide clusters of sizes 6, 3 and 58 are the nodes in the right part. The plots next to slide clusters has x-axis as the slide numbers and shows the slides present in respective clusters with a bar. Lastly, an edge from a student cluster to a slide cluster represents that the respective cohort of students have read the group of slides by spending on an average more than 20 seconds per slide.

As can be seen, out of the students that used Bookroll to review this lecture (S_1, S_2 , and S_3), a larger proportion chose to review the slides in C_1 and C_2 , suggesting that the last portion of the lecture was either the most important part or it was difficult to grasp. Additionally, we can identify that student cluster S_2 reviews only few slides while cluster S_3 are reviewing the whole lecture. Importantly, we can identify the cohort of students in S_4 that doesn't revise the lecture at all, properly.

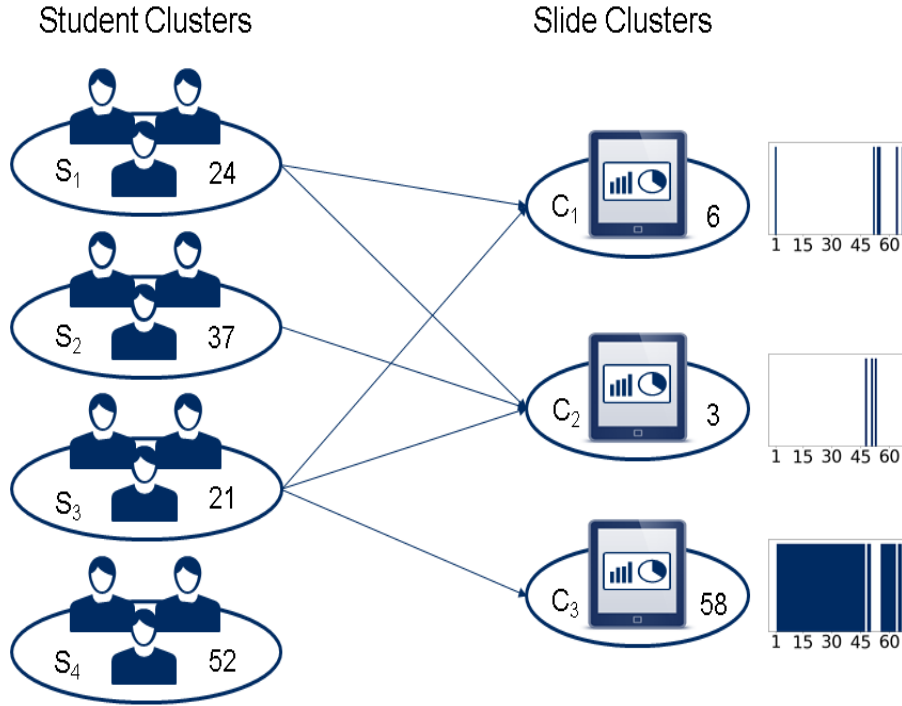


Figure 3: Visualization of relationship between student and slide clusters

4 APPLICATIONS

We emphasize the importance of analyzing the review study pattern of students using the proposed method with two concrete applications.

4.1.1 Improving Course/Material

Through co-clustering, we can identify the slides which students are focusing less upon while reviewing the lectures. In our example, huge cohort of the class is ignoring the initial lectures and reviewing just the last part. In this way, course instructors can modify the lecture material or change the slide order in the middle of the course to increase the visibility of ignored slides for the following review sessions that have important concepts. By analyzing event logs of previous batches, instructors can also improve their teaching pattern by retrospection.

4.1.2 Designing Quiz

To improve the engagement of student in the course across all the lectures, many instructors usually start the class with a small quiz that tests the student's understanding of previous lectures. The proposed method, can be used in designing personalized quizzes for students. In this example, for the quiz at the beginning of lecture 2, questions from the initial slides of lecture 1 should be focused upon more for students in S_1 and S_2 . On the other hand, for S_3 and S_4 , questions covering the whole lecture should be presented.

5 SURVEY FOR VALIDATION

Even though both, our method and NMF based approach, can provide similar insights about the browsing patterns of students in a class, we claim the results of our method are more interpretable and usable by non-experts of data analysis such as K-12 teachers. In order to substantiate our claims

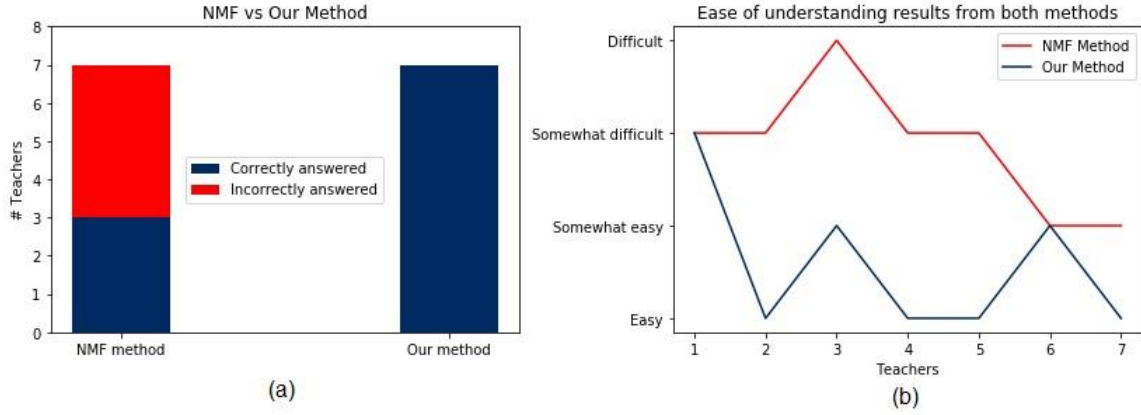


Figure 4: Results from online survey (a)Teacher performance in the quiz to judge their understanding of results (b)Level of difficulty assigned by teachers to both the methods

we conducted a survey among 7 K-12 teachers from India and Japan to get their feedback on both the methods.

5.1.1 Survey Form

Our online survey form consisted of 4 sections and 12 questions in total. First section had 3 questions about the respondent's name, affiliation and grades he/she teaches. Second and third section respectively for NMF-based approach and our method consisted 4 similar questions for comparative study. First question explained the graphical representations, heat map or bipartite graph, that are used in the presentation of results obtained from the two methods and later asked them if the provided explanation was sufficient. With this acquired basic understanding, in the next question we presented the result of the approaches (applied on same dataset) and asked whether they understood the graphs. Third question, a single choice question, was to judge the correctness of their understanding of these results by asking them the slides read by students in a particular cluster. Finally, through fourth question we asked their feedback on the difficulty level of the individual approaches in terms of interpretation and utility in classrooms. Last section which had only 1 question was to validate our method's applications that we present in section 4 of this manuscript.

5.1.2 Survey Results

Feedback obtained from this survey largely supports our claims about both the approaches. Figure 4(a) is a stacked bar plot depicting the number of teachers who gave correct or incorrect answers to third question in sections 2 and 3 of our survey. It clearly validates our assumption of the difficulty in understanding of the results obtained from NMF-based approach as more than half the teachers couldn't interpret the two heat maps correctly. Moreover, for our approach, due to the simplicity of bipartite graph based representation all of the teachers could identify the correct answer. In figure 4(b), we showcase the feedbacks obtained from the last question of section 2 and 3 about the ease of use of both the approaches in practical settings. While 6 out of 7 teachers consider our method easy to use, only 2 teachers thought the same for NMF-based approach. Moreover, no teacher thought our method is difficult to use compared to the counterpart. Lastly, through last section of our survey, more than half of the teachers validated the proposed real-life applications.

6 CONCLUSION

In this work, we have proposed a method to extract browsing patterns of students from reading log of digital textbooks and generate easy-to-interpret results. We have provided an empirical investigation by analyzing the review behavior of university students from the log data of 'Bookroll', a digital content delivery system. Further, we have discussed the practical applications of the proposed method in designing quizzes and improving courses/materials. Lastly, through a survey among K-12 teachers we found our method's precedence over state of the art NMF technique in terms of ease of practical use and interpretability.

ACKNOWLEDGEMENT

We wholeheartedly thank Dr. Koji Ichikawa to set the direction of this research and provide the code for performing inference. We also thank the anonymous reviewers as well as Kazuya Hirata for their helpful comments and support.

REFERENCES

- Akçapınar, G., Majumdar, R., Flanagan, B., and Ogata, H. (2018). Investigating Students' e-Book Reading Patterns with Markov Chains. In the 26rd International Conference on Computers in Education (ICCE 2018)
- Charrad, M., and Ahmed, M. B. 2011. Simultaneous clustering: A survey. In International Conference on Pattern Recognition and Machine Intelligence, 370–375. Springer.
- Flanagan, B., Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In *International Conference on Computers in Education (ICCE2017)* (pp.333-338).
- Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006, July). Learning systems of concepts with an infinite relational model. In AAAI (Vol. 3, p. 5).
- Madeira, S. C. and A. L. Oliveira (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(1), 24–45.
- Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96:455, 1077–1087
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., and Yamada, M. (2015, January). E-Book-based learning analytics in university education. In International Conference on Computer in Education (ICCE 2015) (pp. 401-406).
- Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., and Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In *Smart Sensors at the IoT Frontier* (pp. 327-350). Springer, Cham.
- Oi, M., Okubo, F., Shimada, A., Yin, C., and Ogata, H. (2015). Analysis of preview and review patterns in undergraduates' e-book logs. In 23rd International Conference on Computers in Education, ICCE 2015. Asia-Pacific Society for Computers in Education.
- Shimada, A., Mouri, K., and Ogata, H. (2017, July). Real-Time Learning Analytics of e-Book Operation Logs for On-site Lecture Support. In 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT) (pp. 274-275). IEEE.

- Shimada, A., Okubo, F., and Ogata, H. (2016). Browsing-Pattern Mining from e-Book Logs with Non-negative Matrix Factorization. In EDM (pp. 636-637).
- Shimada, A., Okubo, F., Yin, C. J., Oi, M., Kojima, K., Yamada, M., and Ogata, H. (2015, January). Analysis of preview behavior in E-book system. In the 23rd International Conference on Computers in Education (ICCE 2015) (pp. 593-600).
- Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., Yamada, M., and Ogata, H. (2015, January). Analyzing the features of learning behaviors of students using e-books. In Workshop proceedings of International Conference on Computers in Education (pp. 617-626).
- Tanay, A.; Sharan, R.; and Shamir, R. (2005). Biclustering algorithms: A survey. *Handbook of computational molecular biology* 9(1-20):122–124.
- Y. J. Wang and G. Y. Wong, "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 8–19, 1987.

Knowledge Map Creation for Modeling Learning Behaviors in Digital Learning Environments

Brendan Flanagan^{1*}, Rwitajit Majumdar¹, Gökhan Akçapınar^{1,2},
Jingyun Wang³ and Hiroaki Ogata¹

¹ Kyoto University, Japan

² Hacettepe University, Turkey

³ Kyushu University, Japan

* flanagan.brendanjohn.4n@kyoto-u.ac.jp

ABSTRACT: There has been much research that demonstrates the effectiveness of using ontology to support the construction of knowledge during the learning process. However, the widespread adoption in classrooms of such methods are impeded by the amount of time and effort that is required to create and maintain an ontology by a domain expert. In this paper, we propose a system that supports the creation, management and use of knowledge maps at a learning analytics infrastructure level, integrating with existing systems to provide modeling of learning behaviors based on knowledge structures. Preliminary evaluation of the proposed text mining method to automatically create knowledge maps from digital learning materials is also reported. The process helps retain links between the nodes of the knowledge map and the original learning materials, which is fundamental to the proposed system. Links from concept nodes to other digital learning systems, such as LMS and testing systems also enable users to monitor and access lecture and test items that are relevant to concepts shown in the knowledge map portal.

Keywords: Knowledge map; concept-based analytics; concept maps; knowledge extraction;

1 INTRODUCTION

It has been well documented that learners can benefit from the use of maps to represent the key concepts of knowledge (Lee et al., 2012). Ausubel (1963; 1968) defined the effective assimilation of new knowledge into an existing knowledge framework as the achievement of “meaningful learning”, by which knowledge maps can serve as a kind of scaffold to help learners to organize knowledge and structure their own knowledge framework (Novak et al., 2006). However, the process of creating and maintaining these maps often involves a domain expert manually creating the knowledge map based on their experience and previous knowledge (Wang et al., 2017).

To support the creation and use of knowledge maps by teachers and learners, we propose a knowledge map system that integrates with existing digital learning environments and learning analytics infrastructure. To assist in the creation of knowledge maps from digital learning materials, we propose a process for extracting key concepts from unstructured text to generate knowledge structures. Maps that have been generated are stored in a Knowledge Map Store (KMS) and an authoring system is provided for teachers to create, edit, and manage stored knowledge maps before publishing. The Knowledge Portal provides visualizations of knowledge maps with attributes determined from the analysis of learning behavior event log data from existing learning analytics

infrastructure. In the final section of this paper, we outline the anticipated cases in which the system will be utilized by both teachers and students to monitor individual and group knowledge states.

There are many previous researches into the generation and use of ontologies, concept maps, and knowledge maps in education to show and create knowledge frameworks. Association rules and other data mining techniques have been used to construct concept maps based on the results of test and quizzes to show the relation between knowledge that was tested (Hwang, 2003; Tseng et al., 2007; Chen et al., 2010; Chen et al., 2013). While this technique is applicable to the structured format of tests, it is difficult to apply similar techniques to unstructured text that is contained in digital learning materials.

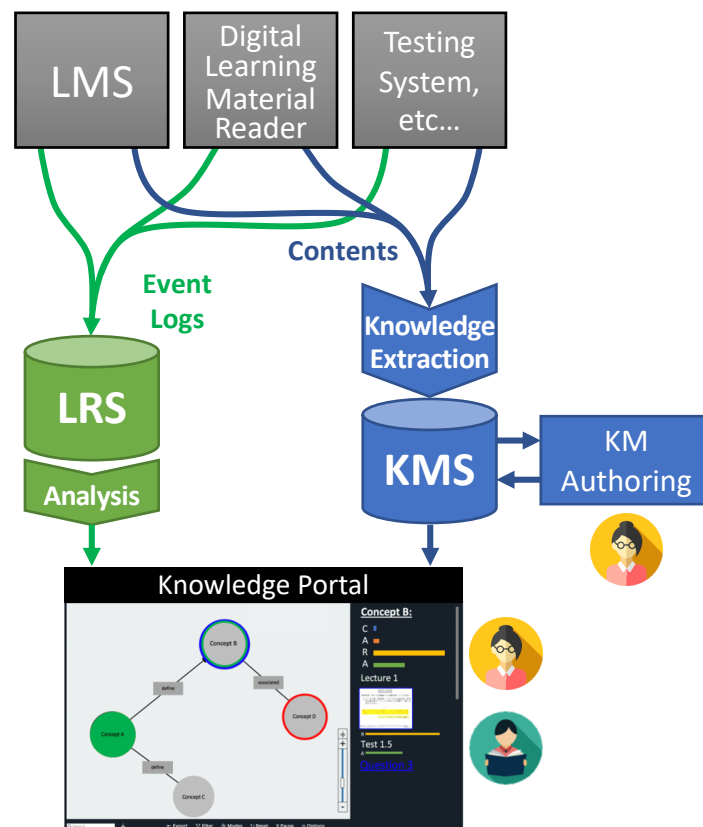


Figure 1: An overview of how the proposed system would integrate with existing LA infrastructure.

2 SYSTEM OVERVIEW

In this section, we provide an outline of the proposed Knowledge Map system, how it integrates with existing LA infrastructure, and how stakeholders will interact with the system. Fig. 1 shows an overview of the system with the main components consisting of:

- Existing user facing LA infrastructure, such as: LMS, Digital Learning Material Reader, Testing system, etc.
- LRS and Analytics Processor.
- Knowledge Extraction Processor.

- KMS (Knowledge Map Store) and a teacher facing Knowledge Map Authoring portal.
- User facing Knowledge Portal.

The existing user facing infrastructure, such as: LMS, Reader, and Testing system serve as an interaction event sensor and also as a source of learning material contents that are sent to the Knowledge Extraction Processor. Recent implementations of LA platforms often utilize an LRS and Analytics Processor as a pipeline for storing and processing event statement data about the use of user facing learning systems (Chatti et al., 2017; Flanagan and Ogata, 2017). We use this existing pipeline to provide information to augment the visualization of knowledge structures representing the underlying learning materials, lecture attendance, and past academic achievement.



Figure 2: Hierarchy of node attributes based on event log analysis.

The main hierarchy of node attributes based on analytics is shown in Fig. 2, where each level is linked to important stages in the formal learning process: lecture attendance, reading learning materials, confirming acquired knowledge through the answering of tests, and attaining a credit for having satisfied the requirements of a course. The most basic form of effort by a learner is to attend a lecture in which learning material related with the concept node was covered. When a learner actively reads the learning material the concept node is attributed as Read. If a learner has correctly answered a test item relating to the concept, then the node is given the Answered attribute. Finally, the if the student passes the course then the Credit attribute is assigned.

The Knowledge Extraction Processor analyzes learning content data from the LMS, Reader, and Testing system. In the present paper, we focus on the extraction of knowledge maps from PDF contents that have been uploaded to the digital learning material reader. The results of this process are then stored in the KMS. Teachers are able to manage knowledge maps stored in the KMS through a teacher facing authoring portal.

3 KNOWLEDGE MAP EXTRACTION FROM CONTENTS

Course curriculum in K-12 education is often well structured and defined by government level organizations that regulate education. However, higher education often is less regulated with the course curriculum being decided by the teacher. In Japanese universities, teachers in charge of courses are busy and course contents are often finished close to when a lecture is due to start, allowing little time to create knowledge maps manually.

The authoring section of the proposed system automatically analyzes contents uploaded by teachers to support the generation of knowledge maps. As a part of the authoring process, the system requires the map to be checked by the teacher before being used by students. The teacher is also able to edit the automatically generated knowledge map to add, remove, or alter required sections.

A knowledge map can be thought of as a graph of key points that are contained within the digital learning material contents that it represents. The relation between nodes of this graph are expressed as a weighted edge representing the strength of the relation between two key points that are in the contents. In this paper, we use a process based on a method previously proposed by Flanagan et al. (2013) as shown in Fig. 3.

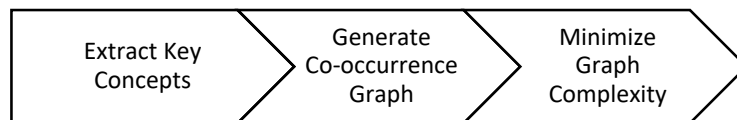


Figure 3: The process used to extract a knowledge map from digital learning material contents.

The lecture slides are usually written in Japanese with sections also in English. The text is extracted from lecture slides PDF files using pdfminer¹ and parsed with MeCab (Kudo, 2006) to separate individual words and parts-of-speech (POS) from a sentence using morphological analysis. Key concept terms are extracted by selecting the longest sequences of nouns and conjugate particles in a sentence. These were then indexed using the GETAssoc² search engine to form a co-occurrence matrix of terms. The link between the concept terms and the sections of the learning material are also included as an attribute in the search engine so relevant learning resources can be retrieved. The final step of the process involves minimizing the complexity of the co-occurrence graph using a minimum spanning tree algorithm to select the strongest concept term relations. In this implementation a thesaurus of technical terms in Japanese and English was used to guide the knowledge map generation process with hierarchical selection.

Table 1: Learning materials for the evaluation.

Lecture	Pages	Concepts (Gold Standard)	Max Concepts (Proposed)
1	30	12	125
2	32	10	153
3	45	6	222

We conducted a preliminary experiment using the proposed method in a university course on Information Science. A knowledge map that includes the concepts of three lecture learning materials was created manually by the course teacher and used as the gold standard for evaluation as shown in Table 1. Knowledge maps were automatically generated for each lecture with the strongest relation calculated using the SMART weight as described in Salton (1983). The precision/recall evaluation when comparing generated maps to the gold standard is shown in Fig. 4 with maximum precision of 0.72 at

¹ <https://euske.github.io/pdfminer/>

² <http://getassoc.cs.nii.ac.jp>

a threshold of 11 nodes for each generated map. As the threshold is increased the precision decreases, however the evaluation shows a majority of correct nodes are extracted at low thresholds. The generated knowledge maps would require some manual editing by a teacher before use in order to represent the same structure as the gold standard, and therefore is an ongoing topic of research.

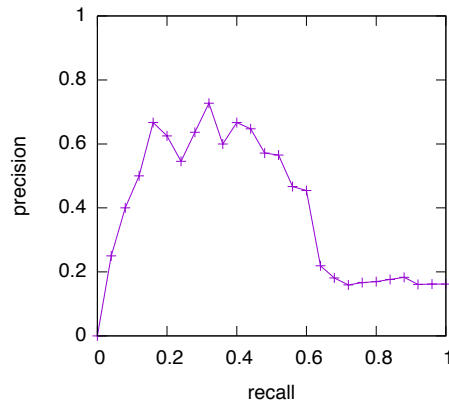


Figure 4: Plot of precision recall of proposed method

4 KNOWLEDGE MAP STORE AND AUTHORING

A centralized storage system of curated knowledge maps is fundamental to the analysis of knowledge accumulated over long-time spans. At the center of the proposed system, a KMS (Knowledge Map Store) acts as an LRS would for a conventional LA platform, collecting data about learning materials from disparate tools and systems to reduce information silos. This could enable the cross referencing and merging of knowledge maps from separate courses, learning materials, and even educational institutions if a KMS is deployed at the inter-institutional level.

The key data that a KMS should store are:

- The structure of knowledge maps that have been generated automatically by the system or created manually by teachers using the authoring interface.
- Links from the concept nodes of a knowledge map to related lecture schedule, learning materials, test items, and learner academic achievement records.

We are proposing that the structure of the knowledge map and links to learning materials/test items should be stored using a standards-based RDF storage service.

The proposed system has an authoring portal to facilitate the creation and management of knowledge maps by teachers. Automatically generated maps are initially stored as a draft and are not publicly available until the course teacher has confirmed the structure and its link to learning materials/test items. Maps can be edited to remove irrelevant nodes and add nodes that are required to cover the concepts in the course. A search function similar to the proposed knowledge map extraction process can be used to support the linking of relevant sections of learning materials and tests items to manually added nodes.

Knowledge maps can also be related with global concepts in the KMS to support large scale knowledge mapping across multiple courses. This feature is intended to facilitate the analysis of prior learner

knowledge, thus allowing a teacher or learner to view what concepts learners have and have not acquired. There is also potential to apply the results of the analysis to recommend learning materials that should be studied to fill in knowledge gaps before attending a course.

5 KNOWLEDGE PORTAL

Once a knowledge map has been published with the authoring tool, it is available for use by students and teachers in the Knowledge Portal. The visualization interface for the proposed Knowledge Portal is based on a web-based open source ontology visualization system called WebVOWL (Lohmann et al., 2014). The interface of the proposed system is shown in Fig. 4 with the main knowledge map visualization on the left, and the right frame displays detailed information about the attributes of the selected node with relation to relevant learning materials. At the top of the right frame the user is given an overview of the percentile rank for each of the attributes: attend, read, and answer. It will also show if a credit has previously been attained in relation to the node concept. The user is able to follow the links to study learning materials or confirm their knowledge by a test item on the node concept. The visualization also features a filter to select specific nodes/relations and reduce the complexity of the knowledge map using varying degrees of edge collapse.

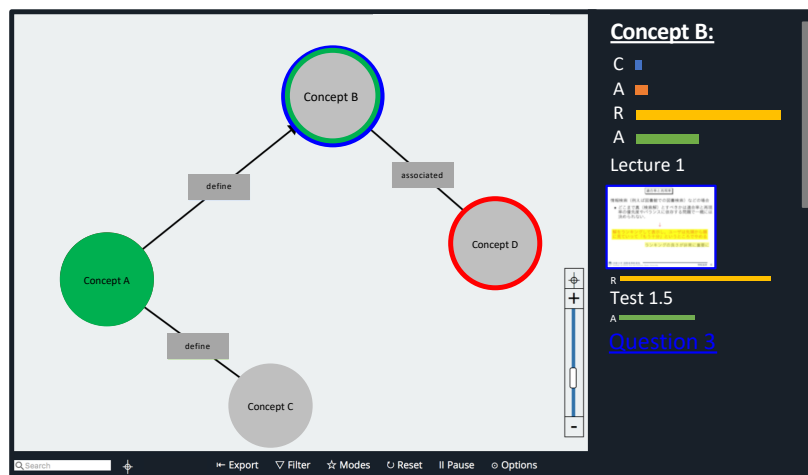


Figure 4: The user interface of the proposed system.

Additional functionality supports the augmentation of the base map structure with analytics results as visual attributes of nodes as was shown in Fig 3, to give users visual cues to the overall knowledge state.

No Attrib	Attended	Read	Credit	
				Not Answered
				Answered

Figure 5: Node visual augmentation definition.

The outline color of a node represents the level of a learner's effort with the relevant learning materials that describe the node concept, and the fill color of the node relates to the learner's knowledge level of the node concept as shown in Fig 5. The degree of coloring in both the outline and fill are displayed to represent the percentile rank of achievement when compared to the whole student cohort. If there is no or very low percentile rank of event data for Attendance/Read/Credit and Answer relating to a node concept, then the outline and fill are displayed as grey.

6 USES OF THE KNOWLEDGE PORTAL

The following section outlines different cases in which the knowledge portal could be used to guide both teaching and learning. An overview of the four main cases is shown in Fig. 6.

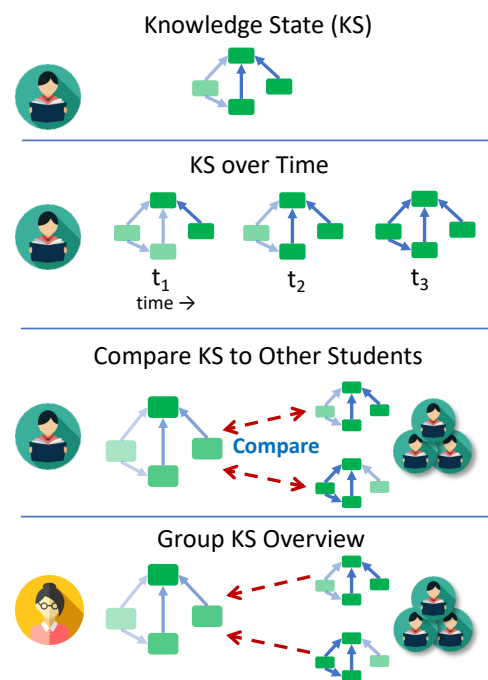


Figure 6: Different cases in which Knowledge Graphs could inform learners and teachers.

The first case is that of a learner confirming their current knowledge state for the support of self-regulated learning. It is intended that the learner could use the knowledge portal for the monitoring and planning of their learning by searching for concepts that they have not yet studied and following the links to appropriate learning materials to reading and test items to confirm their knowledge.

The second case enables the learner to reflect on how their knowledge has evolved over a period of time short or long, such as a student's knowledge at: t_1 = elementary school, t_2 = high school, t_3 = undergraduate university. This could also be used to help students find possible gaps in their knowledge that occurred in the past, and enable the revision of learning materials to resolve knowledge gaps.

The final two cases deal with comparing the knowledge state of groups of learners. For a student, this can enable them to compare their own knowledge to that of the broader student cohort and find possible areas in which their knowledge is lacking. The learner can then study to improve their

knowledge state by working on specific concepts by reading learning materials and testing themselves with linked resources.

Teachers can also benefit from using the proposed knowledge map system to get an overview of the current knowledge state of all of the students in their course. The individual knowledge maps of all of the students are merged into a single aggregated knowledge map. An example use of this would be to check the prior knowledge of students before they attend a lecture, or checking the degree to which students have previewed concepts and the related learning materials to an upcoming class. The teacher then can adjust the lecture to either skip concepts that have been adequately learnt, or focus on concepts that require revision or greater explanation. It is expected that this case will be of particular use when managing courses with large numbers of students.

At a global knowledge map level, the relation between courses could provide insight into what parts of the knowledge map are important and central knowledge to a subject, and highlight what parts are difficult for students to understand and could be incorporated as a filter feature in the knowledge portal. This can be utilized in two different ways: for teachers it gives them an understanding of what knowledge is difficult to understand and may require more thorough explanation, and for students it allows them to see the knowledge that is central to the course and what areas they should pay attention to as it has been difficult for past students.

Knowledge map analysis could also be used in the recommendation of contents both inside and outside the course to learners based on their achievement and focus. Under achieving students may benefit from the recommendation of learning materials that cover concepts that they have yet to master. On the other hand, outperforming students may be interested in exploring extra learning materials outside of the course to expand their knowledge beyond that which would be traditionally offered.

7 CONCLUSIONS

In this paper, we proposed a system to support the creation, management and use of knowledge maps in digital learning environments at a learning analytics infrastructure level. In particular, we proposed processes for the automatic extraction, authoring, storing and use of knowledge maps by students and teachers. For the automatic extraction process, we proposed a text mining method for generating knowledge maps from digital learning materials and conducted a preliminary experiment to evaluate its effectiveness. A key feature of the method is the ability to link extracted concept nodes directly to specific parts of the learning materials from which they were extracted. These links are used to provide not only a reference for users to the original materials, but also as a method of associating learning behavior logs collected in existing system and mapping the analysis of these logs directly onto the knowledge map. This provides feedback to the user about the current learning behavior state overlaid on a knowledge structure.

In future work, the use of the knowledge portal to increase learner knowledge awareness and group formation by knowledge map clustering should be investigated.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 16H06304.

REFERENCES

- Ausubel, D. P. (1963). The psychology of meaningful verbal learning, New York: Grune and Stratton.
- Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). Educational psychology: A cognitive view (Vol. 6). New York: Holt, Rinehart and Winston.
- Chatti, M. A., Muslim, A., & Schroeder, U. (2017). Toward an open learning analytics ecosystem. In *Big data and learning analytics in higher education* (pp. 195-219). Springer, Cham.
- Chen, S. M., & Sue, P. J. (2013). Constructing concept maps for adaptive learning systems based on data mining techniques. *Expert Systems with Applications*, 40(7), 2746-2755.
- Chen, S. M., & Bai, S. M. (2010). Using data mining techniques to automatically construct concept maps for adaptive learning systems. *Expert Systems with Applications*, 37(6), 4496-4503.
- Flanagan, B., Yin, C., Inokuchi, Y., & Hirokawa, S. (2013). Supporting interpersonal communication using mind maps. *The Journal of Information and Systems in Education*, 12(1), 13-18.
- Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy, *Proceedings of the 25th International Conference on Computers in Education (ICCE2017)*, 333-338.
- Hwang, G. J. (2003). A conceptual map model for developing intelligent tutoring systems. *Computers & Education*, 40(3), 217-235.
- Kudo, T. (2006). Mecab: Yet another part-of-speech and morphological analyzer. <http://taku910.github.io/mecab/>.
- Lee, J. H., & Segev, A. (2012). Knowledge maps for e-learning. *Computers & Education*, 59(2), 353-364.
- Lohmann, S., Link, V., Marbach, E., & Negru, S. (2014). WebVOWL: Web-based visualization of ontologies. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 154-158). Springer, Cham.
- Novak, J. D., & Cañas, A. J. (2006). The theory underlying concept maps and how to construct them. *Technical report IHMC CmapTools 2006-01 Rev 01-2008*. Florida Institute for Human and Machine Cognition.
- Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. McGraw-Hill, New York.
- Tseng, S. S., Sue, P. C., Su, J. M., Weng, J. F., & Tsai, W. N. (2007). A new approach for constructing the concept map. *Computers & Education*, 49(3), 691-707.
- Wang, J., Flanagan, B., & Ogata, H. (2017). Semi-automatic construction of ontology based on data mining technique. In *6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 511-515). IEEE.

BoB: A Bag of eBook Click Behavior Based Grade Prediction Approach

Alexander Askinadze, Matthias Liebeck, Stefan Conrad

Heinrich Heine University Düsseldorf, Germany
{askinadze, liebeck, conrad}@cs.uni-duesseldorf.de

ABSTRACT: This paper describes our participation in the LAK 2019 data challenge of predicting student performance. Given a student's clickstream data in the form of actions from the eBook system BookRoll, we predict the score of his or her final test at the end of the course. We propose a method called Bag of Behaviors (BoB) to transform a student's click data into a fixed-size vector by combining a k-Means clustering with localized soft-assignment coding. Using a random forest regressor, we achieve results that are comparable to other aggregation approaches.

Keywords: eBook clickstream behavior, clustering, student performance prediction

1 INTRODUCTION

The Student Watch study (National Association of College Stores, 2018) reported that during the spring term of 2018, 25% of students who purchased at least one course material also bought a digital version. Compared to the spring semester 2016, this represents a growth of 10%. Since digital environment-based learning is steadily increasing, the research area Learning Analytics (LA) is growing in relevance. LA is defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (Siemens, 2010).

The study by Arnold and Pistilli (2012) shows that an automated prediction system can assign a risk score to each student based on their individual data. With the risk level, Purdue University instructors were able to intervene and increased students' retention rates. Since their assessment of risk levels was a success, we are subsequently interested in clickstream behavior as a basis for the automated prediction of student performance as a regression problem.

In this paper, we participate in the LAK 2019 data challenge that deals with the predictions of grades based on clickstream data of an eBook system called BookRoll (Flanagan & Ogata, 2017). A similar challenge was previously carried out in 2018 (Flanagan, Weiqin & Ogata, 2018). The core question considered in our work is: Can clustering be used to create a fixed-size vector representation of the time series click data for the student performance prediction task?

2 RELATED WORK

Experiments on predicting students' final scores using eBook clickstream data were conducted during the 5th ICCE workshop on Learning Analytics (Flanagan et al., 2018). The workshop organizers (Flanagan et al., 2018) reported in their overview paper that most participants of the workshop used

neural networks, random forests, and support vector machines as methods for this task. Additionally, the overview paper reported that the workshop participants often complained about imbalanced data. Hasnine et al. (2018) and Lu, Huang, and Yang (2018) applied oversampling techniques, such as SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) and noise injection, in an attempt to address this issue.

Previous approaches (Hasnine et al., 2018; Askinadze, Liebeck, & Conrad, 2018) on performance prediction based on click data utilized aggregation methods, e.g., sum, mean, or standard deviation, to transform a set of time series data into a single vector per student. In contrast to previous work, we present an approach, which also creates a single vector of fixed length per student, but the raw data is first transformed into a different vector space before applying aggregation methods.

3 METHOD

The core idea of our approach is inspired by the bag-of-words model (BoW) from natural language processing. In this model, text is being represented by a bag. For each word, the number of occurrences in a text is noted. Analogously, the model was also used in computer vision where descriptors—e.g., SIFT features (Lowe, 1999)—of an image are represented by visual words (Csurka, Dance, Fan, Willamowski, & Bray, 2004). In our work, we propose to transfer the model into the research field of learning analytics.

In our case of click data from eBooks, each student s is represented by a time series of actions $X^s = \{x^s_1, \dots, x^s_{n_s}\}$ where n_s denotes the number of actions from student s . Each action $x^s_j \in \mathbb{R}^m$ describes the type of interaction with an eBook, e.g., which page was opened at what time. The clickstream data was derived from xAPI (adlnet, 2017) statements. We applied preprocessing by feature extraction and one-hot encoding so that each action is represented by the features described in Table 1. The relative page number is the only continuous value between zero and one in our feature set and represents where an action was performed in a book. All other features are binary. For example, a one for operationname_ADD_BOOKMARK denotes that a student added a bookmark. We use the features book₁ through book_N to distinguish in which one of the N books an action was performed.

Table 1: Features describing student interactions

operationname_ADD_BOOKMARK	operationname_PAGE_JUMP	operationname_DELETE_BOOKMARK
operationname_ADD_MARKER	operationname_PREV	operationname_DELETE_MARKER
operationname_ADD_MEMO	operationname_SEARCH	operationname_DELETE_MEMO
operationname_BOOKMARK_JUMP	operationname_SEARCH_JUMP	operationname_LINK_CLICK
operationname_CHANGE_MEMO	marker_difficult	operationname_NEXT
operationname_CLOSE	marker_important	operationname_OPEN
devicecode_mobile	devicecode_tablet	xapi_read (PREV, NEXT, PAGE_JUMP
devicecode_pc	book ₁ , ..., book _N	or SEARCH_JUMP)
		relative page number

We present an approach to map the time series $X^s = \{x^s_1, \dots, x^s_{n_s}\}$ to a k -dimensional vector, where k is the same for each student s . Let X^{s_1}, \dots, X^{s_M} be the clickstream data from M students, then the set of all actions can be denoted as $X = X^{s_1} \cup \dots \cup X^{s_M}$. For the creation of our model, actions $X_{\text{train}} \subset X$ of the training subset of the students are taken to perform a k -Means clustering to obtain

k cluster centroids that represent the actions. Since individual clusters contain click actions that are similar regarding a distance measure, we consider the clusters to represent different interaction behaviors. For example, one cluster may contain all read operations at the beginning of book 1, and another cluster centroid may represent all operationname_NEXT events. A priori, it is not possible to determine which clusters will be found and which semantical meaning they bear. Depending on the data, the hyperparameter k must be tuned. We denote the set of all behaviors $B = \{b_1, \dots, b_k\}$ as Bag of Behaviors (BoB) where b_j stands for the j -th cluster centroid.

Now, BoB can be used to transform the actions X^s of student s into a fixed-size vector. For this purpose, we need a function ϕ with $\phi(X^s) \in \mathbb{R}^k$. There are multiple ways of implementing this transformation. We decided to use the localized soft-assignment coding from Liu, Wang, and Liu (2011). With this coding, each student action will be encoded by a subset $B^* \subset B$. This subset is determined by taking the distance from an action x_i^s to all $b_j \in B$ and only taking the nearest l neighbors into account. Let $N_l(x_i^s)$ denote the nearest l cluster centroids to x_i^s , then the localized distance d^* is defined as:

$$d^*(x_i^s, b_j) = \begin{cases} \|x_i^s - b_j\|, & \text{if } b_j \in N_l(x_i^s) \\ \infty, & \text{else} \end{cases}$$

From the time series $X^s = \{x_1^s, \dots, x_{n_s}^s\}$, we derive $h^s \in \mathbb{R}^k$ by setting the t -th dimension of h^s to

$$h^s[t] = \sum_{i=1}^{n_s} \frac{\exp(-\beta d^*(x_i^s, b_t))}{\sum_{j=1}^k \exp(-\beta d^*(x_i^s, b_j))}$$

In case of $b_j \notin N_l(x_i^s)$ and $\beta > 0$, the term $e^{-\beta d^*(x_i^s, b_j)}$ equals 0 since $\lim_{x \rightarrow \infty} e^{-\beta x} = \lim_{x \rightarrow \infty} \frac{1}{e^{\beta x}} = 0$. Since the students' clickstreams have different lengths, we L1-normalized each BoB vector. The function ϕ then transforms each X^s into the resulting vector representing student s with

$$\phi(X^s) = \phi(\{x_1^s, \dots, x_{n_s}^s\}) = \frac{1}{\sum_{t=1}^k h^s[t]} (h^s[1], \dots, h^s[k]) = \frac{1}{\sum_{t=1}^k h^s[t]} h^s \in \mathbb{R}^k.$$

4 EVALUATION

As a vital part of the challenge, several datasets containing clickstream data were provided. For our evaluation, we benchmarked our Bag of Behaviors approach on three datasets: *509a6f75849b* (53 students, 22665 actions in total), *39a67f80f4* (132 students, 207922 actions in total), and *60ab104927* (113 students, 248599 actions in total). The clickstream data was accompanied by the students' scores on the course's final exam. Based on our BoB-representation, we trained a regression model to predict these scores.

In our evaluation, we performed a 3 times 5-fold cross-validation with different seeds and used root-mean-squared error (RMSE) as the evaluation metric for the regression problem. For the regression, we used the random forest (RF) regressor, as well as vectors in the BoB-representation. The results of our approach are listed in Table 2. Additionally, we evaluated the aggregation method approach from Askinadze, Liebeck, and Conrad (2018) using their X_{best} feature set. By directly comparing both approaches, we see that our new feature representation BoB achieves comparable results. By

combining X_{best} with BoB via concatenation of the feature representations, we were able to improve our results across all three datasets slightly.

Table 2: Regression results (RMSE)

Dataset	# students	# actions	BoB	X_{best}	$X_{\text{best}} + \text{BoB}$
509a6f75849b	53	22665	24.79 (∓ 0.53)	24.78 (∓ 1.19)	23.52 (∓ 0.93)
39a67f80f4	132	207922	6.59 (∓ 0.13)	6.62 (∓ 0.08)	6.53 (∓ 0.12)
60ab104927	113	248599	6.13 (∓ 0.12)	6.06 (∓ 0.22)	6.02 (∓ 0.15)

5 CONCLUSION

In this paper, we developed the Bag of Behaviors (BoB) approach, which allows us to transform an arbitrary number of a student's clickstream data into a fixed-size vector. Although we evaluated our approach on students' clickstream behaviors in eBooks, it can also be applied to clickstream data from other e-learning sources.

The evaluation of BoB on dataset 509a6f75849b showed that we achieved RMSE results which are comparable to the results from the aggregation approaches of Lu, Huang, and Yang (2018) and Askinadze, Liebeck, and Conrad (2018). In the future, we will evaluate our approach on more datasets and will perform a more detailed search for hyperparameters. Additionally, we want to experiment with different clustering methods, especially Gaussian Mixture Models (GMM), DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), and agglomerative clustering methods. Furthermore, we want to include temporal information regarding click events.

ACKNOWLEDGEMENTS

This work was partially funded by the IST-Hochschule University of Applied Sciences.

REFERENCES

- adlnet. (2017) xAPI-Spec. Retrieved January 30, 2019, from <https://github.com/adlnet/xAPI-Spec>
- Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267-270). ACM
- Askinadze, A., Liebeck, M., & Conrad, S. (2018). Predicting Student Test Performance based on Time Series Data of eBook Reader Behavior Using the Cluster-Distance Space Transformation. In *Workshop Proceedings. 26th International Conference on Computers in Education (ICCE2018)* (pp. 430-439).
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual Categorization with Bags of Keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV* (pp. 1-22).
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD* (Vol. 96, No. 34, pp. 226-231).

- Flanagan, B., Weiqin, C., & Ogata, H. (2018). Joint Activity on Learner Performance Prediction using the BookRoll Dataset. In *Workshop Proceedings. 26th International Conference on Computers in Education (ICCE2018)* (pp. 487-492)
- Flanagan, B. & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In *International Conference on Computers in Education (ICCE2017)* (pp. 333-338)
- Hasnine, M., Akcapinar, G., Flanagan, B., Majumdar, R., Mouri, K., & Ogata, H. (2018). Towards Final Scores Prediction over Clickstream Using Machine Learning Methods. In *Workshop Proceedings. 26th International Conference on Computers in Education (ICCE2018)* (pp. 399-404)
- Liu, L., Wang, L., & Liu, X. (2011). In defense of soft-assignment coding. In *2011 IEEE International Conference on Computer Vision (ICCV)* (pp. 2486-2493). IEEE.
- Lowe, D. G. (1999). Object Recognition from Local Scale-Invariant Features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 2, pp. 1150-1157). IEEE.
- Lu, O., Huang, A., & Yang, S. (2018). Benchmarking and Tuning Regression Algorithms on Predicting Students' Academic Performance. In *Workshop Proceedings. 26th International Conference on Computers in Education (ICCE2018)* (pp. 477-486)
- National Association of College Stores. (2018) Highlights from Student Watch Attitudes & Behaviors toward Course Materials 2017-18 Report. Retrieved December 3, 2018, from <http://www.nacs.org/research/studentwatchfindings.aspx>
- Siemens, G. (2010). 1st International Conference on Learning Analytics and Knowledge 2011. Retrieved January 16, 2019, from <https://tekri.athabascau.ca/analytics/>

Using machine learning to explore the associations among e-reader operations and their predictive validity of learning performance

Mei-Wen, Nian

National Chiao Tung University, Taiwan
anke801211.eed99@g2.nctu.edu.tw

Yuan-Hsuan, Lee

National Tsing Hua University, Taiwan
jasvi.rms@gmail.com

Jiun-Yu, Wu

National Chiao Tung University, Taiwan
jiunyu.rms@gmail.com

ABSTRACT: While Learning Management Systems (LMS) are largely deployed in higher education. Learning Analytics with Educational Data Mining provides us a powerful tool to extract useful information from those unstructured and messy trace data of students' reading behavior to predict their learning performance. This study aims to investigate the associations among learners' e-reader operations using the dimension reduction and association rule in machine learning and to test the predictive validity of these e-reader operations on learners' performance. 1526 students with their course scores were included from three universities and fifteen courses. With the log files and learning performance datasets, we operationalize learners' e-reader reading behavior by way of the number of ebook operations and reading duration. Five principal components were derived from 17 e-reader operations. The study result showed the more frequent use of the marker-related functions the better the learning performance. Findings of the research can assist instructors and researchers to understand students' latent behaviors in learning with ebooks and can be used to build learning analytics models based on data from learning management systems to predict student performance.

Keywords: machine learning, principal component analysis, predictive validity, e-reader operations, learning performance

1 INTRODUCTION

With the development of technology, students learning progress and their products can be traced and documented on different digital media. For example, researchers have started to pay attention to students' online discussion as indicators of their course performance. Learning management systems (LMS) are frequently applied in blended learning as an extension of the face-to-face classroom to provide course information, assign homework, implement assessment, or serve as a discussion forum. Some researchers also used social media such as Facebook as a platform for blended learning (Wang, Woo, Quek, Yang, & Liu, 2012). The blended nature of learning and instruction led to accumulation of huge amount of student data that is complex and unstructured. Therefore, the use of educational data

mining and machine learning are burgeoning to cluster or classify the student data. For example, in Wu, Hsiao, and Nian (2018), the researchers applied supervised machine learning to classify students' posts and comments on the Facebook learning group and further employed learning analytics to investigate the association between students' messages of the discussion board and their learning performance. In line with the literature, more frequent participation in online discussion is associated with better performance (Wu, Hsiao, & Nian, 2018). Therefore, students' online learning behaviors and traces can be used for modeling student performance.

The current research used data from BookRoll, an e-reader system for students to use in the classroom and in the online learning environment. Students can read the course material on BookRoll uploaded by the instructor. The system have functions including bookmarker, marker, memo, and search and can use these functions for page jump (Flanagan & Ogata, 2017; Ogata et al., 2015). In addition, the system will record users reading events such as the number of pages read, the types of device for reading, and time spent on each operation as well as sequence of operations. The current study would use these data for learning analytics.

This study used datasets from the BookRoll system to investigate the associations among learners' e-reader operations. The principal component analysis was applied to derive latent reading behaviors and to test the predictive validity of these e-reader operations on learners' performance. The research questions in this study include:

RQ1: What are the frequent e-reader operations among students?

RQ2: What are the reading behaviors derived from the principal component analysis (PCA) with the e-reader operations?

RQ3: How are the reading behaviors associated with learners' performance?

RQ4: What is the predictive validity of the reading behaviors on learners' performance?

2 METHOD

The analytic framework of this study consisted of the following stages: 1) exploration of the data structure, 2) generation and selection of the variables, 3) implementation of the principal component analysis, and 4) implementation of the correlation and multiple regression analyses.

2.1 Exploration of the data structure

The data set is comprised of students' e-reader operations and learning performance from 15 courses in 3 universities (3 courses from AU, 10 from KU, and 2 from KyoU). Aside from the 15 e-reader operations described in the manual, two additional operations (BOOKMARK_JUMP and MEMO_JUMP) are included in the data set. Therefore, we would use the 17 e-reader operations for learning analytics. Few students have no data on course performance and thus are excluded for the regression analysis.

2.2 The Generation and Selection of Variables

We calculated the number of each e-reader operation, total duration of reading, and total number of reading pages for each student. Reading time was computed by taking the difference between two operations. These operations include NEXT, PREV, PAGE_JUMP, MEMO_JUMP, BOOKMARK_JUMP, SEARCH_JUMP, OPEN, and CLOSE. Events such as ADD_MARKER and ADD_MEMO were also included. We also conducted outlier analysis to exclude extreme observations, e.g., the operation between NEXT and NEXT was 24 hours apart. Descriptive statistics were computed for the e-reader operations. After confirming the mean, skewness, and kurtosis, ($M=10.80$ hr ($SD=50.03$), skewness=6.81, kurtosis=71.19), we found the observed data did not meet the normality assumption. Thus, we removed the top 5% duration of the corresponding operations.

2.3 Principal Component Analysis

The principal component analysis was performed to reduce the dimension of e-reader operations and to avoid multicollinearity among the operations. We intended to explore the meaningful underlying constructs in students' e-reader operations.

2.4 Correlation and Multiple Regression Analysis

We investigated the association among students' e-reader reading behavior, reading time, and course performance. The principal components derived from the e-reader operations were applied in the multiple regression analysis to test their predictive validity of students' learning performance.

3 RESULTS AND DISCUSSION

3.1 Descriptive Statistics and Correlations

The current sample was from three universities, AU, KU, and KyoU. We calculated the number of each e-reader operation for each student. Students' course grades were standardized within each course. After merging students' course performance and their e-reader operations, we obtained data on 1526 students from 15 courses. There were no duplicate observations across different courses.

Examining the number of each e-reader operations, we found that the most frequent operation was NEXT ($M=731.27$) and PREV ($M=376.56$), followed by ADD_MARKER ($M=39.51$) and PAGE_JUMP ($M=28.67$). The rest of the operations had a mean score less than 10. On average, each student read 8242.92 seconds (or 2.29 hours) of e-book. The average pages of e-book read was 1148.28.

We investigated the pairwise correlations among e-reader operations and course performance. Most of the pairwise e-reader operations were positively correlated ($r = .06 \sim .90$) with possible multicollinearity. Thus, PCA was performed to reduce the dimensionalities and address the issue of multicollinearity. The extracted principal components would then be used in the multiple regression model to predict students learning performance.

3.2 Principal Component Analysis

Principal component analysis with Maximum Likelihood estimation and varimax rotation was conducted with adequate factorization test and index result (KMO = 0.72, Bartlett's K-squared = 153230, df = 16, $p < .001$). Five principal components were synthesized with 66% variance explained from 17 highly-correlated variables (chi square = 2087.26, $p < .001$).

Based on the results of the PCA and functions of the e-reader, we named the five principal components. The first component consisted of OPEN, CLOSE, NEXT, and PREV, which are the basic functions of an e-reader. Thus, we named the first component "basic operation." The second principal component consisted of BOOKMARK and JUMP, which may represent students' intention to revisit specific pages; thus, we named the second component "bookmark and revisit." The third principal component consisted of MARKER-related operation; thus, it was named the "marker operation." The fourth principal component was comprised of search-related functions; thus, it was named the "information search." The fifth principal component was comprised of MEMO-related operation; thus, it was named the "memo operation."

These principal components were reading behaviors, and would then be used to predict students' learning performance in the learning analytical model.

Table 1: Factor loadings for PCA with Varimax rotation.

Operation	Basic operation	Bookmark & revisit	Marker	Information search	Memo	<i>h</i>
OPEN	.89	.14	.01	.04	.16	84%
CLOSE	.89	.14	.02	.05	.14	84%
NEXT	.76	.12	.48	.01	-.11	83%
PREV	.61	.05	.48	-.03	-.21	65%
LINK CLICK	.29	.03	.14	.06	.15	13%
BOOKMARK JUMP	.06	.91	.03	-.02	.07	83%
ADD BOOKMARK	.06	.88	.15	-.01	.10	80%
DELETE BOOKMARK	.12	.65	.27	.04	.01	51%
PAGE JUMP	.47	.63	.10	.22	.02	67%
ADD MARKER	.22	.21	.76	.04	.12	68%
DELETE MARKER	.21	.18	.76	.05	.03	65%
DELETE MEMO	-.02	.04	.43	.13	.38	35%
SEARCH JUMP	.06	.03	.06	.96	.00	93%

SEARCH	.07	.05	.06	.95	.02	92%
CHANGE MEMO	.06	.20	.23	-.03	.71	60%
MEMO JUMP	.09	-.08	-.13	.00	.59	38%
ADD MEMO	.11	.21	.49	.01	.56	61%
Eigenvalue	2.99	2.63	2.24	1.91	1.47	
Var. Explained	18%	15%	13%	11%	9%	

Note. Maximum likelihood estimation with Varimax rotation and Kaiser normalization was used. h = Communality of item. The total variance explained by five principal components is 66%.

3.3 Correlation and Multiple Regression Analysis

The correlations between reading behaviors and standardized course grades were tabulated in Table 2. Among the study variables, only maker operation was significantly correlated with standardized score ($r=.07, p=.004$).

Table2: Correlations of Standardized Score and Principal Component Scores

Variable	Basic operation	Bookmark & revisit	Marker	Information search	Memo
Standardized score	-.02	.02	.07**	.03	.05

Note. * indicates $p < .05$. ** indicates $p < .01$.

Multiple regression analysis was conducted to understand the causal prediction pattern of students' reading behaviors to their learning performance. Five PCA indicators of reading behaviors were used to predict students standardized scores. Result showed that students who use more maker functions would have better learning performance on average ($B_{marker} = 0.072, \beta = 0.073, t = 2.856, p = .004$), while other indicators (i.e. basic operation, bookmark, memo, and search) could not significantly predict learning performance.

Table3: Multiple Regression Results

	B	$SE\ B$	β	t	p	VIF
(Constant)	0.005	0.025		0.207	.836	
PC1: Basic operation	-0.016	0.025	-0.017	-0.653	.514	1
PC2: Bookmark & revisit	0.019	0.025	0.019	0.736	.462	1
PC3: Marker	0.072	0.025	0.073	2.856	.004**	1
PC4: Information search	0.025	0.025	0.025	0.980	.327	1
PC5: Memo	0.047	0.025	0.048	1.864	.063	1

Note: 1.D.V. is standardized score. 2. * $p < .05$, ** $p < .01$.

4 CONCLUSION

This study examined students' latent behavior in reading ebooks as well as the association between the ebook reading behavior and course performance. Data was obtained from BookRoll across 15 university courses consisting of 1526 students' reading progress. On average, each students read 2.29 hours and 1148pages. The most frequent ebook reading events were basic functions such as NEXT (M=731.27) and PREV (M=376.56), followed by ADD_MARKER (M=39.51). We used principal Component Analysis to extract five latent reading indicators of students' reading behaviors, that is basic operation, bookmark and revisit, marker operation, information search, and memo operation. Controlling all other reading behaviors, we found out that students who used the marker functions more frequently in this e-reader system would tend to have higher course grades. Memo operation, which is marginally significant, would be another possible predictor for students' course performance. The analytical results were significant for the marker component; however, the magnitude of the regression coefficient was relatively small. Our study findings suggested that students can have trainings about how to highlight meaningful texts or take self-memo during their reading process to foster deep understanding of the course material that would further improve their learning performance.

Findings of the research can assist instructors and researchers to understand students' latent behaviors in learning with ebooks and can be used to build learning analytics models based on data from learning management systems to predict student performance. In this study, we analyzed the BookRoll dataset from three universities to build a general learning analytics model and find a universal reading behavior to predict students' learning performance. Nevertheless, the data were from 15 courses, which may have different instructional designs and course requirements; thus, it is likely that the effect of reading behaviors and frequency of operations may differ due to differences in course designs and requirements. Future study can be conducted to consider the heterogeneous nature of these courses to yield learning analytics models that fit each individual course or courses of similar kind.

REFERENCES

- Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In International Conference on Computers in Education (ICCE 2017) (pp. 333–338).
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based Learning Analytics in University Education. In International Conference on Computer in Education (ICCE 2015) (pp. 401–406).
- Wang, Q., Woo, H. L., Quek, C. L., Yang, Y., & Liu, M. (2012). Using the Facebook Group as A Learning Management System: An Exploratory Study: Using Facebook Group as An LMS. *British Journal of Educational Technology*, 43(3), 428–438.
- Wu, J.-Y., Hsiao, Y.-C., & Nian, M.-W. (2018). Using Supervised Machine Learning on Large-Scale Online Forums to Classify Course-Related Facebook Messages in Predicting Learning Achievement Within The Personal Learning Environment. *Interactive Learning Environments*, 0(0), 1–16.

Characterization of Fuzziness for Grade Prediction using Deep Neural Networks

Kelvin H. R. Ng

Nanyang Technological University, Singapore

kelvin.ng@ntu.edu.sg

Sivanagaraja Tatinati

Delta-NTU Corporate Lab, Nanyang Technological University, Singapore

tatinati@ntu.edu.sg

Andy W. H. Khong

Nanyang Technological University, Singapore

andy.khong@ntu.edu.sg

ABSTRACT: In this work, we employed deep learning techniques to predict students' grades based on the sequences of their online reading behavior. Based on the students' online experiences, two prediction modes are performed in this work for 1) early detection of at-risk students and 2) continuous monitoring of students' progress in terms of achievable outcomes. Results obtained for both prediction modes highlighted that online reading behavior sequences dataset is multi-valued in nature and the current crop of deep learning techniques are not equipped to deal with this sort of datasets. Furthermore, in this work, we present evidences that underscores the implications in modeling these sequences due to the fuzziness in the education datasets.

Keywords: Deep learning, Grade prediction, Multi-value prediction

1 INTRODUCTION

Instructors in conventional classroom settings often incorporate students' body languages and learning disposition on top of their assessment achievements to evaluate one's learning progress. Due to its low cost and wide reach, since the last decade, online learning environments are becoming prevalent source of education and instructions (Moreno-Marcos, Alario-Hoyos, Muñoz-Merino, & Kloos, 2018). However, in these settings, it is increasingly difficult to evaluate students' learning due to the cyber-physical disconnect as well as the increase in attendees (Daradoumis, Bassi, Xhafa, & Caballé, 2013). In order to focus the attention on at-risk and struggling students, thereby allowing instructors to focus their efforts to provide effective interventions to alter the course of learning, in recent times, learning analytics that are based on machine learning techniques have been developed (Daradoumis et al, 2013). One such framework is to forecast the grades of individuals through their online preparatory activities (Moreno-Marcos et al, 2018). Grade prediction is fundamentally performed using online actions/behaviors from students' interaction with learning resources. These action sequences range from material access and forum participations to fine-grain clickstream interactions with learning materials such as video streams.

Several techniques been developed recently to forecast the course outcomes, such as grades, efforts, learning styles, emotion states. These techniques employ frequencies of interactions (Martínez-Muñoz, & Pulido-Cañabate, 2017) and, more recently, the inclusion of inter-action relationships in

action sequences (Yang, Brinton, Joe-Wong, & Chiang, 2017; Pérez-Lemonche, Martínez-Muñoz, & Pulido-Cañabate, 2017; Brinton, Buccapatnam, Chiang, & Poor, 2016). Grade prediction accuracy using frequency features have shown to vary when the analysis period changes (Ng, Tatinati, & Khong, 2018). On the other hand, instead of aggregating the occurrences of actions, discriminative features can be learnt directly from the action sequences with deep learning techniques. Complex features and relationships extracted by these techniques are more resilient towards these changes and can also easily be transferred onto other prediction tasks.

In this work, students' interactions with digital reading materials are logged as they prepare for class and these action sequences are used to perform grade prediction using various deep neural networks. Two frameworks of grade prediction are performed for early detection of at-risk students and continual monitoring of student progress. The results showed limited prediction performances with all tested models in both frameworks. An in-depth analysis reveals an inherent fuzziness in the dataset that these deep neural networks failed to resolve when trained with conventional approaches in order to differentiate the various grade achievements. Evidences of this fuzziness in education datasets and its implications on the modeling with deep techniques are detailed in the following sections.

2 METHODS AND MATERIALS

Dataset

In this study, interactions with digital reading materials acquired from 1545 students across 15 classes in 3 universities is employed for prediction tasks (Ogata, Yin, Oi, Okubo, Shimada, Kojima, & Yamada, 2015; Flanagan & Ogata, 2017). 19 logged actions captured how students navigated between reading pages, adding bookmarks and memos as well as using markers (marker operators are further differentiated to indicate importance or difficulty). The list of actions is tabulated in Table 1. The actions are ordered temporally into sequences and subsequently used to predict students' grade achievements. On average, these sequences contain 1304 actions.

Table 1: List of logged actions

Operation Types	Navigation Operations	Bookmark Operations	Memo Operations	Search Operation	Marker Operations
Actions	Open Reading Material	Add Bookmark	Add Memo	Search	Add Marker
	Next Page	Delete	Delete Memo	Search Jump	(Important/Difficult)
	Previous Page	Bookmark	Change Memo		Delete Marker
	Close Reading Material	Bookmark Jump	Memo Jump		(Important/Difficult)
	Page Jump				
	Link Click				

Problem Statement: Prediction Tasks

The first framework simulates early detection of at-risk students, denotes this task by Early Detection (ED). This framework is popular with existing grade and drop-out predictions involving massive open online courses (MOOC) due to the prevalent high drop-out and low completion rate

(Moreno-Marcos et al, 2018). With such framework, grade prediction is typically performed when information about one's learning is limited. In this work, the amount of information is limited to the first fifty interactions performed by each student. The latter framework emulates continual detection of at-risk student throughout their learning journeys, denotes this task by Real-time Prediction (RT). As learning behaviors are dynamic in nature and potentially causal towards grade achievements independently, this framework aims to identify how transient behaviors contribute (positively or negatively) towards grade achievements. To emulate this, prediction is performed using sets of contiguous actions. Specifically, sets of 50 actions are extracted from individual students using a sliding window with an overlapping ratio of 50%. All resulting subsequences are associated with the same grade achieved by this student.

Models

Since discrete symbolic inputs such as student interaction logs are not compatible with the continuous numerical space of deep neural networks, the logged actions are first mapped onto numerical vectors using word embeddings. These embeddings are randomly initialized vectors stored in the form of a lookup table (Bengio, 2003). Vectors of 50 dimensions were used to embed these actions.

Grade prediction is performed using two variants of deep networks. The first model follows the architecture of a neural probabilistic language model (Bengio, Ducharme, Vincent & Jauvin, 2003). With this model, input vectors are concatenated into a single vector and the grade is predicted by learning a set of linear weights that map the input to the grade through incremental updates using stochastic gradient descent. Up to three layers of 128, 256, 512, and 1024 hidden nodes were tested and repeated to include 50% drop-out between each layer. Optimal parameters were selected by comparing the average mean absolute error (MAE) achieved by the models from 5-fold cross validation. In this work, training and testing MAE are used to represent the performance of the model and its ability to extract discriminative features and subsequently using these features to perform the prediction for the training set and unseen data, respectively. Three layers of 128 hidden nodes demonstrated a balance between training and testing error. We denote this model as LM.

As the second model, a convolution neural network (CNN) is used. CNNs are good with extracting the local regularities like edges/shapes in images. To exploit this advantage while maintaining contiguity of actions, the numerical vectors of actions from each student are stacked into a 2-dimensional matrix to form an 50-by-50 action image. Two types of kernels were tested – small square kernels like those used for image classification (denoted by Conv), and rectangular kernels where the height matches the embedding length, (denoted by nGramConv). The rectangular kernels are motivated by n-gram models. In a typical n-gram model, extracted word tuples are tabulated for the model learns a probabilistic relationship between word tuple occurrences and the prediction objective. Rather than identifying frequent action pairs, these rectangular kernels identify the locations of frequently occurring action sets. This approach has two advantages. As with CNNs, since kernels are reused throughout the input space, this approach is more memory-efficient compared to maintaining a list of word combinations. Also, as word tuples are represented in a continuous numerical space, word tuple representations associated with similar grade performances can be represented closely in the feature space. For Conv, up to four layers (each with 20 filters of kernel sizes of 3, 5 and 10) were tested and the best model is achieved by 3 layers of 20 3-by-3 kernels with

2-by-2 max-pooling between each layer. For nGramConv, different numbers of kernels representing 3, 4, and 5-grams were tested. The best model is achieved by 10 kernels for each of the 3, 4, 5-grams. Both Models are implemented on Tensorflow and trained for 100 epochs at 0.0001 learning rate with Adam gradient descent optimizer (Kingma, 2014).

3 RESULTS AND DISCUSSIONS

Performance comparison

The performances of various models after training was completed are presented in Table 2. A naïve estimation is computed to compare model performances against the dataset. The naïve estimation is computed as the average MAE when the model predicts the dataset average grades for all inputs. In both ED and RT modes, all three models performed better than naïve for training but not for testing. The best testing errors are achieved by LM and Conv for ED and RT, respectively. The training progresses of individual models, represented by the mean squared errors, are illustrated in Fig. 1 and 2, respectively. While the training errors decreases, the testing errors plateaued after some training epochs. In the case of LM and Conv for RT, the testing errors diverged.

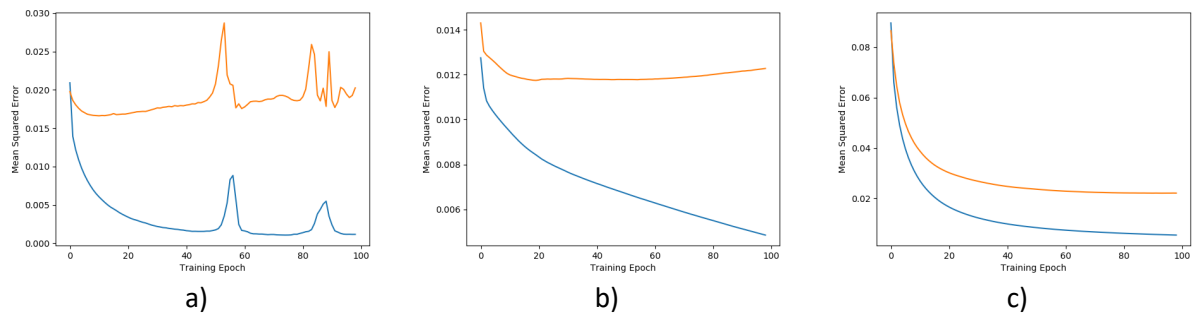


Figure 1: ED mode Training and testing errors for a) LM, b) Conv and c) nGramConv

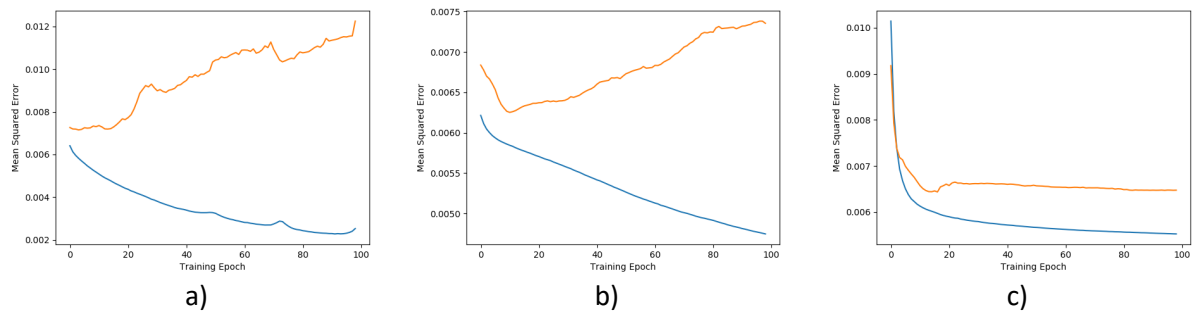


Figure 2: RT mode Training and testing errors a) LM, b) Conv and c) nGramConv

Specifically, it is shown that the absolute errors of the prediction can attain as low as 7% simply by producing the target data average for all inputs. In other cases, where the score assessments are more varied and cover a wider range, the errors may become so large the model is assumed to have not learnt from the dataset. An in-depth analysis is performed in the following to understand why deep learning models are incapable of learning such prediction tasks.

Fuzziness in education datasets and its modeling

The objective of machine learning algorithms is to identify hyperplane that better differentiates the inputs according to their target values. Most of the machine learning techniques, while dealing with nonlinear objective functions, map input features into higher dimensional feature space and identify linear boundaries there to differentiate the inputs. These algorithms can often be re-formulated to be represented by a linear regression model with nonlinear basis functions and learns according to the maximum likelihood (Bishop, 2006). Owing to the normal distribution assumption on modeling errors, identical inputs are expected to have identical target values described by the mean of the distribution while differences in target values are explained by the variance of this distribution. However, when target values are very different the algorithm is forced to increase the variance to provide sufficient coverage. As the spread of the distribution increases, the current estimation is now required to represent new inputs whose target values are more probable under the new distribution. The resulting model may be forced to predict the mean value of the training dataset rather than identify individual grades attained by different inputs.

Table 1: Grade prediction performances for ED and RT.

		Model			
Task		Naïve	LM	Conv	nGramConv
ED	Training	0.0701	0.0594	0.0546	0.0403
	Testing	0.0700	0.0719	0.1091	0.1013
RT	Training	0.0589	0.0559	0.0567	0.0592
	Testing	0.0589	0.0847	0.0598	0.0651

An exploratory analysis performed on the study-interactions dataset shows that it violates this assumption. In this analysis, identical input sets are identified by comparing the composition of actions as well as the order of actions performed by each student. As it is intractable to identify causality between differences in inputs to students' achieved grades, for illustration purposes, only exactly identical inputs are considered in this analysis. Each set of identical inputs are decomposed into input pairs formed by all possible combinations with inputs other than itself within the set. A 2-tuple (g_1, g_2) is created for each input pair, where g_1 and g_2 are the grade achieved by students performing these actions. These 2-tuples are distributed in a 2-dimensional space and the distributions of these tuples are illustrated using heat maps in Figure 1 for both ED and RT tasks.

For both tasks, majority of the input pairs are distributed along the diagonals. Specifically, most of the input pairs occur when the grade is 0.8. For ED, as depicted in Figure 1(a), this is further broken down into two distributions (centered around 0.8 and 0.9). As shown in Figure 1(b), input pairs for RT follows a single distribution of a much larger variance. This implies that hidden representations learnt by models have to be able to predict all possible grades between 0.6 and 1.0 for ED and 0.5 and 1.0 for RT. As such, with high probability, the models would learn the weighted average of the grades based on the distributions. Furthermore, as there are much lesser inputs with grades below 0.5 in both cases, these inputs may be poorly represented in the models.

Machine learning techniques discern inputs that have different prediction outputs and identify similar feature sets for those having similar outputs. Therefore, some inputs, although having some

differences, are represented similarly due to their prediction outputs. We hypothesize that this fuzziness exists in a trained model. As feature sets are refined through the multiple layers of a deep neural network, the output of the last hidden layer is extracted for each input. A similar analysis is performed with these inputs by defining the similarity of input pairs as the sum of square differences between the hidden representations. Only input pairs that are within the first percentile similarity are considered, depicted in Figure 4 (heat maps). The distribution of these similar input pairs follows that in Figure 3. This implies that the hidden representation, although linearly related to a predicted grade, is associated with more than one grade prediction. This in turns result in either 1) some inputs having smaller outputs while others have very large errors or 2) all associated inputs have moderately large errors. In any scenario, the model fails to perform the prediction task well.

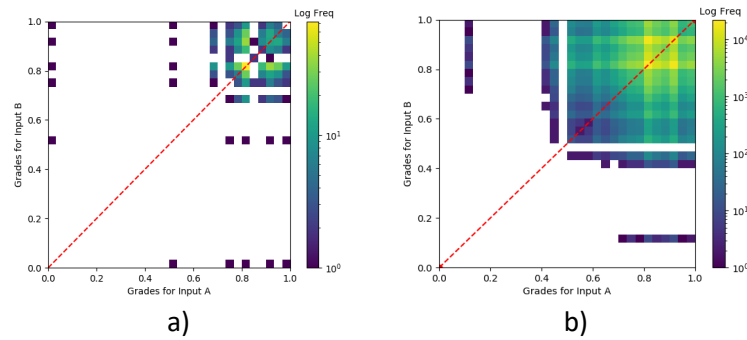


Figure 3: Frequency of grade pairs for identical inputs a) early detection, b) real-time prediction

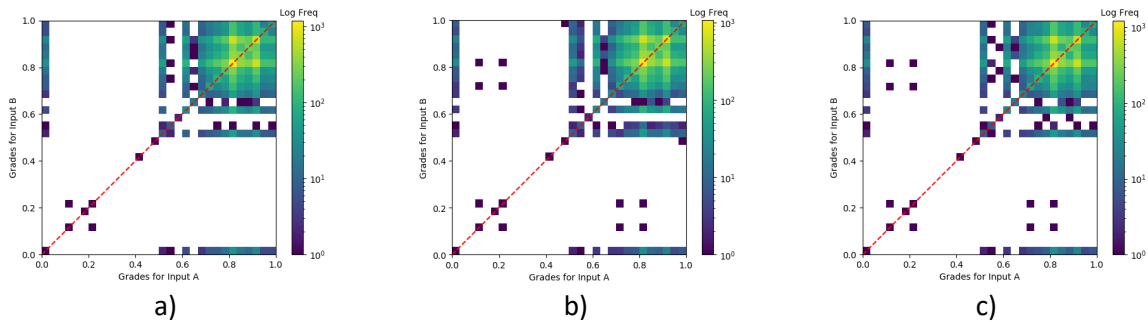


Figure 4: Frequency of grade pairs of exactly identical inputs for a) LM, b) Conv and c) nGramConv

4 CONCLUSION

In this work, we employed deep learning techniques for grade prediction based on the students' online interactions. Results underscore that education datasets are multi-valued by nature. This inherent property can easily be masked when we analyze only the prediction errors produced by the model without taking into consideration the innate distribution of target data. To demonstrate this issue, we conducted an exploratory analysis and highlighted the fuzziness in education datasets and limited capabilities of current crop of machine learning techniques to deal with this fuzziness.

REFERENCES

Angelo, T. A., & Cross, K. P. (1988). *Classroom assessment techniques. A handbook for faculty*, Office of Educational Research and Improvement, Washington, DC.

- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., and Yamada, M. (2015). *E-Book-based learning analytics in university education, Proceedings of the 23rd International Conference on Computer in Education* pp.401-406, 2015.
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb), 1137-1155.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2016). Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance. *IEEE Transactions on Signal Processing*, 64(14), 3677-3692.
- Fei, M., & Yeung, D. Y. (2015). Temporal models for predicting student dropout in massive open online courses. *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop*, (pp. 256-263). IEEE.
- Flanagan B., Ogata H., Integration of Learning Analytics Research and Production Systems While Protecting Privacy, *Proceedings of the 25th International Conference on Computers in Education*, pp.333-338, 2017.
- Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.
- Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*.
- Ng, K. H., Tatinati, S., & Khong, A. W. (2018). Online Education Evaluation for Signal Processing Course Through Student Learning Pathways. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6458-6462). IEEE.
- Pérez-Lemonche, Á., Martínez-Muñoz, G., & Pulido-Cañabate, E. (2017). Analysing Event Transitions to Discover Student Roles and Predict Grades in MOOCs. *Proceedings of the International Conference on Artificial Neural Networks* (pp. 224-232). Springer, Cham.
- Yang, T. Y., Brinton, C. G., Joe-Wong, C., & Chiang, M. (2017). Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(5), 716-728.

Investigating Reading Behaviors within Student Reading Sessions to Predict their Performance

Makhlouf Jihed and Tsunenori Mine

Kyushu University

makhlouf_jihed@yahoo.fr, mine@ait.kyushu-u.ac.jp

ABSTRACT: In this research paper, we predict students' performance using data gathered from their usage of electronic-books-reading platforms. We manage to detect their reading sessions from a stream of their interaction data. Then, we analyze their reading behaviors to build models that predict their performance. We try two different experiments to predict high grades using two different score delimiters. The first delimiter of having high grades is set to 85, then we set it back to 80. For each delimiter, we build a model for predicting which students have high grades. We used genetic programming to build our predictive models. The experimental results shows that we have good predictions for when the delimiter is set to 80, in fact our model attained a ROC AUC of 0.63. But the choice of using 85 as a delimiter was not a good choice as the model have bad results in almost all metrics.

Keywords: Electronic Book, Learning Analytics, Performance Predictions, Reading Behaviors

1 INTRODUCTION

Understanding students' behaviors and provide them with a better learning experience has always been a driving motivation in learning science and educational technologies. Thanks to the continuous increase of the adoption of educational software, these goals are easier to achieve. With the introduction of Information and Communications Technology (ICT) to education, different types of educational software and teaching techniques have grown up. Learning Management Systems, Intelligent Tutoring Systems, Blended learning and many more have been applied to educate people in K12 education. Nevertheless, higher education is also taking advantage of the advances in educational technology.

Learning Management Systems such as Moodle are being used in different educational institutions. They are even part of a bigger infrastructure which include different systems that are in cooperation such as the work of (Flanagan & Ogata, 2018) (Flanagan & Ogata, 2017) where they integrate an e-book learning system called BookRoll (Ogata, et al., 2015) and a system for Share and Reuse Ubiquitous Learning Log named SCROLL (Ogata, Li, Hou, & Yano, 2011) within an integrated system for learning analytics.

These Digital-Learning-Materials Readers are useful in different ways. First, they are a good means of distributing the course material, second, they are a valuable data collection source for learning analytics as it serves to gather students' reading behaviors. Finally, it also provides feedback to teachers about the students' learning experience. They also provide several usability advantages (Ogata, et al., 2017) (Nakajima, Shinohara, & Tamura, 2013).

In this paper, we will explain how we used data coming from the “BookRoll” Online Reading System (Ogata, et al., 2015) to aggregate data and use it to investigate the students’ reading behaviors and use them to predict the students’ performance. We proceed to different features transformations and aggregations, then apply a new approach for feature selection before using genetic programming to find the best machine learning pipeline with its best hyper-parameters to predict ‘high’ or ‘low’ grades depending in two different scores threshold.

2 METHODOLOGY

2.1 Data Acquisition

The dataset we used consists of data files provided by three different universities. For each university, the files are organized by course. The data of each course are split into four different files: Event Stream, Lecture Material, Lecture Time and Quiz Score files. The event stream files are the actions log files representing click-stream interactions of students with the “Book Roll” software; the lecture material files describing the materials used for the respective course; the lecture time files contain the lectures schedule; finally, the quiz scores files represent students’ final scores in the related course.

2.2 Initial Data Analysis

Regardless of the different universities, in overall, the dataset contains almost 2 million rows of events, each one of them describes an action done by the student within the system. Different types of actions are recorded such as a request to open a file, a jump to a specific page, saving a bookmark and many more. The dataset contains 126 lectures, 15 different courses and 1531 unique students.

Since the files have the same structure despite coming from different universities, we inserted them in three different data structures based on the entity described by the file. Therefore, we had a data structure describing the event log containing all the actions of the whole dataset. We, also, merged the lecture files describing the material used and the schedule of the lectures into one data structure. Finally, we created another data structure containing the students’ scores of the whole dataset. We use the whole data set, without taking into account the university, for two reasons. Firstly, there is one university which is responsible for more than 95% of the actions stream data and 86% of the students’ data, accordingly, we did not have enough data points from the other universities to be analyzed separately in a university-based process. Secondly, we wanted to build models that can perform well with data coming from different sources and for that it is better to use the data regardless of the university. Nevertheless, the university information was helpful in the model building, which will be described later in the paper.

2.3 Feature Exploration

For each event, the system stores several information, such as the anonymized student ID, the page number where the action happened, the device (PC or Mobile), the timestamp, the action type and some other information that depends on the action type. The lecture is defined by an id, start and end time, the content used and its number of pages. When it comes to students, the data set contains only their anonymized ID and their score in the respective course.

The main interesting part of the dataset is the actions log data. Each action is characterized by its type. The action type (named 'operationname' in the dataset) is a categorical feature having 17 possible values describing the possible types of actions that the student can perform within the system. Using this feature with other information defining students' actions, we extract students' reading sessions and generate many other features that help us gather more insights about students' reading behaviors.

2.4 Reading Sessions and Feature Transformation

To make predictions related to students' performance, we need to change the granularity of our data from the action level to the student level. But before that, we wanted to investigate students' reading behavior. To this end, we extract the period of time in which a student is reading a specific document and we call that a reading session. Basically, a reading session is related to opening a document and being engaged with it until the student closes it or the time when we detect an inactivity period exceeding a predefined 'Inactivity threshold'. If the student closes the document, then the session is closed normally; if the student is inactive then we terminate the session, but we keep track of the opened document, and we start another session when the student is back using the respective document. Since the student can open multiple documents, he can be engaged in many different reading sessions, but we only close the session when the student doesn't use a specific document during a period of time.

The choice of the adequate inactivity threshold, after which we consider a reading session closed, is subject to some experimentations. We wanted to find the most reasonable value which is not too long that it won't detect the inactivity behavior, but also in the same time don't be too short that we mark students as inactive when they come back to the document shortly after. Therefore, we investigate 4 different values of the inactivity threshold and compare the number of detected reading sessions. We chose 30 minutes, 60 minutes, 90 minutes and 120 minutes as the inactivity thresholds.

In Figure 1, we see how the number of detected reading sessions is influenced by the choice of the inactivity threshold. The first choice, which is 30 minutes, detected the greatest number of reading sessions. But, when we increased that threshold to 60 minutes, we experienced a big reduction of 6% of the number of the detected reading sessions. From this change, we can see that setting the threshold to 30 minutes was very short and many students were labeled 'inactive' and closed their sessions while they came back again to the document and continued their activities shortly after that. Therefore, in the 30 minute threshold we detected more sessions simply because many of them were the same session, but they were split into two sessions due to the small limit of time. So, when we fixed the threshold to 60 minutes, a big number of these wrongly labeled inactive students kept their session open. We continue to investigate another threshold of 90 minutes and we remarked that the reduction in the number of detected sessions was not very significant. In fact, the difference in the number of sessions detected by 60 minutes and 90 minutes is about 1.75%. Moreover, when we selected 120 minutes as the threshold, the reduction was only 0.33% compared to the 90 minute threshold. So, we can say that choosing 120 minutes is somehow high and do not grasp the inactivity of students until they close normally the document. 60 minutes and 90 minutes are credible choices, but we chose 60 minutes. The reason is that 90 minutes is the duration of a

lecture, and it is less likely for students to be inactive concerning the lecture material for the whole period of the class.

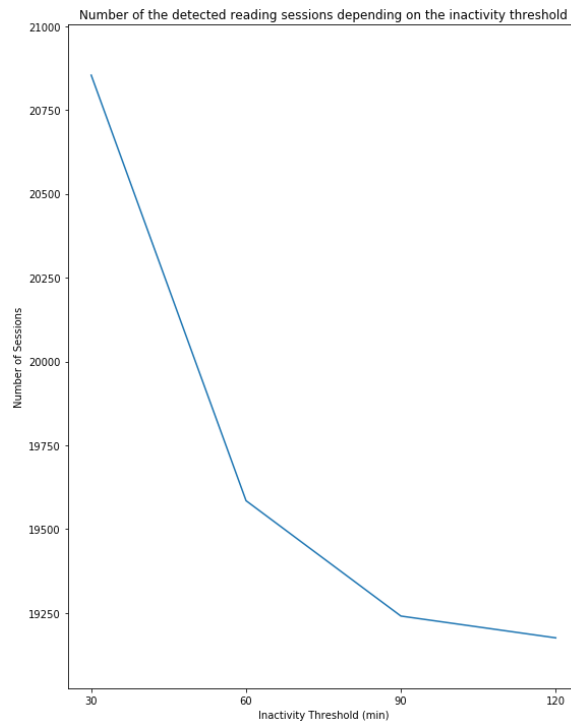


Figure 1 Investigation of different inactivity thresholds

Once we extract the reading sessions, we then define them by:

- The session id, the content (document) id, and the student id
- The start time, which is the time of the first action inside the session
- The last event time, which is the time of the last action within this reading session
- The set of features that we generate using the initial dataset features

For the features generation, we started by counting separately each different type of action within a reading session. Here, we do not count “OPEN” and “CLOSE” actions since they are meant to delimit the reading sessions. We also keep track of the time of each action to check whether the event (thus the reading session) happened in the lecture or not.

Once the reading sessions were detected, we proceed to some features aggregation and transformation for each session as follows:

Table 1 Constructed features and their composition

<i>Column</i>	<i>Meaning and composition</i>
Session length	Session length in seconds

Actions per page	Number of actions divided by the number of pages
Bookmark actions ratio	Number of actions related to bookmarks (add, delete) divided by the number of actions
Bookmark actions per page	Number of actions related to bookmarks (add, delete) divided by the number of pages
Memo actions ratio	Number of actions related to memos (add, change, delete) divided by the number of actions
Memo actions per page	Number of action related to memos (add, change, delete) divided by the number of pages
Link actions ratio	Number of link click actions divided by the number of actions
Link actions per page	Number of link click actions divided by the number of pages
Search actions ratio	Number of actions related to search (action, jump) divided by the number of actions
Search actions per page	Number of action related to search (action, jump) divided by the number of pages
Important actions ratio	Number of important marker actions (add, delete) divided by the number of actions
Important actions per page	Number of important marker actions (add, delete) divided by the number of pages
Difficult actions ratio	Number of difficult marker actions (add, delete) divided by the number of actions
Difficult actions per page	Number of difficult marker actions (add, delete) divided by the number of pages
Browsing actions ratio	Number of 'NEXT' or 'PREV' actions divided by the number of actions
Browsing actions per page	Number of 'NEXT' or 'PREV' actions divided by the number of pages
Jumping actions ratio	Number of jumping actions (from bookmark, memo or page) divided by the number of actions
Jumping actions per page	Number of jumping actions (from bookmark, memo or page) divided by the number of pages

At this level, we have accumulated 44 features. Here it is time to change again the granularity of our data from the session level to the student level. For that, we took the students one by one, and we measured the number of sessions and the total number of actions within the system, then for each other feature, we measure the average across all the student's reading sessions. At the student level, we merge the student's data with their score data structure.

Predicting students' performance had to be made by considering this problem as a classification problem. Thus, we had to transform the numerical score feature to a binary feature. However, we had to choose the score delimiter with which we can distinguish between students who had high grades and who did not. So, we tested different values as the score delimiter, with considering the imbalance of our data.

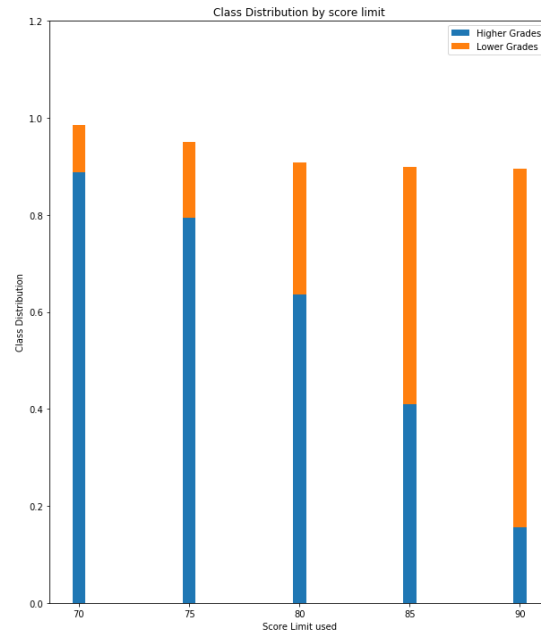


Figure 2 Class distribution depending on the score delimiter

In Figure 2, we see that the score delimiter that obtains the best class balance is 85. However, because it is high, we decided to try both 80 and 85 as score delimiters for the distinction of higher and lower grade students.

2.5 Overall Approaches

For our predictions, we try both score delimiters. For each one, we run separately the same process of feature selection, model training and validation.

2.5.1 Features Selection

With up to 41 features in the student level, we have to reduce the number of features. For that, we try a novel approach consisting of a selection based on three of the most famous features selection techniques. In fact, we use a combination of the Univariate Features Selection, the Forward Features Selection, and the Recursive Features Elimination. Basically, we give a score to each feature based on its rank and whether or not it was selected in the respective features selection method. The aggregation of the score and ranks of all features selection is used then to select the subset of features to be chosen.

We applied this method separately to both approaches (i.e. 80 and 85 score limits). As a result, we had almost the same subset of features extracted. For the 85 score limit we have 12 features selected, and for the 80 score limit, we have 11 features chosen. All of them are the same except for the extra feature which is written in bold in Table 2.

Table 2 Features Selection Results

<i>Avg browsing actions per page</i>	<i>Avg browsing actions ratio</i>	<i>Avg difficult actions per page</i>
<i>Avg difficult actions ratio</i>	<i>Avg important actions ratio</i>	<i>Avg in lecture</i>
<i>Avg jumping actions per page</i>	<i>Avg jumping actions ratio</i>	<i>Avg memo actions per page</i>
<i>Sessions count</i>	<i>Total number of actions</i>	Avg session length

2.5.2 Splitting the data

After the features selection phase, we split the data into training and testing sets. Furthermore, the split is made following a stratified way. In fact, we want to keep the proportions of the label (higher or lower grades) and also the proportion of the university within the splits themselves.

2.5.3 Optimization and Genetic Programming

In order to simplify the process of finding the adequate machine learning technique with its best hyper parameters, we use genetic programming.

Genetic programming is a technique derived from genetic algorithms in which instructions are encoded into a population of genes. The goal is to evolve this population using genetic algorithm operators to constantly update the population until a predefined condition is met. The most common ways of updating the population are to use two famous genetic operators called crossover and mutation. Crossover is used to diversify the research in the research space by taking some parts of the parent individuals and mixing them into the offspring. On the other hand, mutation is the process of updating only some part of an individual and it is used to maintain the actual diversity, in other words, intensify the research in a certain area of the research space. The population is evolving from one generation to another while keeping the fittest individuals in regard to one or many objectives. When using genetic programming for machine learning optimization, we used the model's prediction score as the objective function; the pipeline accuracy score is an example of an objective function which has to be maximized.

In our case, we used genetic programming by searching through a multitude of machine learning techniques and their respective hyper-parameters to find out which combination gives the best results. To achieve our goals we used the python library TPOT (Olson, Bartley, Urbanowicz, & Moore, 2016). However, in order to use genetic programming, there are several hyper-parameters that we need to initialize.

Table 3 Genetic Programming Hyper-parameters

<i>Generations count</i>	<i>Population size</i>	<i>Offspring size</i>	<i>Scoring</i>
200	150	100	ROC AUC
Mutation rate	Cross over rate	Internal Cross Validation	
0.9	0.1	5-fold	

Table 3 explores the principal hyper-parameters that we have to initialize. The Generations count is the number of iterations of the whole optimization process. A bigger number gives better results, but also takes more time to finish. The Population size is the number of individuals which will evolve in each iteration. The offspring size is the number of individuals that are supposed to be generated from the previous population using the genetic algorithm operators. After executing the operators and generating the offspring, the individuals from the population and the offspring compete to survive and be part of the next population. When the individuals compete against each other, we only keep the fittest ones, meaning the individuals with the best score. The method used to measure the score is defined in the scoring hyper-parameters. We used the Area under the Receiver Operating Characteristic Curve (ROC AUC) as our scoring method. That means we only keep the individuals which have the highest ROC AUC values. Mutation and Crossover rates are the probabilities of having respectively a Mutation or a Crossover operation to evolve one or more individuals. We set them to be 90% chance of having a mutation against 10% of having a crossover operation. Finally, the TPOT tool gives us the possibility to cross-validate our pipelines internally, therefore we set the number of folds to 5.

2.5.4 Validation

After the end of the optimization process, the result is a machine learning pipeline and its best hyper-parameters. Since we run the optimization process separately for each score delimiter we have different results. Using the hyper-parameters we train the models using 5-fold cross-validation and measure different performance scores.

Table 4 Scores after testing with the validation data

	<i>Score delimiter of 85</i>	<i>Score delimiter of 80</i>
Machine learning method	Gradient Boosting Classifier	Gradient Boosting Classifier
Accuracy	0.53	0.68
ROC AUC	0.59	0.63
Precision	0.52	0.75
Recall	0.7	0.84

As shown in Table 4, Gradient Boosting Classifiers were chosen after the optimization process. With the score delimiter of 85, it does not have good performance on all metrics. In fact, the accuracy is low attaining 0.53, but the ROC AUC is fair since it attains 0.59. The precision is low too, approaching 0.52 and the Recall is 0.7. While with the score delimiter of 80 the model has better results. Attaining 0.84 in Recall, 0.75 in Precision, 0.63 in ROC AUC and an accuracy of 0.68.

More details are shown in the confusion matrixes in Figure 3 where the values are normalized. With the score delimiter of 85 the rate of True Positive is 0.7 and for the True Negative is 0.36, while the False Positive and False Negative rates are 0.64 and 0.3 respectively. When it is with the score delimiter of 80, the True Positive rate is better, reaching 0.84, but the True Negative rate is 0.25 while the False Positive rate attains 0.75. Finally, the False Negative rate is 0.16.

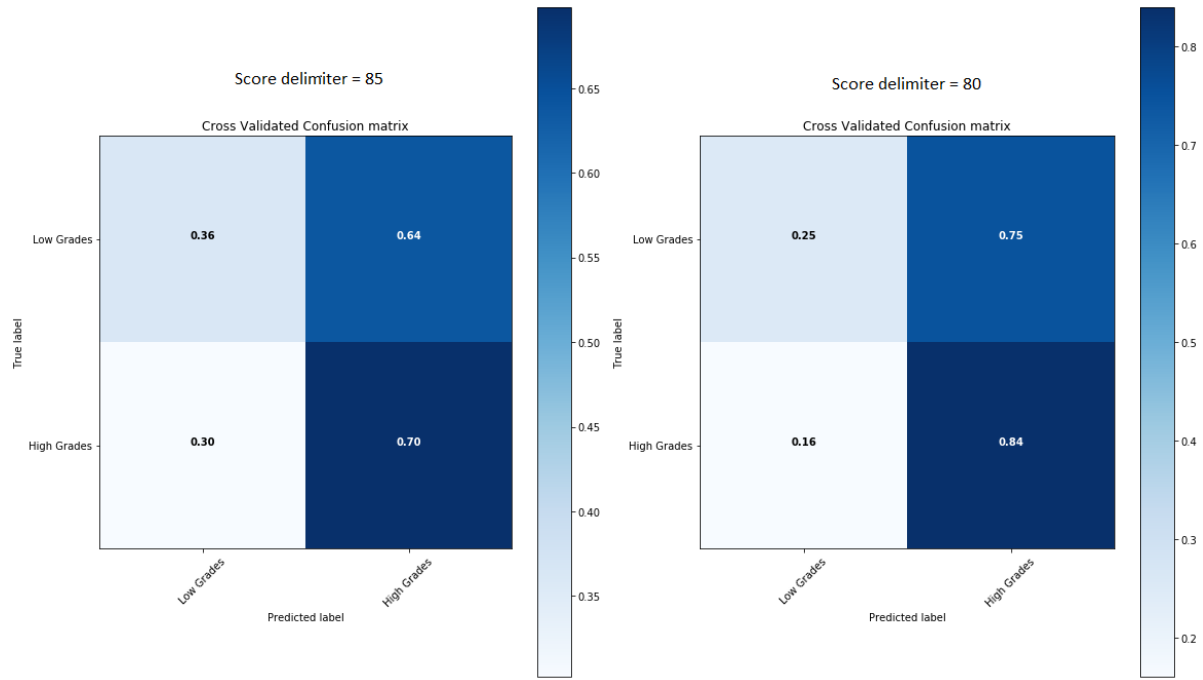


Figure 3 Cross validated confusion matrixes

3 DISCUSSION AND CONCLUSION

Throughout this research, we tried to investigate the students' reading behaviors that can distinguish between students who achieve high scores and those who don't. We used a dataset containing log files of students' interactions with an electronic books reading platform called "BookRoll". From the stream log files, we extracted the students' reading sessions which consist of a period of time in which students engaged with a document without inactivity for a period of time. We transformed the features and aggregated them to build a robust set of predictors. In the feature selection process we used a combination of the three most used features selection methods: The Univariate Features Selection, the Forward Features Selection and the Reverse Features Elimination. The features chosen are different aggregations of the usage of markers, bookmarks, memo, the usage of the next and previous buttons, the reading session length and the number of reading sessions and whether or not the reading session is within a lecture.

Generating the label was also a transformation of the students' grade to a binary feature using a score delimiter. We tried two different delimiters while running the whole process of feature selection, optimization, training, and validation separately for both. The optimization phase allowed us to get the best machine learning pipeline with the best hyper-parameters using genetic programming.

Using a score of 85 as a delimiter gave us a balanced class distribution of higher and lower grades, but in fact, it was as if we reformulated the problem to predict which students are excellent using the dataset. Further, using this approach actually deluded the difference that the features had between 'good' and 'bad' students in terms of their behavior. Furthermore, students who have less than 85 and more than 80 are numerous and share several traits with students who have more than

85. That's why the model build to predict students with 'high' or 'low' grades using the delimiter set to 80 had better performance since the difference in features is more significant.

However, the confusion matrix results suggest another problem, since we have a quite high rate of False Positives. This can perhaps be explained by some overfitting which can be fixed with some dimensionality reduction. The dilemma is to pick up which features to be removed since the features were chosen after using a combination of 3 famous features selection techniques. Since we aggregated some features depending on the number of actions of the students in the corresponding session, and the same features with the number of pages of the content (e.g. Avg browsing actions ratio and Avg browsing actions per page), it would be interesting to investigate which type of aggregation gives better results while simultaneously reducing the dimensionality of the dataset.

4 ACKNOWLEDGEMENT

We would like to thank the data challenge organizing committee for giving us the opportunity to work with the dataset.

This work is partially supported by JSPS KAKENHI No.JP16H02926, JP17H01843 and JP18K18656.

REFERENCES

- Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. *Proceedings of the 25th International Conference on Computers in Education (ICCE2017)*, (pp. 333-338).
- Flanagan, B., & Ogata, H. (2018). Learning Analytics Infrastructure for Seamless Learning. .
- Nakajima, T., Shinohara, S., & Tamura, Y. (2013). Typical Functions of e-Textbook, Implementation, and Compatibility Verification with Use of ePub3 Materials. *Procedia Computer Science*, (pp. 1344-1353).
- Ogata, H., Li, M., Hou, B. U.-B., & Yano, Y. (2011). SCROLL: supporting to share and reuse ubiquitous learning log in the context of language learning. *Research and Practice in Technology Enhanced Learning*.
- Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., & Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In *Smart Sensors at the IoT Frontier* (pp. 327-350). Springer.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-book-based learning analytics in University education. *Proceedings of the 23rd International Conference on Computers in Education (ICCE 2015)*, (pp. 401-406.).
- Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 485-492). ACM.

Investigating Subpopulation of Students in Digital Textbook Reading Logs by Clustering

Christopher C.Y. Yang¹, Brendan Flanagan¹, Gökhan Akçapınar^{1,2}, Hiroaki Ogata¹
Kyoto University¹, Hacettepe University²
yang.yuan.57e@st.kyoto-u.ac.jp, flanagan.brendanjohn.4n@kyoto-u.ac.jp,
akcapinar.gokhan.2m@kyoto-u.ac.jp, ogata.hiroaki.3e@kyoto-u.ac.jp

ABSTRACT: The increasing volume of student reading logs from virtual learning environment (VLE) provides opportunities for mining student' engagement pattern in digital textbook reading. In order to mine and measure students' engagement pattern, in this paper, we extract several students' reading interaction variables from the digital textbook as metrics for the measurement of reading engagement. Moreover, in order to explore the presence of subpopulation of students that can be differentiated based on their engagement patterns and academic performances, we cluster students into different groups. Students are clustered based on their reading interactions such as total session of reading, total notes adding, etc. Accordingly, we identify students' engagement patterns from different groups based on the clustering analysis results. Several student subpopulations such as low engagement high academic performances and low engagement low academic performances are identified based on students' reading interaction characteristics by clustering analysis. The obtained results can be used to provide researchers with opportunities to intervene in the specific group of students and also an optimal choice for student grouping.

Keywords: Student engagement pattern, academic performance, clustering, digital textbook

1 INTRODUCTION

1.1 Student Engagement Pattern

Student engagement can be considered as the extent of students' involvement and active participation in learning activities (Cole & Chan, 1994). In addition, student engagement through active classroom participation is an important ingredient for learning that has many educational benefits for students (Berman, 2014; Lippmann, 2013; Kuh, 2009). Hence educational data mining (EDM) techniques help researchers with the extraction of students' behavioral features in various domains including e-book reading, MOOCs learning, etc. Moreover, reading interaction variables representing student engagement have been used to prove the relation to self-regulated learning theory (Yamada, Oi, & Konomi, 2017). Therefore, in this paper, we extract several reading interaction variables as metrics for the measurement of reading engagement in digital textbook. We then analyze students reading engagement pattern and the corresponding academic performance (test scores given by the lecturers during lecture time).

1.2 Student Grouping

In terms of exploring subpopulation of students in higher educational domains, researchers often face problem on how to properly, comprehensively group students according to different demands based on tracking logs or self-report assessments. In context of learning analytics, the combination of

students with different learning styles in specific groups may have in the final results of the tasks accomplished by them collaboratively (Alfonseca et al., 2006). Therefore, many of the researchers applied clustering algorithms for optimal student grouping such as k-means (Kizilcec, Piech, & Schneider, 2013) or Ward's method (Pardo, Han, & Ellis, 2017) in order to explore a subgroup of learners with specific learning pattern in the context of digital textbook reading, MOOCs learning, and Self-Regulated Learning (SRL) theory. Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets (Connolly, 1999). The goal of clustering is to identify structure in dataset by objectively organizing data into homogeneous groups where the within-group-object similarity is minimized, and the between-group-object dissimilarity is maximized (Liao, 2005).

In this paper, we group students by a standard centroid-based clustering algorithm k-means method, to explore the presence of subpopulation of students that can be differentiated based on their interaction characteristics and academic performances in digital textbook reading. Moreover, we identify subpopulation of students based on their engagement pattern observed in clustering analysis results.

1.3 Digital Textbook System

BookRoll is a digital textbook reading system which is able to offer many kinds of interaction between users and system, including adding memos and highlighting text, etc (Flanagan & Ogata, 2017; Ogata et al., 2015). In BookRoll, student reading behaviors can be tracked and recorded into the learning analytics system (Flanagan & Ogata, 2017). By analyzing students' reading interactions recorded in BookRoll, in this paper, we expect to answer the following two research questions:

1. How many subpopulations of students can be identified based on reading interactions?
2. How do students' academic performances differ in different subpopulations of students?

2 METHOD

2.1 Data Collection and Variable Extraction

In this paper, we cluster and explore students' engagement pattern in digital textbook reading based on their reading interaction variables and identify the subpopulation of students based on reading characteristics. We used KU dataset¹ which is one of the given datasets that contains around 1.9 million students' click-stream reading events from ten classrooms with totally 1326 students. All classrooms used the same learning materials and quizzes. Students' reading events are collected by BookRoll system. In KU dataset, students from ten classrooms were provided the same learning contents with the same curriculum designs during the semester. Therefore, we combined ten classrooms into one then compared students reading interactions. Moreover, in order to analyze students' engagement pattern and the corresponding academic performance in digital textbook, we extracted seven variables from reading events collected in BookRoll as shown in Table 1. We also included students' test scores (academic performance) as one of the variables for clustering.

¹ <https://sites.google.com/view/lak19datachallenge>

Furthermore, since we wanted to obtain a better distribution of population for the following clustering analysis, a two-stage approach for the outlier removal when using k-means (Hautamäki et al., 2005) was performed. The first stage consist of purely k-means process, while the second stage iteratively removes vectors which are far away from the cluster centroid, resulting of 9 outliers were removed from 1326 students.

Table 1: Description of digital textbook reading variables (N=1317).

Variable	Description of Variable	Average	SD
SESSION	Total number of reading session	16.30	7.60
NEXT	Total times students turn to next page	856.26	468.00
PREV	Total times students turn to previous page	425.84	320.66
PREV/NEXT	Clicking ratio of PREV and NEXT	0.46	0.17
NOTE	Total times students add notes	78.08	120.14
SEARCH	Total times students search for contents	1.27	3.82
JUMP	Total times students jump to another page	36.69	3.82
SCORE	Students' test score given by lecturers	83.70	7.82

2.2 Clustering Analysis

In this paper, k-means method (MacQueen, 1967) from Python packages was applied to cluster 1317 students into different groups based on their digital textbook reading variables as shown in Table 1. Reading variables from 1317 students were normalized in advance by using Z-score normalization. We determined the optimal number of clusters for k-means method by applying Elbow method which is one of the most popular method for determining the optimal number of clusters in a data set (Ng, 2012). The Elbow method maps the within-cluster sum of squares onto the number of possible clusters. The location of the elbow in the resulting plot suggests an optimal number of clusters objectively. We then computed the average score for each individual cluster for the representation of the corresponding academic performance. The optimal number of clusters by Elbow method and the results of clustering analysis are shown in Figure 1 and Table 2 and explained in the next section.

3 RESULTS AND DISCUSSIONS

In this section we present the optimal number of clusters determined by Elbow method and results of clustering analysis. By applying Elbow method for the optimal number of clusters, we obtained several possible optimal numbers of clusters which were 2, 5, and 8 as shown in Figure 1. We then clustered students' reading interaction variables based on those obtained number of clusters accordingly. We finally chose 5 as the optimal number of clusters since we observed the most explainable results of students' engagement pattern and corresponding academic performance. Based on the optimal number of clusters, we clustered 1317 students into 5 groups, the average value and standard deviation of each variable for each group are shown in Table 2. The number of students from cluster 1 to cluster 5 are 512 (38.9%), 177 (13.4%), 256 (19.4%), 338 (25.7%), 34 (2.6%), respectively. As shown in Table 2, we identified 5 student subpopulations based on the engagement patterns in digital textbook reading and the characteristics of each subpopulation of students are described below.

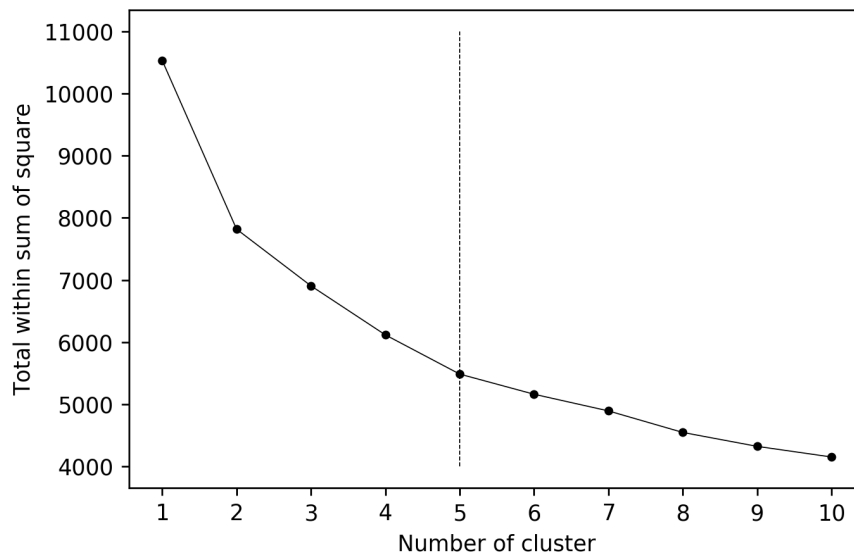


Figure 1: Optimal number of clusters by Elbow method

Table 2: Students' reading engagements and academic performances in different cluster (N=1317).

		Average (SD)							
Cluster	n	SESSION	NEXT	PREV	PREV/NEXT	NOTE	SEARCH	JUMP	SCORE
1	512	12.02	548.97	236.20	0.41	35.36	0.38	22.89	87.78
		(4.45)	(233.51)	(129.84)	(0.15)	(58.94)	(1.14)	(15.74)	(4.58)
2	177	25.48	1333.75	586.34	0.43	271.53	1.49	88.77	85.73
		(8.93)	(471.62)	(305.92)	(0.12)	(174.28)	(2.44)	(48.85)	(6.00)
3	256	14.74	625.61	231.15	0.36	27.54	0.46	27.70	72.97
		(5.95)	(253.32)	(136.90)	(0.15)	(51.94)	(1.33)	(20.74)	(6.52)
4	338	18.46	1225.40	772.90	0.63	74.20	1.18	32.48	84.34
		(5.81)	(382.17)	(297.66)	(0.10)	(82.17)	(2.14)	(18.35)	(6.11)
5	34	23.09	1064.91	461.65	0.43	133.47	20.50	83.00	86.21
		(10.16)	(418.82)	(255.66)	(0.15)	(133.43)	(8.44)	(43.66)	(5.68)

Cluster 1: Students in cluster 1 engaged the least on digital textbook reading compared to other four groups such as session reading, contents searching, etc. Surprisingly, students in this group obtained the highest academic performances as shown in Table 2. For now, we do not know the reason of it, still, the observation in this cluster showed us the subpopulation of Low Engagement High Academic Performance.

Cluster 2: Students in cluster 2 engaged more on session reading, NEXT events, note adding, and page jumping compared to other groups. Students in this group obtained similar academic performances to cluster 1. The observation in this cluster showed us the subpopulation of High Engagement (SESSION, NEXT, NOTE and JUMP) High Academic Performance.

Cluster 3: Students in cluster 3 also engaged very few on digital textbook reading compared to cluster 2, 4, and 5, such as sessions of reading, note adding, etc. Unsurprisingly, students in this group obtained the worst academic performances as shown in Table 2. To mention an interesting finding in this paper, the clicking ratio of PREV event and NEXT event (PREV/NEXT) in this group is significantly lower than other groups as shown in Table 2, indicating that students in this group tended to turn to next page frequently but rarely turned back to previous pages for review while reading. The observation in this cluster showed us the subpopulation of Low Engagement Low Academic Performance.

Cluster 4: Students in cluster 4 engaged more on NEXT event PREV events and clicking ratio of PREV events and NEXT events, indicating that students in this group tended to turn to next page frequently and also turned back to previous pages frequently for review while reading. Although students in this group engaged not as much as cluster 2 and cluster 5 on session reading, note adding, and page jumping, they engaged more comprehensive than cluster 1 and cluster 3 and the academic performances are similar to cluster 1, cluster 2 and cluster 5. The observation in this cluster showed us the subpopulation of High Engagement (NEXT, PREV and PREV/NEXT) High Academic Performance.

Cluster 5: Students in cluster 5 engaged more on sessions of reading, note adding, contents searching, and page jumping compared to other groups. Students in cluster 5 obtained similar academic performances to cluster 1, cluster 2, and cluster 4. The observation in this cluster showed us the subpopulation of High Engagement (SESSION, NOTE, SEARCH and JUMP) High Academic Performance.

4 CONCLUSION

In this paper, we investigated subpopulation of students in digital textbook reading. Students' engagement pattern and the corresponding academic performance in digital textbook reading are analyzed by applying k-means algorithm for clustering. We clustered 1317 students into 5 different groups based on reading variables extracted from BookRoll. To answer two research questions above, we identified 5 students' reading characteristics to represent different subpopulation of students in digital textbook reading which are Low Engagement High Academic Performance, High Engagement (SESSION, NEXT, NOTE and JUMP) High Academic Performance, Low Engagement Low Academic Performance, High Engagement (NEXT, PREV and PREV/NEXT) High Academic Performance, and High Engagement (SESSION, NOTE, SEARCH and JUMP) High Academic Performance. The results showed us that subpopulation of students in digital textbook reading can be identified by clustering students into different groups as students' engagement patterns and academic performances differ while learning. Lastly, the obtained results provide researchers opportunities to find homogeneous groups for collaborative group activities and also demonstrated the importance of student grouping with respect to learning analytics. As an implication, we hope that the results provide chances for instructors to consider different kinds of intervention for the improvement of engagement for different subpopulation of students in digital textbook reading.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 16H06304.

REFERENCES

- Alfonseca, E., Carro, R. M., Martín, E., Ortigosa, A., & Paredes, P. (2006). The impact of learning styles on student grouping for collaborative learning: a case study. *User Modeling and User-Adapted Interaction*, 16(3-4), 377-401.
- Berman, R. A. (2014). Engaging students requires a renewed focus on teaching. *Chronicle of Higher Education*, 61(3), 28-30.
- Cole, P. G., & Chan, L. (1994). *Teaching principles and practice*. Prentice Hall.
- Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687
- Flanagan, B., Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In International Conference on Computers in Education (ICCE2017) (pp.333-338).
- Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T., & Fränti, P. (2005, June). Improving k-means by outlier removal. In *Scandinavian Conference on Image Analysis* (pp. 978-987). Springer, Berlin, Heidelberg.
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM.
- Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New directions for institutional research*, 2009(141), 5-20.
- Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11), 1857-1874.
- Lippmann, S. (2013). Facilitating Class Sessions for Ego-Piercing Engagement. *New Directions for Teaching and Learning*, 2013(135), 43-48.
- MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
- Ng, A. (2012). Clustering with the k-means algorithm. *Machine Learning*.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In International Conference on Computer in Education (ICCE 2015) (pp. 401-406).
- Pardo, A., Han, F., & Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, 10(1), 82-92.
- Yamada, M., Oi, M., & Konomi, S. I. (2017). Are Learning Logs Related to Procrastination? From the Viewpoint of Self-Regulated Learning. *International Association for Development of the Information Society*.

Using Learning Analytics to Detect Off-Task Reading Behaviors in Class

Gökhan Akcapinar^{1,2}, Mohammad Nehal Hasnine¹, Rwitajit Majumdar¹,
Brendan Flanagan¹, Hiroaki Ogata¹

Kyoto University¹, Hacettepe University²
akcapinar.gokhan.2m@kyoto-u.ac.jp

ABSTRACT: In this paper, we aimed at detecting off-task behaviors of the students by analyzing logs from a digital textbook reader. We analyzed 47 students' reading logs from a 60-minutes long in-class reading activity. During the preprocess, we extracted each student's reading patterns as a single vector. Then we used cluster analysis to find the most common reading patterns. Our results indicated that there are two major reading patterns in data. The first pattern is, the students who are following the instructor from the beginning until the end of the lecture. The second pattern is, students who are following the instructor's pattern until the first 17th minute but not during the rest of the lecture. Based on these patterns we labeled first group as *on-task* students while the other group as *off-task* students. We also investigated academic performance of students in these two groups. Obtained results can be used to design data-driven support for in-class teaching. Instructors can plan interventions when off-task behaviors occur while the lecture is in progress.

Keywords: learning analytics, educational data mining, in-class decision making, off-task behavior, reading pattern analysis, clustering

1 INTRODUCTION

Off-task behaviors can be defined as any actions that a student exhibit in the learning environment that are not according to the tasks given by the lecturer (McElroy-Yeider & Courtney, 2016). Off-task behavior is a common problem that intelligent tutoring systems and traditional classrooms often face (Hughes, 2010). According to Hofer (2007), there are two types of off-task behaviors in traditional classrooms, that are: active and passive. Active off-task behaviors include physical activities that students exhibit in a learning environment which often considered to be distributing to their surroundings and consequently effects teaching process negatively (e.g. disturbing other students, making noise, etc.). On the other hand, passive off-task behavior means that students are cognitively disengaged from ongoing learning activities (e.g. daydreaming, texting to other students etc.). Passive off-task behaviors may be harder to notice since students are not disturbing their surroundings (McElroy-Yeider & Courtney, 2016).

With regard to online learning environments, abovementioned problems remain when technology is used to support in-class learning. In addition, devices like computers, mobiles, tablets etc. can be a reason of distraction because students may play games, use other applications, and browse internet (Hughes, 2010).

Both active and passive off-task behaviors require teachers' attention that can lead to frustration for teachers and limit the learning scopes within a classroom (Hofer, 2007). Engaging with off-task

behaviors has also been shown to be associated with poor learning (Baker, Corbett, Koedinger, & Wagner, 2004; Cocea, HersHKovitz, & Baker, 2009). Therefore, both traditional classrooms and online learning environments should consider reducing off-task behaviors while promoting on-task behaviors.

Previous researches have focused on developing detectors for off-task behaviors for intelligent tutoring systems (Cetintas, Si, Xin, & Hord, 2010; Walonoski & Heffernan, 2006). However, without using biological sensors such as eye trackers or EEG headsets, detecting off-task behaviors in traditional classrooms is a challenging task (Baker, 2007). In this paper, we aimed at detecting passive off-task behaviors in a classroom setting by using students' reading logs that were collected from a digital textbook reader.

2 METHOD

2.1 Data

As the data source, we used reading logs collected from a 60-minutes long in-class activity. In the class, there were 47 students. Both students and instructor used the digital textbook reader (BookRoll) during the lecture. BookRoll is a system that allows to view digital materials used for delivering lecture (Ogata et al., 2018). It is an online environment that allows teachers to upload contents as pdf file. Students can browse anytime and anywhere from a web browser in their personal devices (computer or smartphone).

In the BookRoll system, there are features like bookmark, markers, memo function etc. that students can use for learning. In the collection of data for this study, students used their mobile devices or laptops to access the BookRoll system. Reading logs collected automatically by the learning analytics system developed by Flanagan and Ogata (2017). After 60 minutes learning session, 4430 rows of click-stream were recorded in a database that are related to students' interaction with the system. At the end of the lecture, students took part in the quiz session related to content.

2.2 Preprocess

The collected click-stream data contained the following fields: *userid* (anonymized student userid), *contentsid* (the id of the e-book that is being read), *operationname* (the action that was done, e.g. open, close, next, previous, jump, add marker, add bookmark, etc.), *pageno* (the current page where the action was performed), *marker* (the reason for the marker added to a page, e.g. important, difficult), *memo_length* (the length of the memo that was written on the page), *devicecode* (type of device used to view BookRoll, e.g. mobile, pc), and *eventtime* (the timestamp of when the event occurred). For the analysis, we used *eventtime* and *pageno* columns. We grouped the data into 1-minute intervals, and extracted the pages for each student for all time intervals. If a student does not have a log for the specific time interval, we assumed that the student is in the same page where s/he was in the last time.

2.3 Data Analysis

For the data analysis, first we visualized all students' reading patterns. Later, we calculated relative reading patterns of all students. To do this we took instructor's reading pattern as a baseline since

expected reading behavior of students is to follow the instructor during the lecture. Finally, to find off-task reading behaviors we used cluster analysis. Since we do not have prior knowledge about the number of clusters in the data, we conducted GAP statistics (Hastie, Tibshirani, & Walther, 2001) to find the optimal number of clusters. Students took part in the open-book quiz during the last 15 minutes of the lecture, therefore reading patterns of the students during this time is varying. We eliminated quiz part and limited our cluster analysis with the first 45 minute of the course.

3 RESULTS

Visualization of reading patterns of all students can be seen in Fig.1. In the Fig.1, X-axis shows the time, Y-axis shows the page of the books. Intersection of the Time and Page shows the current page of the student in a specific time. Each line shows reading patterns of the different students. Expected reading pattern is, to increase the number of page as the time progresses. As observed from the Fig.1, while most of the students are following the this expected pattern, there are some students who has different reading patterns.

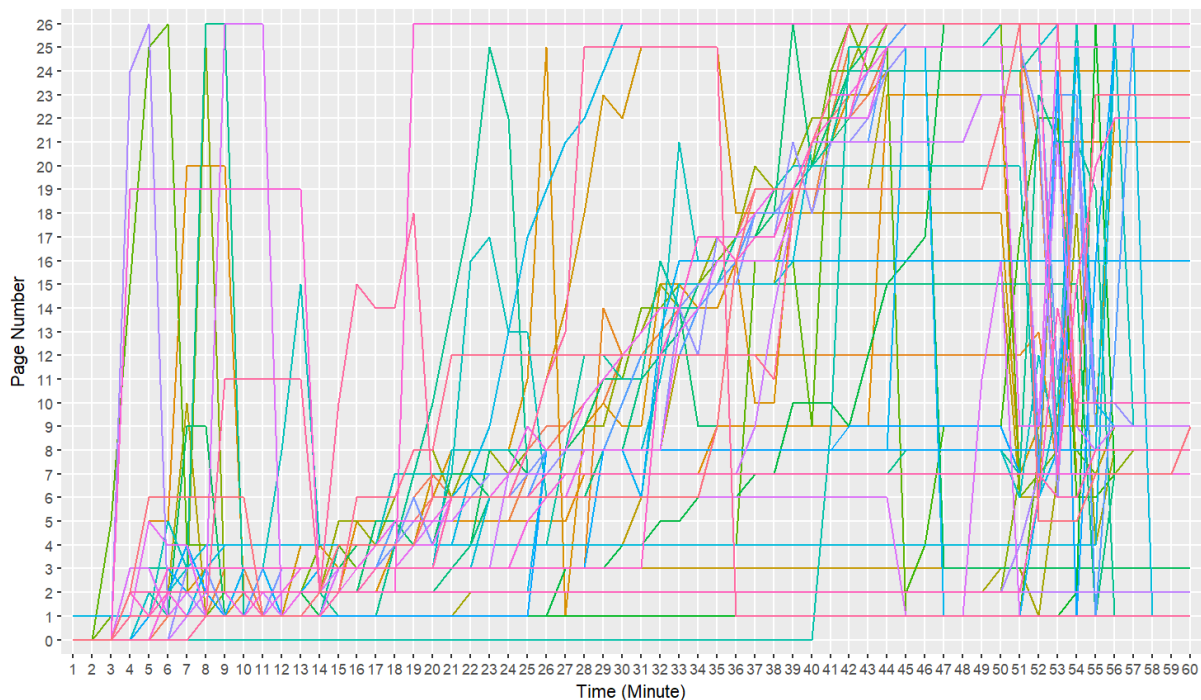


Figure 1: Students' reading patterns across the lecture

To standardize students reading patterns, we calculated relative reading patterns. For instance, if a student is in page 4 while the instructor is in page 6, that student's relative distance will be -2. If a student is in page 8 that student's relative distance will be +2. And if the student is in page 4 (same page as instructor) it will be 0. Results of this calculation is shown in Fig.2. Here again X-axis shows the time of the lecture, while Y-axis shows the students relative distance from the page where instructor is currently in. After calculating students' relative distances from the instructor's pattern, we conducted cluster analysis to find the common reading patterns.

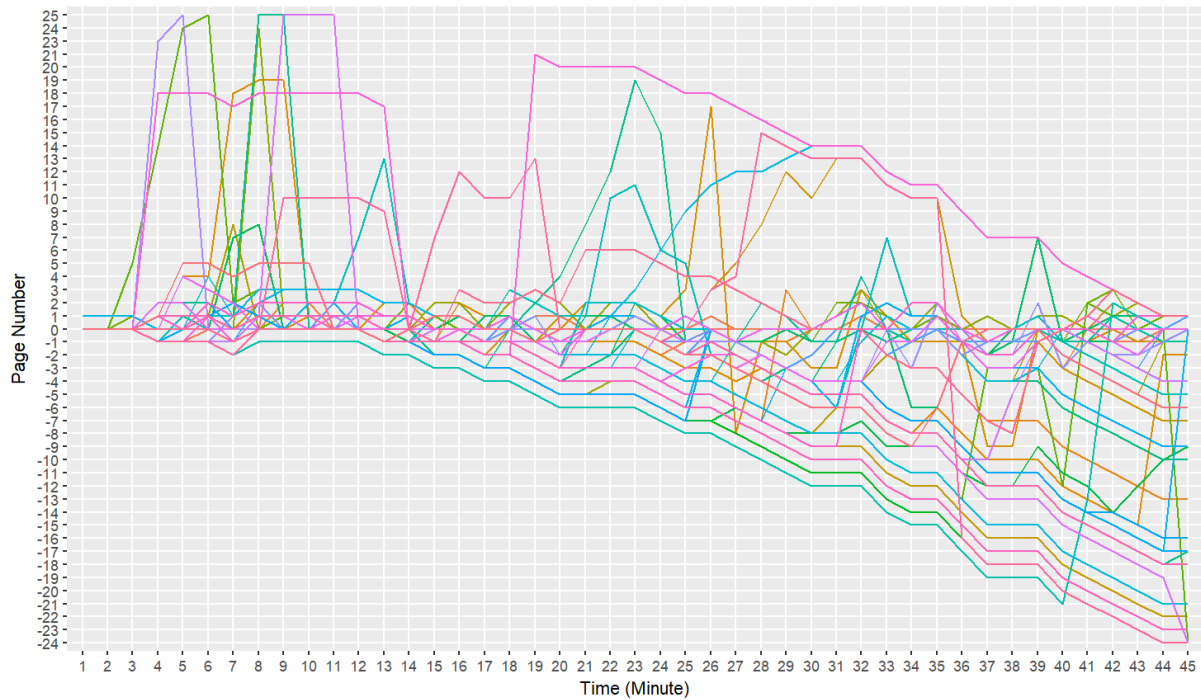


Figure 2: Relative reading patterns of students

3.1 Cluster Analysis Results

Results of the GAP statistics can be seen in Fig.3 (left). According to the results, optimal number of cluster was found as 2. Fig.3 (right) shows the visualization of these two clusters.

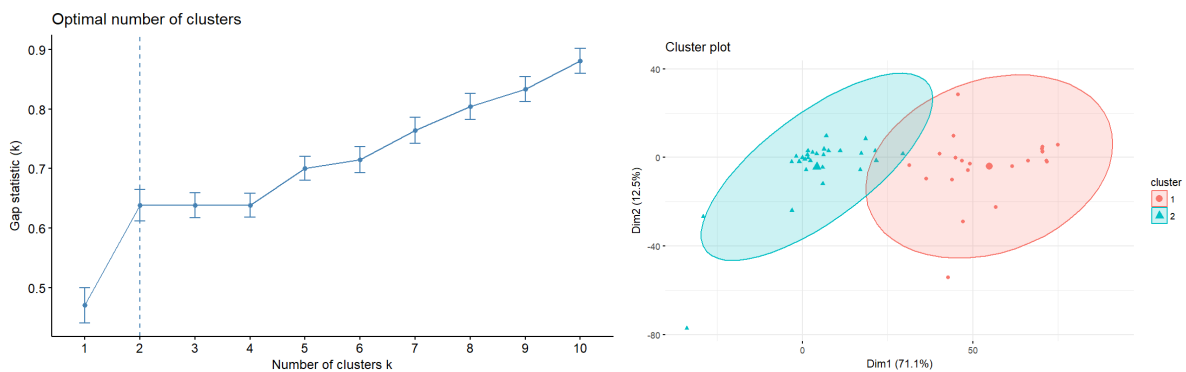


Figure 3: GAP Statistics (left) – Cluster Centers (right)

To see the common reading patterns of the students in these two clusters, we visualized cluster means as well. Results can be seen in Fig.4. From Fig.4, we found two different patterns. Based on these patterns, students in Cluster 2 labelled as On-Task students since they are following the instructor until at the end of the lecture. On the other hand, students in Cluster 1 labelled as Off-Task students since after 17th minute of the lecture those students could not follow the instructor. In addition, distance between Cluster 2 and Cluster 1 increased towards the end of the course.

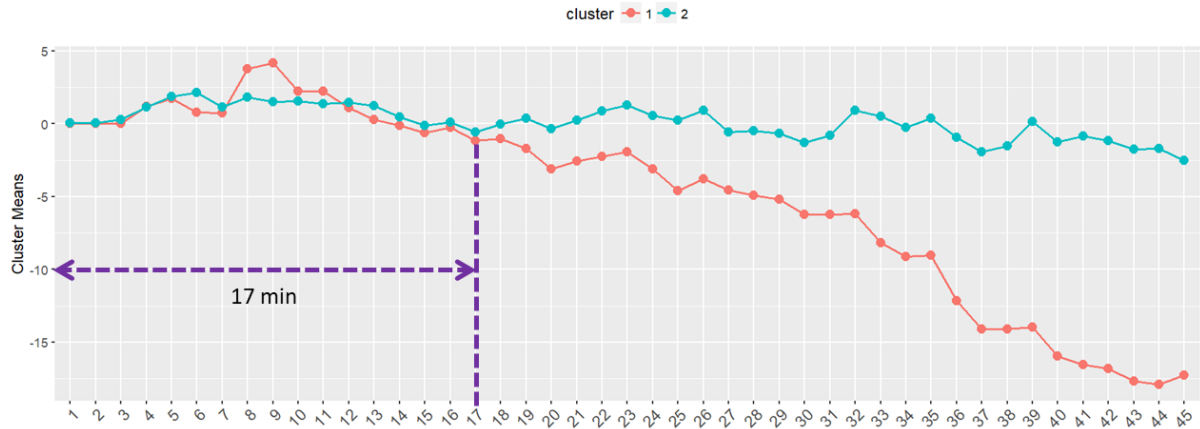


Figure 4: Clustered reading patterns

3.2 Student Academic Performance

As mentioned before students took the quiz in last 15 minutes of lecture. Even it was an open-book quiz, we compared the quiz performance of the students in two different clusters. Since the data were not normally distributed, we used Wilcoxon Signed-ranks test to compare two groups. We compared students' quiz scores and the time they spend on quiz. Results is shown in Table 1. In terms of scores, a Wilcoxon Signed-ranks test indicated no significant difference between Cluster 1 (Mdn = 10) and Cluster 2 (Mdn = 10), $W = 296$, $p = .58$. The time spent on quiz was also not significantly differed among Cluster 1 (Mdn = 146) and Cluster 2 (Mdn = 120), $W = 320$, $p = .31$.

Table 1: Descriptive Statistics

Variable	Cluster 1 (n = 21)		Cluster 2 (n = 26)	
	Mean (Sd)	Median	Mean (Sd)	Median
Score	9.14 (1.20)	10	8.62 (2.10)	10
Time	145 (65)	146	124 (42)	120

4 CONCLUSION

In this study, we identified off-task students by analyzing their reading patterns, however, in terms of academic performance there was no significant difference between off-task and on-task students is noted. In literature, there are many studies found no relationship between off-task behavior and learning outcomes and the reasons for this are not yet known (Coccea et al., 2009; DeFalco, Baker, & D'Mello, 2014). In our case, there might be two possible explanations. First, the quiz sessions students took part in was open book. Therefore, even off-task students might find the answers during the quiz since their average time on quiz higher than students on-task. On the other hand, high-knowledge or high-ability students might also exhibit off-task behaviors since they find the task too easy for them. Simonsen, Little, and Fairbanks (2010) found that less challenging tasks may be less engaging the high-ability students. However, further research is required to test these hypotheses.

Teachers cannot observe and interact with every student at the same time, however, an off-task behavior detector built into the learning environment can observe every student at every moment (Baker, 2007). The obtained results can be used to develop real-time detector for off-task students. Interventions can also be designed to help off-task students (Walonoski & Heffernan, 2006).

ACKNOWLEDGMENT

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

REFERENCES

- Baker, R. S. (2007). *Modeling and understanding students' off-task behavior in intelligent tutoring systems*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA.
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). *Off-task behavior in the cognitive tutor classroom: when students "game the system"*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria.
- Cetintas, S., Si, L., Xin, Y. P. P., & Hord, C. (2010). Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Transactions on Learning Technologies*, 3(3), 228-236. doi:10.1109/TLT.2009.44
- Cocca, M., Hershkovitz, A., & Baker, R. S. J. d. (2009). *The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate?* Paper presented at the Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling.
- DeFalco, J., Baker, R. S. J. d., & D'Mello, S. K. (2014). *Addressing Behavioral Disengagement in Online Learning*.
- Flanagan, B., & Ogata, H. (2017). *Integration of Learning Analytics Research and Production Systems While Protecting Privacy*. Paper presented at the 25th International Conference on Computers in Education, ICCE 2017, New Zealand.
- Hastie, T., Tibshirani, R., & Walther, G. (2001). Estimating the number of data clusters via the Gap statistic. *J Roy Stat Soc B*, 63, 411-423.
- Hofer, M. (2007). Goal conflicts and self-regulation: A new look at pupils' off-task behaviour in the classroom. *Educational Research Review*, 2(1), 28-38. doi:10.1016/j.edurev.2007.02.002
- Hughes, K. M. (2010). *Educational Software for Off-Task Behavior*. Unpublished Major Qualifying Project. Worcester Polytechnic Institute, MA.
- McElroy-Yeider, & Courtney, Q. (2016). *On-task Behavior During a Reading Task: Effect of Quality Physical Education*. BSU Master's Theses and Projects, Retrieved from <http://vc.bridgew.edu/theses/31>
- Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., . . . Hirokawa, S. (2018). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In H. Yasuura, C.-M. Kyung, Y. Liu, & Y.-L. Lin (Eds.), *Smart Sensors at the IoT Frontier* (pp. 327-350). Cham: Springer International Publishing.
- Simonsen, B., Little, C. A., & Fairbanks, S. (2010). Effects of Task Difficulty and Teacher Attention on the Off-Task Behavior of High-Ability Students with Behavior Issues. *Journal for the Education of the Gifted*, 34(2), 245-260.
- Walonoski, J. A., & Heffernan, N. T. (2006). *Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems*. Paper presented at the Intelligent Tutoring Systems, Berlin, Heidelberg.

An Investigation of Academic Performance, Mindless Reading and Its Reading Behavior Indicators

Michelle P. Banawan

Ateneo de Davao University
mpbanawan@addu.edu.ph

Ma. Mercedes T. Rodrigo

Ateneo de Manila University
mrodrigo@ateneo.edu

ABSTRACT: We present initial research results for our work on classifying students according to performance using reading clickstream indicators. We also make an initial attempt to model the phenomenon of mindless reading or zoning out while reading as it has been found to have a negative impact on comprehension and other reading failures. Researchers in the field of psychology and education have investigated this common phenomenon mostly found in reading within and outside academic settings and have devised ways to increase or decrease zoning out which they also referred to as mind wandering or lack of attentiveness. We used univariate, bivariate and multivariate statistical analysis to derive insights from the dataset. We then operationalized zoning out or mindless reading using references from education, psychology and social science literature. We found that zoning-out as a construct may be determined from the clickstream data. Since this is an initial study, we propose that a self-report mechanism be used to validate the ground truth of zoning out. As a minimum contribution we were able to find that the reading click-stream data are indicators of academic performance by implementing a neural network classification model. The neural network classification model for the academic performance (i.e. score) outperformed the neural network classification model for the zoning-out construct. The zoning-out model can be improved if self-report data will be gathered in the subsequent phases of this work.

Keywords: Reading behavior, Academic Performance Indicators, Zoning-out, Neural Network

1 INTRODUCTION

Reading within academic settings has been deemed as an important activity. It has been found to contribute to the overall academic performance of students. When reading failures occur, a negative impact on academic performance is seen. An example of reading failure is non-comprehension of the reading material and this has been attributed to mindless reading. Mindless reading or zoning-out while reading is a common phenomenon within academic settings. Being able to detect this phenomenon will be beneficial as it has been reported to be detrimental to comprehension and related to a number of reading failures (Reichle, Reineberg and Schooler, 2010). It is in this context that we investigate academic performance from reading clickstream data and a related construct referred to as zoning out within the datasets provided by Flanagan, et.al. which contains clickstream reading data of students in three different universities.

1.1 Research Questions

We would like to answer the research questions: What are the indicators of academic performance within the clickstream data of Bookroll as measured by the final total score received by the student at the end of the course? Does zoning-out or mindless reading occur within this clickstream dataset?

2 THE DATASET

We used the datasets provided by Flanagan, et.al. for this data challenge as our testbed to answer this paper's research questions. The first sub-dataset is comprised of BookRoll (Flanagan and Ogata, 2017) clickstream data, lecture time, lecture material and quizscore data of students from three Japanese universities. The clickstream data (Eventstream) is comprised of user actions with timestamps while reading online content. These actions include opening the book, closing the book, navigating to the next and previous pages, which includes also jumping to particular pages, adding bookmarks, markers and memos and editing these, and other reading data. The second sub-dataset we used was the quizscore dataset. This contains data on the final total score that the student has received for the course.

3 PRIOR WORK ON MINDLESS READING AND MIND-WANDERING

The phenomenon of zoning out or mind-wandering while reading has been only very recently investigated as it was previously deemed as subjective and too elusive to scientifically study (Zimmer, 2009). Investigating why readers zone out has been found to be particularly interesting to researchers for its utility and costs (Mooneyham and Schooler, 2013). Zoning out has been found to be a potential reason for compromised comprehension and other reading failures (Franklin, Smallwood and Schooler, 2011). Experiments investigating zoning-out related to reading have used self-report mechanisms to record their zoning-out episodes and detectors of these phenomena have been studied (Drummond and Litman, 2010). Schooler, Reichle and Halpern have found that detecting inconsistencies or errors in the text being read is sensitive to zoning out episodes which explain why reading success can be predicted by monitoring the errors or inconsistencies detected by the reader more than determining reading comprehension (Schooler, 2004). A related construct that has been to zoning out episodes is Task Unrelated Images and Thoughts (TUIT). According to Giambra and Grodsky, 1990, the difficulty of the text was not an indicator of TUIT occurrence or frequency, i.e. both difficult and easy texts may produce TUITs.

Mindfulness and mind wandering have been studied in prior work and found that the indicators of these two constructs reveal opposing relationships. In addition, interventions towards mindfulness reduce mind-wandering (Mrazek, Smallwood and Schooler, 2012).

The authors have investigated carefulness or mindfulness in an educational game for Physics and have empirically validated its in-game predictors using Philippine samples from three different universities in the Philippines (Banawan, Rodrigo and Andres, 2017). In this study carefulness/mindfulness was found to be predicted by mastery, novelty, caution, and control. In another work using student annotated reading data, self-reported zoning out episodes while reading

reveal that too much difficulty and irrelevant texts were indicators of zoning-out episodes (Banawan, 2018).

4 OPERATIONALIZING MIND-WANDERING OR ZONING OUT GROUND TRUTH

Mind-wandering has been described to be happening when the reader's eyes move across the page and only little, or none, of what has been read is processed meaningfully or when the eyes continue to read the words without due attention to their meaning (Reichle, Reineberg and Schooler, 2010) (Smallwood, 2011).

Taking off from prior work on mind-wandering or zoning-out when reading, a possible indicator of this construct could be the student's (absence of) state of flow (Csikszentmihalyi, 1997) and (lack of) engagement to the task at hand, i.e. reading. There has been a number of prior work investigating the different indicators of flow or engagement (Buselle and Bilandzic, 2009), (Fredricks, Blumenfeld, Friedel and Paris, 2005). After inspecting the dataset, we propose that zoning-out can be (reversely) determined by the number of difficult and important markers (also reverse-coded) that the reader annotated. This implies that among all the candidate indicators in the feature set, if the reader placed a marker then the reader is in a state of flow and is not zoning-out. Hence, we used both important and difficult markers (reversely-coded) as the basis for determining zoning-out. In addition, since prior work has also proven that zoning-out is negatively related to academic performance, we also used score as an (reversely-coded) indicator for zoning-out. In summary, we propose that zoning-out while reading is determined by engagement (with conscientious marking of texts as a reverse indicator) and academic performance (with score as a reverse indicator).

5 METHODS

This section discusses the methods that we used to answer our proposed research questions and derive related insights from the dataset.

5.1.1 Descriptive Analysis

The summary of descriptive statistics on the numeric features of the dataset is presented in table 1.

Table 1: Descriptive Statistics of Numeric Features

Feature	Mean	Standard Deviation
Score	82.39	10.344
Add Bookmark	11.99	20.01
Add Marker	80.58	111.67
Difficult Markers	17.39	38.32
Important Markers	92.15	118.56
Add Memo	12.25	20.47
Bookmark_Jump	17.42	29.02
Change_Memo	7.22	22.41

Close	8.55	5.96
Delete_Bookmark	2.52	2.34
Delete_Marker	13.47	18.32
Delete_Memo	1.34	1.02
Link_Click	2.26	2.85
Memo_Jump	6.33	8.39
Next	780.19	485.02
Open	15.27	8.10
Page_Jump	31.83	32.23
Prev	386.17	319.71
Search	5.06	6.81
Search_Jump	2.08	6.43
Grand_Total	1,335.36	867.42

5.1.2 Assumptions

Pearson correlation was performed on the aggregated dataset and we found no significant correlations existing between the candidate predictors.

5.1.3 Preprocessing

We wrote scripts to append the different data subsets and aggregated the values of each feature at the student level. Examples with missing scores were eliminated from the resulting aggregated dataset. We also applied z-transformation as a normalization method. The cleaned and preprocessed aggregated dataset had 1,430 examples and 24 attributes (22 regular attributes, 2 special attributes: id and label). Three attributes were added to the attributes presented in table 1, i.e. User_Id (as ID attribute), Cluster (as label), and total memo length (as regular label).

5.1.4 Clustering the Data According to Zoning-Out Ground Truth

We then clustered the aggregated student datasets according to the score and markers. We used the resulting cluster as the ground truth to label the degree of zoning-out of the students. Using X-means clustering algorithm, three well-separated clusters emerged.

Table 2: Zoning-Out Cluster Centroids

	Cluster 0	Cluster 1	Cluster 2
Score	0.32	0.22	3.45
Important Markers	0.98	-0.13	0.30
Difficult Markers	2.11	-0.29	-0.48

Doing a qualitative inspection and evaluation of the resulting clusters, we can characterize each cluster by reviewing their cluster centroids for the score, important markers and difficult markers. Clusters 0 and 2 as those students who had the least zoning-out episodes as evidenced by their high

scores and high number of markers placed. Cluster 1, however, seem to be comprised of those students who had the most number of zoning-out episodes as evidenced by the least number of markers and the lowest score (normalized) values. The more difficult part is now characterizing as to which students were better (had fewer zoning-out episodes) between Cluster 0 and 2. If we reference prior work, zoning out has been found to have a negative impact on academic performance. It was also previously established that zoning out is similar to disengagement. From their cluster centroids, we find that cluster exhibited more engagement than cluster 2 (as evidenced by the higher mean values for important and difficult markers). Given the prior findings, we establish the following ranking in terms of zoning-out : Cluster 1 (most number of zoning out) → Cluster 2 (have some zoning out / disengaged episodes → Cluster 0 (least number of zoning out/most engaged).

Because the data did not have self-reported values on actual zoning-out episodes, we used these three clusters as labels or ground truth for zoning-out.

5.1.5 Neural Network Classification Model

We built a neural network model using (Breuel and Shafait, 2010) Auto MLP where learning rate adjustments are done during training and the parameters for the best networks during validation are used. The performance of the neural network classification model is shown in table 3. This model shows that zoning-out or mindless reading existed in the data and can be predicted by the action-based features and the score-based feature (i.e. cluster).

Table 3: Performance Metrics of ANN Classification Model

	Values
Accuracy	84.10 %
Classification Error	15.90%
kappa	0.374

5.1.6 Establishing Clusters According to Academic Performance

We also used the normalized (z-transformed) score to establish clusters of academic performance. After running X-means (k-means based) algorithm for the scores of the student aggregated dataset, two well-separated clusters were formed (see table 4).

Table 4: Academic Performance (Score-based) Clusters

	Cluster 0	Cluster 1
Score	-3.51	0.23

Looking at the mean values of both clusters, we characterized Cluster 0 as the least performing cluster and cluster 1 as the better performing cluster of students.

5.1.7 Neural Network Classification Model for Academic Performance

Using (Breuel and Shafait, 2010) Auto MLP to build the ANN classification model for academic performance, we derived the following model performance shown in table 5.

Table 5: Performance Metrics of ANN Classification Model for Academic Performance

	Values
Accuracy	94.52 %
Classification Error	5.48%
kappa	0.029

6 SUMMARY OF FINDINGS AND RECOMMENDATIONS

We found that academic performance (scores) can be determined by the reading clickstream features which implies that a student's reading behavior are indicators of his academic success/performance. We also found that zoning-out episodes can be determined by a student's academic performance and level of engagement. However, the classification model for zoning-out episodes performance was not at par with the model performance for academic success. This could mean that zoning-out can still be attributed to other factors not determined in this study. Further, the features investigated in this study predicted mastery or academic success more than zoning-out.

In comparison to the authors' prior work, mindfulness in a problem-solving environment such as Physics Playground has been found to exist and can be predicted by mastery, novelty, caution and control indicators while in the context of eBook reading, (the reversely coded) mindfulness (or zoning-out) can be robustly predicted using indicators of mastery, difficulty and the reader's perceived importance (of content).

As a recommendation, self-report data can validate and improve this study's zoning-out model. We also recommend that the time-based features be investigated and included in future models of mindfulness as the predictors considered in this study only focused on action-based features.

As a possible contribution, the findings of this study can help learners achieve higher academic success when features like degree of difficulty, mastery and importance of the reading material be incorporated in the cognitive and meta-cognitive scaffolding for teaching and learning environments that use reading as a learning activity.

REFERENCES

- Banawan, M. P., Rodrigo, M. M. T., & Andres, J. M. L. (2017). Predicting Student Carefulness within an Educational Game for Physics using Support Vector Machines. In *Proc. of the 25th International Conference on Computers in Education*(pp. 62-67).
- Banawan, M. P., Andres, J. M. L., & Rodrigo, M. M. T. Predicting Student Carefulness in an Educational Game for Physics Using Semi-supervised Learning.
- Banawan, M. P. (2018). Detectors of Zoning-out Episodes While Reading Among Computer Studies Students. Manuscript n preparation.

- Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, 12(4), 321-347.
- Breuel, T., & Shafait, F. (2010, April). Automlpl: Simple, effective, fully automated learning rate and size adjustment. In *The Learning Workshop* (Vol. 4, p. 51). Utah.
- Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.
- Drummond, J., & Litman, D. (2010, June). In the zone: Towards detecting student zoning out using supervised machine learning. In *International Conference on Intelligent Tutoring Systems* (pp. 306-308). Springer, Berlin, Heidelberg.
- Flanagan, B., Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy, *Proceedings of the 25th International Conference on Computers in Education (ICCE2017)* (pp. 333-338).
- Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*, 18(5), 992-997.
- Fredricks, J. A., Blumenfeld, P., Friedel, J., & Paris, A. (2005). School engagement. In *What do children need to flourish?* (pp. 305-321). Springer, Boston, MA.
- Grodsky, A., & Giambra, L. M. (1990). The consistency across vigilance and reading tasks of individual differences in the occurrence of task-unrelated and task-related images and thoughts. *Imagination, Cognition and Personality*, 10(1), 39-52.
- Ogata, H., Yin, C., Oi, C., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education, *Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015)* (pp.401-406).
- Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., & Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In *Smart Sensors at the IoT Frontier* (pp. 327-350). Springer, Cham.
- Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2012). Mindfulness and mind-wandering: finding convergence through opposing constructs. *Emotion*, 12(3), 442.
- Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: a review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(1), 11.
- Smallwood, J. (2011). Mind-wandering while reading: Attentional decoupling, mindless reading and the cascade model of inattention. *Language and Linguistics Compass*, 5(2), 63-77.
- Schooler, J. W. (2004). Zoning Out while Reading: Evidence for Dissociations between Experience and Metacognition Jonathan W. Schooler, Erik D. Reichle, and David V. Halpern. *Thinking and seeing: Visual metacognition in adults and children*, 203.
- Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, 21(9), 1300-1310.
- Zimmer, C. (2009, June). *The Brain: Stop Paying Attention: Zoning Out Is a Crucial Mental State*. Retrieved from <http://discovermagazine.com/2009/jul-aug/15-brain-stop-paying-attention-zoning-out-crucial-mental-state>

Feature Engineering for Learning Log Analysis

Sachio Hirokawa

Kyushu University

hirokawa@cc.kyushu-u.ac.jp

Chengjiu Yin

Kobe University

yin@lion.kobe-u.ac.jp

ABSTRACT: Estimating students' final grades from learning behavior is one of the important issues in learning analytics. Two tasks are to build a model with high estimation performance and to clarify important features which affect the final results. In previous studies it is known that the students' behavioral characteristics are more important factors than demographic features such as sex and age. In this paper, not only the browsed pages and browsing operations but also the paths of the browsing pages are added to the vector data of each student's learning behavior as high dimensional data. We applied SVM (support vector machine) and feature selection to gain high prediction performance with interpretability. Specifically, we analyzed a total of 2,014,652 learning logs from 1,1545 students at three universities and predicted if a student obtained the score of 80. We identified with high performance of F-measure 92%. This result was obtained with the feature selection from all combination of pages, operations and the page transition paths of the length less than 6 pages. This performance is better than that by pages (77%) and than that by operations (84%).

Keywords: Learning Log, E-book Book-Roll, SVM, Feature Selection, Prediction Performance

1 INTRODUCTION

The Japanese government schedules to use e-books for elementary, middle and high schools by 2020 (Yin et al., 2014; Yin et al., 2015; Yin et al., 2016; Yin et al., 2018). Much attention is paid to make good use of data kept as student learning logs. In fact, the research on learning analytics (LA) and on educational data mining (EDM) attracts many researchers, for example in conferences of Learning Analytics (<https://solaresearch.org/events/lak/>) and Educational Data Mining (<http://educationaldatamining.org/>). Another reason of this current status is the progress of learning analytics platform. Actually, the data of the present paper analyses is provided by the BookRoll system (Flanagan & Ogata 2017, Ogata et al. 2015, Ogata et al. 2017).

LA results can be used to optimize and increase educational benefits for education (Colvin et al., 2015; Yin & Hwang, 2018). Many of researches used LA to do prediction. Some of them want to know who might fail a class; some of them want to know whether s/he mastered the skill to solve the next problem. There are many prediction analysis methods such as Classification, Regression, and Latent Knowledge Estimation (Yin & Hwang, 2018).

The present author proposed a visualization of page view transition in (Hirokawa et al. 2015). However, the quantitative evaluation of the visualization was not given then. The present paper applies the

machine learning method SVM and feature selection to predict the high performance students who achieved more than 80 points in their evaluation.

One record of the learning log analyzed in this paper consists of items as shown in Table 1. Since we focus on the operations of browsed pages, the three attributes are not used in this research: marker (highlight or underline) of the 4th item in table1, memo length (the 5th item in table1), device type (pc or mobile, and the 6th item in table1) . In addition, the seventh eventtime (7th item) was used for the order of browsing pages, but it was not used as a single attribute. The total number of userids (0th item) was 11, 545. There were 42 kinds of teaching materials for contentsid as a whole. In the supplied data, contentsid was a long character string, but in this paper it is represented by numbers 1 to 42. The page number ranges from 1 page up to 147 pages.

	no	attribute	sample
*	0	userid	166
*	1	contentsid	5ef059938ba799aaa845e1c2e8a762bd
*	2	operationname	NEXT
*	3	pageno	1
	4	marker	NULL
	5	memo_length	0
	6	devicecode	pc
*	7	eventtime	2018/04/11 0:20:49

Table 1: Attributes of Log

The operations (2nd item) have 15 types of operations as below. In this paper, we first constructed search engines for 2,014,652 access

OPEN : opened the book
 CLOSE : closed the book
 NEXT : went to the next page
 PREV : went to the previous page
 PAGE_JUMP : jumped to a particular page
 ADD BOOKMARK : added a bookmark to current page
 ADD MARKER : added a marker to current page
 ADD MEMO : added a memo to current page
 CHANGE MEMO : edited an existing memo
 DELETE BOOKMARK : deleted a bookmark on current page
 DELETE MARKER : deleted a marker on current page
 DELETE_MEMO : deleted a memo on current page
 LINK_CLICK : clicked a link contained in the e-book current page
 SEARCH : searched for something within the e-book
 SEARCH_JUMP : jumped to a page from the search results

2 DESIGN OF FEATURES AND SEARCH ENGINES OF STUDENTS' LEARNING BEHAVIOR

We analyzed how 1,545 students used Book-Roll. Log information of each student was expressed as a word as shown in Table 2. We vectorized each student as BOW (bag of words) containing those words.

Attributes are roughly divided into three types -- page information, operation information, and path information. Page information is represented as a pattern of the form "contentsid:page". An example of the pattern "33:2" in the second row of Table 2 shows that the student viewed the second page and the third page of the teaching materials no.33. Operation information was set to start with "o" or "O". The one that starts with "o" is followed by the teaching material number, page number, and operation. For items starting with "O", the teaching material number and the operation continue. This allows us to analyze which teaching materials are affecting students' grades and which pages of which teaching materials are affecting students' grades. A path is a sequence of pages consecutively viewed. We used two kinds of patterns, with teaching material number and without teaching material number. The length was set to 5 or less.

Using these attributes, we built four kinds of search engines shown in Table 3. We applied machine learning and feature selection method (Sakai & Hirokawa2012) with a positive example of students with a final grade of 80 or more.

Table 2: Extended Log Features

attributes	samples
contentsid:page	33:2, 33:3
contentsid:page:operation	o:32.19.search_jump, o:33.1.next
contentsid:operation	O:32.search_jump, O:33.next
path(length 2)	46-45
path(length 3)	46-45-46
path(length 4)	46-45-46-47
path(length 5)	46-45-46-47-46
contentsid:path(length 2)	30:10-9
contentsid:path(length 3)	30:46-45-46
contentsid:path(length 4)	30:46-45-46-47
contentsid:path(length 5)	30:46-45-46-47-46

Table 3: Search Engines for Student Learning Behavior

Search Engine	attributess
page	uses contentsid&page number
ope	uses contentsid, page number and operation
ope+path2	uses contensid, page number, opration and path of length two
ope+path5	uses contensid, page number, opration and path of length less than six

3 PREDICTION OF HIGH SCORE STUDENTS WITH OPTIMAL FEATURE SELECTION

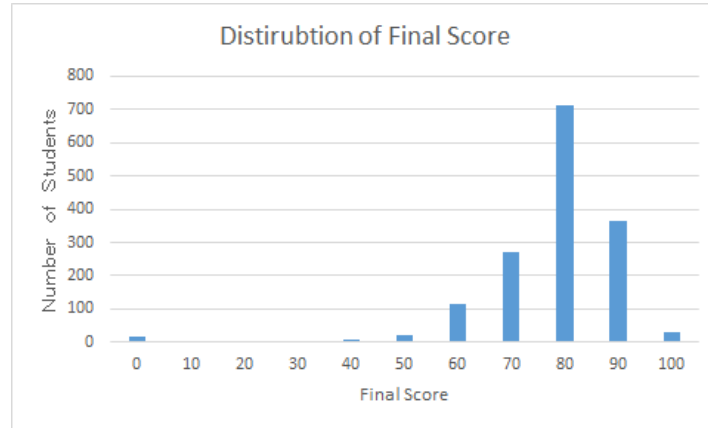


Figure 1: Distribution of Final Score

Figure 1 shows the distribution of students by the final grades. For example, at score 60, it indicates that there are 114 students with a score of over 60 but less than 70. As we can see from Figure 1, it follows normal distribution. In this paper, we consider students with the score over 80 as top students. We applied machine learning and took these students as positive data.

Table 4 shows the number of attributes and the discrimination performance in vectorization by the four attribute sets described in the previous chapter. It is understood that the dimension is 10 times (ope + path 2) and 50 times (ope + path 5) higher than the vectorization on the page by introducing the path. However, in Table 4 using all the attributes, there is no big difference in the discrimination performance, and all are about 70%.

Table 4: Prediction Performance (baseline with all feature)

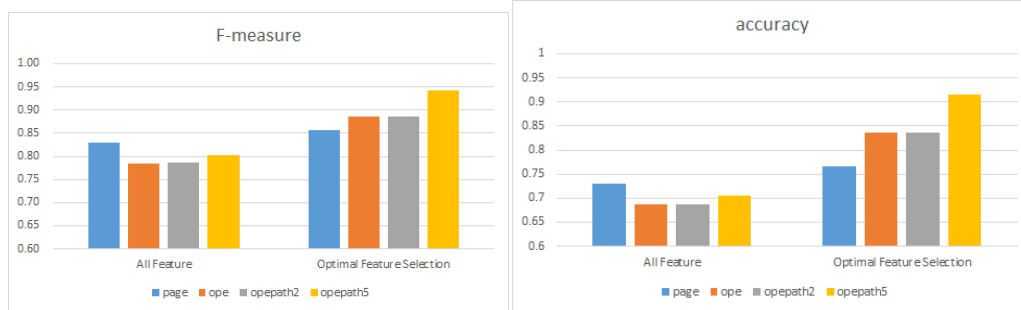
vectorization	dimension	precision	recall	F-measure	accuracy
page	1872	0.7572	0.9185	0.8300	0.7300
ope	8548	0.7775	0.7907	0.7838	0.6873
ope+path2	18377	0.7713	0.8006	0.7854	0.6865
ope+path5	101437	0.7733	0.8323	0.8012	0.7042

On the other hand, Table 5 shows the discrimination performance in vectorization by feature selection that optimizes F-measure. The second column shows the optimum number of attributes. For example, in the ope of the third line, the F-measure is 0.8861 when $N = 400$, which is 10% better than when using all attributes. It should be noted that $N = 400$ uses 800 attributes for the top 400 attributes of the positive score and the top 400 attributes of the negative score. Figures 2 (a) and (b) show F-measure and Accuracy by optimum attribute selection. It can be confirmed that the discrimination performance is improved by attribute selection. Furthermore, it is understood that the performance is the highest in the path of length 5.

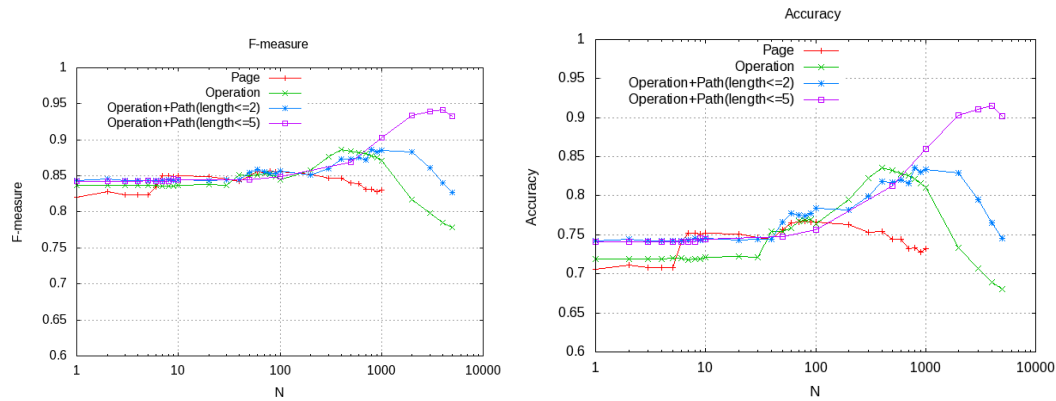
Table 5: Prediction Performance (Optimal Feature Selection wrt F-measure)

vectorization	opt-FS	precision	recall	F-measure	accuracy
page	80	0.7686	0.9667	0.8562	0.7671
ope	400	0.8777	0.8956	0.8861	0.8355

ope+path2	800	0.878	0.895	0.8862	0.8354
ope+path5	4000	0.9201	0.9656	0.9421	0.9154



(a) F-measure (baseline & Optimal) (b) Accuracy (baseline & Optimal)



Effect of Feature Selection for (c) F-measure and (d) Accuracy

Figure 2: Prediction Performance of High Score Students

Figures 2 (c) and (d) show the change in F-measure and accuracy when N is changed. There is no effect of attribute selection with page information alone. There is no big difference between ope and oope + path 2. In ope + path 5 which also includes a path of length 5, the performance has continued to improve even with over 1000, which will degrade the performance with ope and oope + path 2.

4 FEATURES OF HIGH SCORE STUDENTS

Figures 3 (a), (b), (c) and (d) are the top ten highest positive attributes in the vectorization by page, operation, ope+path2, and ope+path5. It indicates a negative attribute. For example, you can see that the positive top-most attribute of page has a frequency of 1251 at "30:31", and the negative top-most attribute has a frequency of 1208 at "31:10". "30:31" means the 31-st page of content with ID 30. Note that the content ID is for identification purposes only and there is no point in the order. Even for the same content number 30, the pages 1, 31 and 57 are positive features, whereas the page 51 and 53 are negative features. Even "31:10" the highest negative feature, the page "31:12" is positive

feature. In this way, even with the same teaching materials, we can see the difference in the effects to high score students.

Prior research [Hirokawa 2018] reported the transition to the previous page as a feature of the top students. However, in Figure 3 (b), the operation "prev", which represents the transition to the previous page, appears in both positive and negative attributes. However, "next" does not appear at the higher level. Therefore, it can be said that the difference between the high score student and the low score students locates in where they look back.

The most characteristic of Figure 3 (c) and (d), where with page transitions are added as features, is that that the operation of "page_jump" appears as the most important feature. The operation does not appear in (a), (b). In (c) of ope+path 2, "page_jump" from page 66 of material 29 is also in the higher rank. However, in (d), only "page_jump" with no contents and no page is included in the top ten. The present paper is a preliminary analysis of features. Further analysis and interpretation will be reported in the final version.

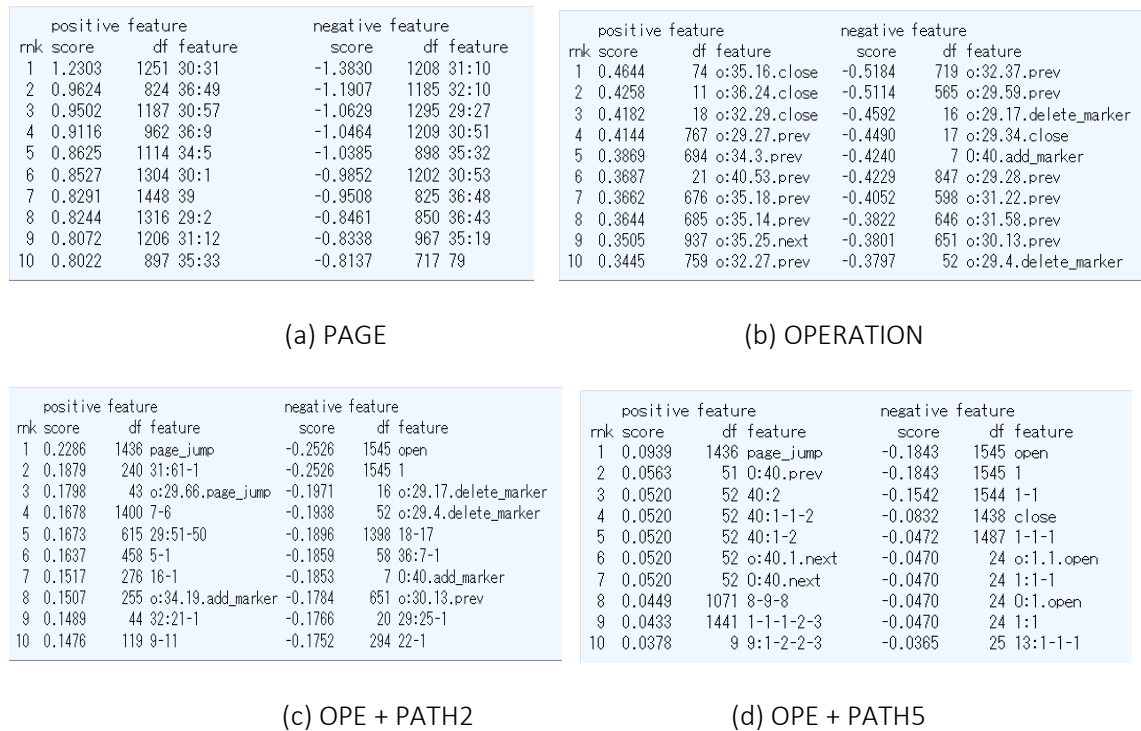


Figure 3: Characteristic Feature

5 CONCLUSION AND FURTHER WORK

In this paper, we analyzed the learning logs on the e-book reading system Book-Roll to estimate the students' final scores. In this paper, three attributes of "learning material", "pages", "operation", such as , prev, next, jump, and "page transition" are used to vectorize the student's learning behavior. Those features are used as words and a student is described by a BOW (bog of words). We applied machine learning method SVM and feature selection (Sakai & Hirokawa2012) to predict high score students who gained 80 points or over as the final score.

The optimum feature selection with respect to the "page2 information alone, the accuracy is only 0.7671. But when we considered "operation" as features, the accuracy reached 0.8355. Moreover, it turned out that the attribute with the highest score was "page_jump". When considering the "page transition" up to length 5, the operation "page_jump" remained as the most important feature and the prediction performance is improved to 0.9154.

The prediction model constructed in this paper has both high discrimination performances and attributes interpretability, which is a satisfactory result. For example, looking at the attribute obtained as a feature in detail, you can see which page of which content has higher grades and different students. However, those interpretations as learning theories are future tasks. In the model with paths up to length 5, the optimum number of attributes N is 4000, and it is necessary to confirm whether it is a model common to all subjects or a union of many different learning materials in different universities. This can be confirmed by constructing a model through limiting the target to one teaching material and verifying it with students using other teaching materials. In other words, we think that it can be confirmed by cross validation with the teaching material as a unit.

REFERENCES

- Colvin, C., Rogers, T., Wade, A., Dawson, S., Gašević, D., Buckingham Shum, S., Fisher, J. (2015). Student retention and learning analytics: A snapshot of Australian practices and a framework for advancement (Research Report). Canberra, Australia: Office of Learning and Teaching, Australian Government.
- Flanagan, B. and Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy, Proceedings of the 25th International Conference on Computers in Education (ICCE2017), pp.333-338
- Hirokawa, S., Yin, C. , Wang, J., Oi, M., and Ogata, H. (2015). Visualization of e-Book Learning Logs, Proc. International Conference of Computers on Education, pp.659-664
- Ogata, H. , Yin, C. , Oi, M. , Okubo, F., Shimada, A. , Kojima, K., and Yamada, M. (2015), E-Book-based learning analytics in university education, Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015) pp.401-406
- Ogata, H., Oi, M. , Mohri, K. , Okubo, F. , Shimada, A. , Yamada, M. , Wang, J. , and Hirokawa, S. (2017), Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In Smart Sensors at the IoT Frontier, pp.327-350, Springer, Cham
- Sakai, T. , and Hirokawa, S. (2012), Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, pp.360-367
- Hirokawa, S. (2018), Good Students Look Back Previous Pages, ICCE2018 Workshop Proceedings, pp.457-466
- Yin, C., Okubo, F., Shimada, A., Kojima, K., Yamada, M., Ogata, H., & Fujimura, N. (2014). Smart phone based data collecting system for analyzing learning behaviors. Proceedings of the International Conference of Computers on Education, Nov.30-Dec.4, Nara, Japan, 575- 577.

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., Yamada, M., Kojima, K., & Ogata, H. (2015), Identifying and Analyzing the Learning Behaviors of Students Using e-books, Proc. of 23rd International Conference on Computers in Education, 118-120.

Yin, C., Yau, J.Y.K., Uosaki, N., Hirokawa, S., Kumamoto, E. (2016), Measuring & evaluating digital textbooks through quizzes, Proc. of 24th International Conference on Computers in Education, 374-379.

Yin, C., & Hwang, G. J. (2018). Roles and strategies of learning analytics in the e-publication era. Knowledge Management & E-Learning, 10(4), 455–468.

Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K. & Ogata, H. (2018): Exploring the Relationships between Reading Behavior Patterns and Learning Outcomes Based on Log Data from E-Books: A Human Factor Approach, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2018.1543077

Beyond Identifying Areas for Improvement in Schools: Using the NILS™ Online Platform to Accelerate Improvement Work

Jojo Manai

The Carnegie Foundation for the Advancement of Teaching
manai@carnegiefoundation.org

Hiro Yamada

The Carnegie Foundation for the Advancement of Teaching
yamada@carnegiefoundation.org

Susan Haynes

The Carnegie Foundation for the Advancement of Teaching
haynes@carnegiefoundation.org

ABSTRACT: We confront a growing chasm between rising aspirations for our educational systems and what schools can routinely accomplish. Although educators at the classroom, school, and district levels are expending significant energy generating and testing promising interventions, we often observe a failure to scale up research-based knowledge across varied contexts. This interactive half-day workshop presents a way to move from trying to get better to getting *good* at getting better. We will introduce an improvement science approach that focuses on learning-by-doing to make progress toward a specific aim on a shared problem of practice by leveraging the power of networked communities. We will present how to apply the six core principles of improvement and organize improvement work through an online technology called NILS™ (Networked Improvement Learning and Support platform), emphasizing that (a) knowledge about the innovation itself and associated know-how around effective implementation flow through the interpersonal relationships between different actors; (b) attending to variation in performance and seeing the system that produces the current outcomes help us to identify areas for improvement. Utilizing NILS, participants will engage in structured activities and data exercises, learn how to identify areas for improvement from data, and create a driver diagram as a theory of practice improvement.

Keywords: Networked Improvement Community, Improvement Science, Social Learning, See the System, Systems Thinking, Scaling Up, Variation in Performance, Knowledge Dissemination.

1 BACKGROUND

We currently face a growing rift between rising expectations of what we want schools to achieve and what they can realistically accomplish. For instance, one of the main challenges in education is the failure to scale up research-based knowledge across varied contexts (Lewis, 2015). Bryk (2015) argues that we need an improvement paradigm that recognizes the complexity of educational work and the variability in educational outcomes that the current systems generate. Following this posit,

1

over the past decade the Carnegie Foundation for the Advancement of Teaching has pioneered a fundamentally new vision for the research and development enterprise in education, seeking to join the discipline of improvement science with the powerful capacities of networks to foster innovation and social learning for education reform (Bryk, Gomez, Grunow, & LeMahieu, 2015).

Improvement work is organized around six core principles (Bryk et al., 2015): (a) make the work problem-specific and user centered (*what specifically is the problem we are trying to solve?*); (b) focus on variation in performance (*what works, for whom, and under what set of conditions*); (c) see the system that produces the current outcomes (*how local conditions shape work processes*); (d) embrace practical measurement (*we cannot improve at scale what we cannot measure*); (e) anchor practice improvement in disciplined inquiry (*engage in rapid cycles of PDSA [Plan, Do, Study, Act]*); (f) accelerate improvements through networked communities (*connect members of professional communities to harness the wisdom of crowds*). Carnegie's approach to improvement is embodied in Networked Improvement Communities (NICs; Bryk et al., 2015). A NIC comprises a group of practitioners, administrators, researchers, and improvement specialists that works to improve a specific problem, shares a working theory of improvement embedded in systems thinking, and uses common measures and inquiry tools for learning whether the changes introduced are moving in the right direction towards improvement.

NILS™ is the **Networked Improvement Learning and Support** online system developed by the Carnegie Foundation to accelerate the initiation and development of work in NICs. The impetus for building NILS emerged from listening to the needs and challenges addressed in various improvement communities including teachers, district leads, and state heads of education, where technology could be of great help in surmounting obstacles or catalyzing the improvement work. This platform is designed to align with the six core principles of improvement and follows the *SECI* model of promoting social, organizational learning and disseminating tacit and explicit knowledge (Nonaka & Takeuchi, 1995) for improvement in education by moving much of what we currently do face-to-face into a virtual learning environment. NICs are communities of practice and learning. Accordingly, NILS enables NIC members to learn improvement methods and culture within a system without the need for high-lift, in-person training. NICs initiate their work through seeing the system in the *Chartering* phase with in-site scaffolding for chartering activities. NICs then progress to system work in the *Improvement Testing* phase with a driver diagram, through which members test and record results for change ideas by running PDSA cycles. Ideas and individual learnings from PDSAs are then spread to the community for social learning via site- or role-based work groups and topic teams. Social learning occurs through school-to-school, school-to-network, and network-to-network conversations among NIC members, which in turn enhance collaboration across the NIC both horizontally and vertically. As a change idea is tested across a variety of contexts, improvement ramps form and individual learnings converge as system knowledge. Members of a network hub curate knowledge gleaned from testing under varied contexts and share findings with the rest of the network, which prompts ideation for further changes. At its core, NILS attempts to address the question of how to derive knowledge from a NIC's data collection cycles: specifically, how does a system surface knowledge and wisdom to the right person at the right time? The platform aims to provide relevant

data to testers based on their contexts and site interactions, and enhance connected learning for professional communities, thus enabling participating educators to take evidence-driven next steps towards achieving a collective aim.

At the LAK17 conference we introduced the initial version of NELS (Author, 2017) and received much interest from participants. We are now ready to show the enhanced version of NELS to participants at LAK19. Through this proposed workshop, participants will learn to utilize NELS for their own problems in a deliberate and systematic manner by acting as members of an actual NIC. We will embed data exercises with a focus on variation in performance throughout the workshop, so that participants will learn how to identify areas for improvement and share their data observations through NELS. In sum, our aim is to promote a more disciplined approach to improvement in schools by leveraging technology that supports continuous improvement.

2 ORGANIZATIONAL DETAILS

We propose a half-day, open workshop for applied researchers, evaluators, practitioners, and school leaders. Expected workshop activities are data exercises, discussion, and practice using NELS. We expect up to 40 participants, and plan to recruit attendees via email and Twitter with the message, “Unleash the power of a Networked Improvement Community to coordinate your research efforts in a practical manner”. Required materials include the Internet, a laptop, and a browser (Chrome, Firefox, Safari, or Edge). The proposed agenda is presented below in Table 1.

Table 1: Proposed agenda

Session	Time	Content
1. Introduction	10 min	● Introduction to Improvement Science and NIC life cycles
2. Launch the Simulation	30 min	<ul style="list-style-type: none"> ● Introduction to the NELS platform ● Get everyone logged in ● Example: chronic absenteeism crisis ● Data conversation protocol
3. Understanding the Problem	25 min	<ul style="list-style-type: none"> ● Why are we getting our current outcomes? ● Hypothesize causes to form theory of improvement ● Solicit feedback from other groups via NELS
Break	20 min	● Sip tea/coffee (with snack)
4. Focusing Collective Efforts	25 min	<ul style="list-style-type: none"> ● Share and comment via NELS ● Craft an aim statement for chronic absenteeism crisis
5. Change Ideas and Testing	35 min	● Example: chronic absenteeism crisis

		<ul style="list-style-type: none"> ● Craft a driver diagram for chronic absenteeism crisis ● PDSA cycle: Family Meeting Protocol ● Share and comment via NILS
6. Evidence in Improvement Science	25 min	<ul style="list-style-type: none"> ● Building evidence for change ● Assess confidence in change bundles
7. Unpacking NILS	10 min	<ul style="list-style-type: none"> ● Current features and future roadmap
8. Closing remarks	10 min	<ul style="list-style-type: none"> ● Summarize key takeaways ● Q & A

3 OBJECTIVE

Through this workshop, participants will learn ways to identify areas for improvement and formalize a theory of improvement. We will focus primarily on two of the six improvement principles: (a) focus on variation in performance (including an identification of positive deviance to learn from) and (b) see the system that produces the current outcomes. Under each principle, we will introduce tools, examples, and data exercises to help participants illuminate variation and see the system. By the end of the workshop, participants will understand how to identify improvement priorities from data and create a driver diagram as a representation of their theory of improvement. We will introduce NILS as an online tool designed to support this work process. We will provide an overview of these learning outcomes through series of tweets using the #improve hashtag. In addition, we will send an email reminder with a detailed agenda and logistics for the workshop.

REFERENCES

- Author. (2017, March).
- Bryk, A. S. (2015). Accelerating How We Learn to Improve. *Educational Researcher*, 44(9), 467-477. <https://doi.org/10.3102/0013189X15621543>
- Bryk A. S., Gomez L. M., Grunow A., LeMahieu P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.
- Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, 44(1), 54-61. <https://doi.org/10.3102/0013189X15570388>
- Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. New York, NY: Oxford University Press.

Interdisciplinary Learning Analytics: What to Know, Who to Talk To, and How It's Done

ABSTRACT: This half-day workshop develops interdisciplinary collaboration among new scholars. In particular, this workshop exists to address the substantial interests expressed by graduate students at past LAKs and members of the SoLAR Student SIG. This workshop addresses three components of interdisciplinary collaboration: what you know, what your collaborators know, and collective interactions. Specific topics include an overview of research areas in learning analytics, exposure to domain specific social-science and computer-science methods and mindsets, and activities building self-reflective and joint collaboration capacity. Workshop time will be split between presentations and structured collaboration. With the long-term goal to create new collaborative research in addition to researcher capacity. We discuss how these community building efforts will be sustained beyond this workshop through the SoLAR Student SIG. We anticipate these topics are widely applicable, but have designed this workshop with an emphasis on new scholars and graduate students.

Keywords: interdisciplinarity, collaboration, graduate students, professional development, communities of practice, learning analytics

1 WORKSHOP BACKGROUND

The learning analytics community began as a self-identified interdisciplinary field (Siemens & Baker, 2012), with interdisciplinarity pervading its areas of expertise, research questions, and designed solutions (Romero & Ventura, 2013). To date, concerns about supporting interdisciplinarity and collaboration have culminated in proposed standards, ethical guidelines, and frameworks (e.g. Berland, Baker, & Blikstein, 2014; Piety, Hickey, & Bishop, 2014). Beyond these solutions, we also see a need to directly foster interdisciplinarity skills and mindset. To that end, this workshop centers around developing interdisciplinary capacity in learning analytics with the goals of discussing disciplinary ideas across domains and concretely promoting future collaborations.

Interdisciplinarity can be simplified into three components: what you know, what your collaborators know, and your shared interactions (author, 2018; author, in press; Klein, 2010). Further, in research collaborations, what you or your collaborators know is further divided into disciplinary methods and epistemologies. Often pitfalls in collaboration relate to one of these areas and result in a lack of shared understanding, not seeing what your skills offer, confusion about what those outside your discipline do, or simply not knowing how to get started. In fact, simply working in an interdisciplinary field does not guarantee you do interdisciplinary work—interdisciplinarity is an intentionally honed skill (author, 2018). To directly and actively promote interdisciplinarity in learning analytics, we address these three areas in the context of building collaboration among new scholars.

Addressing the first component of interdisciplinarity—what an individual does and does not know—we outline key areas in learning analytics (e.g. Lang, Siemens, Wise, & Gašević, 2017). Many pre-career scholars work with advisors with little to no expertise in learning analytics. Often, a student's initial interest in the learning analytics body of research is stymied by not knowing where to start. Our goal is not to teach everything, but outline the field's diversity and provide resources to learn more. In the second component—what other collaborators know—the workshop addresses how techniques and theory differ between computational and learning sciences. Creating these camps, though perhaps artificial, allows consideration of the methods and epistemological differences between fields. Identifying these differences creates a common ground to approach working with collaborators beyond an individual's discipline. Finally, addressing the third component—the interactions at the core of interdisciplinary collaboration—the workshop's activities foster connection with other scholars. By

strengthening a network of young scholars, we believe the future of learning analytics research will grow. As described below, activities center around identifying pathways to collaboration and beginning concrete work within the workshop.

In sum, learning analytics is an increasingly interdisciplinary field, and steps must be taken to ensure graduate students with limited university or advisor support receive the training necessary to thrive as learning analytics scholars (see Dawson, Gašević, Siemens, & Joksimovic, 2014). The material for this workshop builds on peer information shared from the community-of-practice model (Wenger, 2011). We (the organizers) took initiative starting the SoLAR Graduate Student SIG because the interests and struggles expressed by others resonated with our own experiences. Through joining the LAK community, we (the organizers) curated information useful in our own processes of joining the field. To validate and supplement our ideas, we have also conducted informal interviews with experienced learning analytics scholars. Additionally, the organizer's own research in interdisciplinary collaboration provides a guiding theoretical framework.

In conclusion, this workshop extends work initiated at LAK17 and LAK18, to induct new scholars into the field of learning analytics and build interdisciplinary collaborations. At LAK17, the organizers met with over 20 graduate student and postdoctoral scholars to discuss the need for training. In response, at LAK 18, the newly formed SoLAR Student SIG offered a collaborative mentoring workshop. Now, for LAK19, we build on our LAK18 workshop to both address the need of introductory training and promote further development through interdisciplinarity. Our workshop does not replace the existing SoLAR doctoral consortium, which offers an excellent platform for graduate students to receive feedback on their specific work. Instead, this workshop supports collaborative work. It addresses the questions new scholars may have about how to get started, where to learn germane topics outside their departments' training, and how to initiate collaborations.

2 ORGANISATIONAL DETAILS

2.1 Type of Event

This proposed event is a half-day workshop. The coordinators will balance interactive activities with information presentation. Workshop attendees will also collectively participate in several un-workshop style collaboration-forming activities. For such activities, we invite participants to bring an extended abstract of their current or proposed research.

2.2 Schedule and Activities

2.2.1 Welcome and Survey—30 Minutes

To begin, we want participants to introduce themselves since a goal of this workshop is to build interdisciplinary collaboration. The participants will have completed a reflective survey prior to the workshop, including general questions they bring to the workshop, research interests, areas of expertise, training needs, and future career plans. The survey will also include an interdisciplinary self-efficacy measure (Author, 2018). Results from this survey and instrument will be discussed and set the stage for the workshop.

2.2.2 What Comprises the Multidisciplinary Landscape of Learning Analytics—50 Minutes

Next, we will present a brief talk summarizing the broad topics subsumed within learning analytics. For this talk, we draw on resources like *The Handbook of Learning Analytics* (Lang, Siemens, Wise, & Gašević, 2017) and

InfoHub to identify themes in the field and create a “short list” of preeminent articles relevant to each area for dissemination. The remaining time will involve discussion around what areas of work are developing in learning analytics. Using our interdisciplinary framework, the group will consider the deep differences in these research areas, with an underlying emphasis that in learning analytics, no one scholar does it all. We leave participants with the understanding that interdisciplinary collaboration is essential because of the field’s diversity.

2.2.3 *Learning Science Versus Computer Science—50 Minutes*

From the initial broad presentation of areas within the field, we transition to focus on particular orientations between disciplines. In particular, we have found that many graduate programs training learning analytics scholars draw on either social science or computational backgrounds. Thus, we anticipate splitting participants into two groups based on their discipline area. Then, we have two different brief talks and collaborative activities planned to expose participants to the “other side’s” methodology and epistemology. We start with the methodological orientation and perspectives, discuss particular analytical skills, and finally conclude with some advice on how to collaborate across disciplines based on interdisciplinary work research (Klein, 2010; Mansilla & Duraising, 2007).

2.2.4 *Building Interdisciplinary Collaborations—60+ Minutes*

The remaining workshop time involves interactions to develop concrete collaborations. Often, starting a research project is daunting and exciting as a graduate student. Furthermore, when we, the presenters, initially joined the field, we felt unconnected and without collaboration opportunities. To concretely promote interdisciplinary work, we address two issues. First, what is the practical research process behind the articles that we read? Drawing on the struggles in our own dissertation work and especially on interviews with established authors, we will present collaborative workflows from several learning analytics projects. Particular attention will be placed on strategies from interdisciplinarity literature, including adopting roles that integrate different expertises. Our second goal is to begin real collaborations. Participants will be invited to submit an extended abstract of their current or planned work, and will then be grouped based on their interests and heterogeneity of skills. Participants will be presented with research design challenges to brainstorm in their group and then given the opportunity to share out. The objective is to have individuals identify other participants with shared interests to build network connections.

2.3 Recruitment and Dissemination

This event will be promoted through the SoLAR Graduate Student SIG email list and the learning analytics Slack channel. We believe this workshop will hold special interest to graduate students and new scholars. Our intended participant size is 10 to 30 attendees. To further motivate students to attend our workshop, the SoLAR Graduate Student SIG will offer five micro-scholarships to a few workshop attendees. These micro-scholarships (\$150 USD) will effectively cover the cost of this workshop. We especially encourage first-time LAK attendees that do not have access or funding for learning analytics support through their university.

2.4 Equipment

No special equipment will be needed beyond audio / visual presentation equipment.

3 INTENDED OUTCOMES

An intangible outcome of this workshop will be connecting new scholars interested in learning analytics to others, promoting interdisciplinary self-efficacy. Concrete measures of this goal include recruiting new members to the SoLAR Graduate Student SIG, recruiting attendees to future LASI events, and forming research connections leading

to presentations at future LAK conferences or JLA publications. We expect a number of individual outcomes for participants. Though the workshop will not teach a specific analytic tool or method, we hope participants leave understanding the scope of learning analytics research, the various types of skills utilized in this interdisciplinary work, and how to find information and mentorship. Finally, at a more affective level, we hope that the individual outcomes for participants include increased confidence in participating in learning analytics research, connections to other scholars with similar research interests, and a sense of belonging in the SoLAR community. The coordinators for this workshop plan to provide continued support for collaborations formed through resources and connections to mentors. Additionally, we will publish the resources through the SoLAR Student SIG network. This report will be reflective, including survey responses from participants (and other Graduate Student SIG respondents) regarding level of interdisciplinarity, research interests, areas of expertise, training needs, and future career plans.

REFERENCES

- Author. (In Press).
- Author. (2018).
- Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2), 205-220.
- Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014, March). Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231-240). ACM.
- Klein, J. T. (2010). A taxonomy of interdisciplinarity. In R. Frodeman, J. T. Klein, & C. Mitcham (Eds.), *Oxford handbook of interdisciplinarity* (pp. 15–30). Oxford, United Kingdom: Oxford University Press.
- Lang, C., Siemens, G., Wise, A., & Gašević, D. (Eds.). (2017). *Handbook of learning analytics* (1st ed.). Society for Learning Analytics Research. 10.18608/hla17
- Mansilla, V. B., & Duraising, E. D. (2007). Targeted assessment of students' interdisciplinary work: An empirically grounded framework proposed. *The Journal of Higher Education*, 78(2), 215-237.
- Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014, March). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 193-202). ACM.)
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.
- Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.
- Wenger, E. (2011). Communities of practice: A brief introduction. National Science Foundation.

Fairness and Equity in Learning Analytics Systems (FairLAK)

Kenneth Holstein¹ and Shayan Doroudi²

¹Human Computer Interaction Institute, ²Computer Science Department
Carnegie Mellon University
{kjholste, shayand}@cs.cmu.edu

ABSTRACT: The potential for data-driven algorithmic systems to amplify existing social inequities, or create new ones, is receiving increasing popular and academic attention. A surge of recent work, across multiple researcher and practitioner communities, has focused on the development of design strategies and algorithmic methods to monitor and mitigate bias in such systems. Yet relatively little of this work has addressed the unique challenges raised in the design, development, and real-world deployment of learning analytics systems. This interactive workshop aims to provide a venue for researchers and practitioners to share work-in-progress related to fairness and equity in the design of learning analytics and to develop new research and design collaborations around these topics. The workshop will begin with a brief overview of research in fair AI and machine learning, followed by presentations of accepted and invited contributions. In addition, a key outcome of the workshop will be a *research agenda* for the LAK community, around fairness and equity. Workshop participants will collaboratively construct this agenda through a sequence of small- and whole-group design activities. At the end of the workshop, participating researchers and practitioners will then explore opportunities for collaboration around specific research and design thrusts within this agenda.

Keywords: fairness; equity; algorithmic bias; real-world impact; critical perspectives; human factors; design; ethics; AI; machine learning; cross-disciplinarity

1 BACKGROUND

Data-driven algorithmic systems increasingly influence every facet of our lives, including the quality of healthcare we receive, who receives a job or a loan, whose livelihoods are automated away, who is released from jail, and who is subjected to increased policing (e.g., Barocas & Selbst, 2016; Veale, Van Kleek, & Binns, 2018). In recent years, the potential of such systems to amplify existing social inequities, or even to create new ones, has received a surge of popular and academic attention. It is now commonplace to see popular press articles about algorithmic bias in high-stakes applications such as loan granting, hiring, recidivism prediction, and predictive policing (e.g., Giang, 2018; Lohr, 2018). Interdisciplinary research communities have emerged with a focus on understanding and mitigating such risks – most notably the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)¹ and the nascent FAT* community².

¹ <https://www.fatml.org/>

² <https://fatconference.org/>

Despite this widespread attention to fairness and bias in data-driven algorithmic systems, communities such as FAT/ML and FAT* have thus far tended to focus heavily on a relatively small set of high-stakes application domains such as the examples mentioned above (Green & Hu, 2018; Holstein et al, 2018; Veale et al., 2018). In particular, relatively little work has focused on *educational* contexts, where increasing use of learning analytics and AI raises unique challenges not commonly faced in other domains (Ocumpaugh, Baker, Gowda, Hansen & Reich, 2015; Heffernan, & Heffernan, 2014; Ito, 2017). For example, while most existing fairness auditing and “de-biasing” methods require access to sensitive demographic information (e.g., age, race, gender) at an individual-level, such information is often unavailable to learning analytics practitioners in practice (Holstein et al., 2018; Kilbertus et al., 2018). In addition, it can sometimes be challenging to define what “equitable” outcomes might look like (Hansen & Reich, 2015; Ito, 2017), in contexts where a learning analytics system results in disparate outcomes across student subpopulations (e.g., students coming in with lower or higher prior knowledge).

The Learning Analytics and Knowledge (LAK) community has long been interested in the ethical dimensions of data-driven educational systems (e.g., Draschler et al., 2015; Sclater & Bailey, 2015; Tsai & Gasevic, 2017). However, the focus has often been on institutional and policy level considerations, including concerns around data ownership and privacy. As multidisciplinary conversations around algorithmic fairness and bias proceed at a rapid pace, it is critical that they are not proceeding without us. It is crucial not only that the learning analytics community is *aware of* advances in understanding and mitigating undesirable algorithmic bias, but also that our community is *actively contributing* to these conversations. In addition to advancing the field of learning analytics, such direct engagement may help push the broader literature on algorithmic fairness forward, by presenting domain-specific nuances that need to be addressed or by challenging some of the literature’s core assumptions from an educational perspective. This workshop is particularly well suited for this year’s LAK conference, given the theme of promoting inclusion and success.

2 WORKSHOP OBJECTIVES AND INTENDED OUTCOMES

The primary goals of this workshop are as follows:

- A. Cross-disciplinary ‘translation’:** Introduce LAK researchers and practitioners to the state-of-the-art in fairness and bias in data-driven algorithmic systems.
- B. A venue to share relevant research and practice:** Provide a venue for researchers and practitioners to share in-progress research/design work or on-the-ground experiences related to algorithmic fairness and bias in learning analytics systems.
- C. Visioning / Developing a research agenda:** Collaboratively develop a research agenda for more equitable learning analytics, based on the open problems and directions identified by workshop participants.
- D. Researcher and practitioner ‘matchmaking’:** Helping participants identify opportunities for fruitful researcher-researcher and/or researcher-practitioner collaborations.

We will disseminate the shared research agenda developed at the workshop, along with other workshop outcomes, via a Twitter hashtag (#FairLAK). In addition, outcomes will be disseminated via

one or more blog posts (which will also be shared over social media, such as Twitter) and through a potential joint paper with workshop participants for LAK 2020 or the Journal of Learning Analytics.

3 WORKSHOP ORGANIZATION

Type of event: Half-day workshop

Type of participation: Participation for the first FairLAK workshop will be ‘mixed’: both participants with a paper submission (following an open call) and other interested members of the LAK community will be welcome to attend.

Schedule:

A. Introductions and background (~30 minutes): Workshop organizers will present high-level workshop objectives. Participants will briefly introduce themselves and share their personal objectives for the workshop. Then the organizers will provide a rapid overview of existing work on fairness in data-driven algorithmic systems. Researchers and practitioners will learn about existing, state-of-the-art methods (from FAT/ML and related literatures in machine learning, statistics, and human-computer interaction) to audit real-world learning analytics systems for potentially harmful biases, and strategies/methods to mitigate such biases.

B. Presentations of accepted and invited contributions (~80 minutes): Three accepted presentations and three invited presentations (8 minutes each, with 5 minutes for questions and discussion)

C. Collaborative group work (60 minutes):

C.1 Small-group discussions: Problem-finding (20 minutes): Participants will identify pressing open issues around fairness and equity in learning analytics systems, collecting issues on sticky notes in small-group discussions

C.2 Whole-group discussion: Sharing open problems and envisioning possible solutions (20 minutes): Groups will share the issues they have identified, synthesizing issues through affinity diagramming

C.3 Small-group discussions: Turning ‘possible solutions’ into research agendas for the LAK community (20 minutes): Groups will gather around particular areas of the growing affinity diagram (dynamically and self-selected, based on areas of interest), to discuss specific issues that interest them in greater detail – this time generating ideas for possible solutions and/or research projects

D. Synthesis, speed dating, and next steps (40 minutes):

D.1 Whole-group discussion: Developing a shared research agenda for fair learning analytics (20 minutes): Based on the activities above, the organizers will help groups

synthesize their ideas into a shared research agenda (i.e., a call to action for the LAK community, consisting of several concrete research and design directions)

D.2 Speed Dating and Closing Notes (20 minutes): Researchers and practitioners will circulate throughout the room, engaging in brief conversations with others to begin exploring concrete opportunities for collaboration

REFERENCES

- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *Cal. L. Rev.*, 104, 671.
- Drachsler, H., Hoel, T., Scheffel, M., Kismihók, G., Berg, A., Ferguson, R., Chen, W., Cooper, A., & Manderveld, J. (2015). Ethical and privacy issues in the application of learning analytics. In *LAK'15* (pp. 390-391). ACM. <http://dx.doi.org/10.1145/2723576.2723642>
- Giang, V. (2018, May 07). The Potential Hidden Bias In Automated Hiring Systems. Retrieved September 15, 2018, from <https://tinyurl.com/yd3uc5zq>
- Green, B., & Hu, L. (2018). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *ICML'18*.
- Hansen, J. D., & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, 350(6265), 1245-1248.
- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? To appear in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM.
- Ito, M. (2017, December 26). From Good Intentions to Real Shortcomings: An Edtech Reckoning [Web log post]. Retrieved September 28, 2018 from <https://www.edsurge.com/news/2017-12-26-from-good-intentions-to-real-shortcomings-an-edtech-reckoning>
- Lohr, S. (2018, February 09). Facial Recognition Is Accurate, if You're a White Guy. Retrieved September 15, 2018, from <https://tinyurl.com/GenderShadesNYTimes>
- Ocuppaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *BJET*, 45(3), 487-501. <http://dx.doi.org/10.1111/bjet.12156>
- Sclater, N., & Bailey, P. (2015). Code of practice for learning analytics. *JISC*.
- Tsai, Y. S., & Gasevic, D. (2017). Learning analytics in higher education---challenges and policies: a review of eight learning analytics policies. In *LAK'17*. ACM. <http://dx.doi.org/10.1145/3027385.3027400>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *CHI'18* (p. 440). ACM. <http://dx.doi.org/10.1145/3173574.3174014>

Ethics in Praxis: Socio-Technical Integration Research in Learning Analytics

Kyle M. L. Jones

Indiana University-Indianapolis (IUPUI)

kmlj@iupui.edu

Chase McCoy

Indiana University-Bloomington

chamccoy@indiana.edu

ABSTRACT: Learning analytics create beneficial opportunities to reimagine educational institutions, pedagogy, and learning experiences. However, it is unclear whether or not the processes that will create these benefits take into consideration ethical issues, such as fairness and privacy. For learning analytics to be legitimate, there is a need to make sure that data practices and technological designs limit downstream harms as much as possible. To do so, some have argued that transparency, auditing, and participatory design can achieve this goal; we argue that there is an opportunity to address ethical concerns upstream using a method from science and technology studies: Socio-technical integration research.

Keywords: Ethics, fairness, interventions, socio-technical integration research, social science methods

1 INTRODUCTION

Universities are enmeshed in ubiquitous information technologies that serve their educational aims and the administration of highly bureaucratic institutions. These assemblages of databases, applications, systems, sensors, and other technical artifacts have led to an undeniable increase in data quantity and with it the potential to transform data into actionable insights using machine learning, descriptive statistics, and predictive models. Often under the umbrella term of learning analytics, researchers, practitioners, and administrators are working to explore how methods derived from data science can potentially impact, if not transform, higher education. While learning analytics present a significant opportunity to examine pedagogy and learning outcomes, in addition to institutional structure and management practices, the positive benefits learning analytics may reap come with significant ethical questions. Research has emerged to address these questions, but the scholarly field and practitioner discipline need methods to identify and resolve these problems in the day-to-day work of learning analytics—not just at a theoretical level.

Our approach in this paper is to present a method from science and technology studies (STS)—socio-technical integration research, or STIR—and its potential to identify and influence ethics in praxis. What we mean by “ethics in praxis” is not *just* the instantiation of a moral choice in everyday learning analytics work. Instead, our focus is on how a STIR participant (a learning analytics

practitioner) defines, justifies, and acts on a moral perspective, embedding that perspective in and using that perspective to guide the work the practitioner does as a means to ends aligned with learning analytics. In brief, a STIR study situates social scientists alongside learning analytics practitioners to engage the latter in questions about their ethics in praxis and the social consequences of their work.

We begin this paper by arguing that the ethical questions surrounding learning analytics have been focused primarily on privacy concerns, and rightfully so. However, there are other problems worth investigation, including how practitioners make choices that protect students from harmful consequences and treat them fairly. Instead of focusing on effects downstream of learning analytics, we contend that addressing ethics in praxis upstream could be useful; to do just that, we outline the STIR method. Potential applications of STIR studies with learning analytics practitioners follow. Finally, we summarize our STIR study of an institutional researcher and conclude the paper with recommendations for the learning analytics community.

2 TOWARDS FAIRNESS IN LEARNING ANALYTICS

2.1. More Than Privacy

Information ethics scholars and learning analytics researchers have taken up some ethical concerns as they relate to informational privacy, but less so questions of fairness. Naturally, the creation of new information flows—many of which contain granular, identifiable data about student life—in support of learning analytics have raised student privacy concerns, and this area of the literature has demonstrated conceptual and theoretical rigor in ways that are having notable impacts on, inter alia, how institutions grapple with privacy problems in their policies (see Pardo & Siemens, 2014; Rubel & Jones, 2016). Other researchers are examining socio-technical solutions to scaffold important informed consent strategies in technological designs (see Prinsloo & Slade, 2015). However, the privacy literature has, with notable exceptions, only touched on issues of fairness (see Prinsloo & Slade, 2016; West, Huijser, & Heath, 2016). One way to understand fairness issues is to consider who benefits from learning analytics and whether or not the distribution is just.

2.2. Fairness and Just Distributions of Benefits

The capture and analysis of data representing students' social, intellectual, and physical behaviors that drive learning analytics lead us to ask two important questions related to fairness. First, what benefits accrue, for whom are they distributed, and is the distribution justifiable? We can easily imagine situations where data derived from student life is used to support administrative aims (e.g., efficiency, effectiveness, political gains)—but not positive learning outcomes and experiences for those whose lives are made transparent for data analysis purposes. Our second question homes in on processes informing learning analytics. Regardless of the *actual* benefits and how they are distributed, will the processes by which learning analytics insights are created directly benefit students and protect them from harm? If learning analytics are not beneficent and attuned to particular harmful consequences, then they cannot be considered fair practices. The first question attends to distributive justice, while this second question raises concerns about procedural justice and ethics in praxis—our focus for this paper.

2.3. Examining Downstream Effects

One way to shore up the legitimacy of learning analytics is to bring to light data practices and their results (e.g., interventions, predictions) to determine whether or not such things are justifiable. In information ethics and critical data studies, research efforts have focused on improving transparency around black-boxed data artifacts (see Citron & Pasquale, 2014). The general argument for doing so is a Brandeisian one: Transparency will hold those who create, distribute, and implement data artifacts more accountable; consequently, accountability will resolve discriminatory and/or deceptive practices and increase fairness (Ananny & Crawford, 2016). One weakness of this approach is that it tends to place its attention on *downstream* practices and artifacts that are already established and mature. As a result, technological recommendations and policy suggestions attempt to slow down and reverse that which has technological momentum. While we support this type of research and these ongoing initiatives, we also believe directing research on learning analytics *upstream* could lead to fairer, ethically sensitive technological designs and practices.

2.4. Addressing Ethics in Praxis Upstream

Important efforts are being made by researchers and designers to design learning analytics systems and data artifacts with particular users in mind, and other work—such as that which takes a participatory/co-design strategy—develops learning analytics hand-in-hand with actual users in the design stage. For instance, Zhu, Yu, Halfaker, and Terveen (2018, p. 2) suggest a novel “value-sensitive algorithm design” process, which “engages relevant stakeholders in the early stages of algorithm creation and incorporates stakeholders’ tacit values, knowledge, and insights into the abstract and analytical process of creating an algorithm.” These efforts are *crucial* for identifying and resolving ethical problems upstream before they are baked into learning analytics technologies. However successful these approaches may become, they cannot fully account for socially situated practice.

Socio-technical user studies have time and again demonstrated that tool use is dependent on the social context in which the user is situated (see Oudshoorn & Pinch, 2003). And while participatory/co-design/value-sensitive design strategies of learning analytics can account for one aspect of upstream ethics, particular uses (or non-uses as may be the case) of these technologies depends on conditions, norms, and values that are sometimes hard to identify and account for in design. Moreover, not all practices depend on specific learning analytics technologies. In fact, it is still commonplace that data visualizations, statistical models, and other analytic practices are done using off-the-shelf applications (e.g., Tableau, SPSS, Excel). As a result, upstream interventions also need to address how practitioners interact with tools in support of learning analytics and account for the social context in which practitioners work day to day.

An approach of this sort requires researchers to get into the very spaces and places where consequential decisions are made about how to make students into data and consider them as data artifacts (Jones & McCoy, 2018). More importantly, such an approach would need to go beyond descriptive studies of what is happening and move towards actively intervening in analytic work. In so doing, practitioners would be prompted by researchers to become reflexive about their practices

and the consequences thereof to make responsive practical and ethical modulations. We argue that the socio-technical integration research (STIR) method can lead to positive upstream engagement and useful modulations.

3 THE SOCIO-TECHNICAL INTEGRATION RESEARCH METHOD

3.1. Who and What to STIR

In socio-technical integration research (STIR), social scientists embed themselves within a research context to actively engage with researchers by probing and encouraging them to reflect on the societal dimensions and implications of their practices. STIR was initially developed to provide laboratory scientists the opportunity to pair with social scientists in order to enable collaboration between them and to aid laboratory scientists in unpacking "the social and ethical dimensions of research and innovation in real time and to document and analyse [*sic*] the results" (Fisher, n.d., p. 76).

Social scientists STIR participants (practitioners participating in a study) to provide opportunities for them to reflect on what they are doing, why they are doing it, and how they could do things differently, with the end goal being that participants will actively modulate their behavior by considering the social aspects of their work. During their time together, the STIR researcher seeks to elicit "reflexive awareness," or an attentiveness to "the nested processes, structures, interactions, and interdependencies, both immediate and more removed, within which they operate" (Fisher, Mahajan, & Mitcham, 2006, p. 492) for the STIR participants. Participants are encouraged to reflect upon three areas of their practice: considerations; alternatives, and outcomes.

Considerations refer to the particulars of their practice, including the goals and values of their work, as well as the social, political, and technological resources from which they draw for support. Alternatives are practices that differ from the participants' current ones but could impact the trajectory of their work if they were to be adopted. Finally, with outcomes participants are encouraged to reflect upon the outcomes of their work and if different decisions, approach, resources, and people could influence their practice. As STIR participants reflect upon these three areas and related societal and ethical dimensions, opportunities emerge for participants to recognize their socio-ethical position, which in turn leads to "goal-directed" (p. 492) modulations that directly impact the participant's current practices.

3.2. Modulations: De Facto, Reflexive, and Deliberate

Modulations in STIR occur in three stages: de facto, reflexive, and deliberate. De facto modulations are the implicit societal and ethical dimensions that shape research participants' everyday work practices and exist prior to a STIR. The STIR approach assumes that participants do not actively reflect on whether these dimensions are efficacious or in alignment with their norms and values or those of the social context that guides their practices, because there is no incentive to do so. Reflexive modulations are those that arise because of heightened awareness as the participant is probed to consider the societal and ethical dimensions of their practice and the consequences. In these cases, these dimensions are made explicit by the participant, and they begin to notice how

social influences (e.g., actors, politics, values, resources, etc.) interact with their given practice. In the final stage, deliberate modulations, participants act upon their reflexive modulations to make changes to their practices. These deliberate modulations may simply influence the efficiency and effectiveness of their work; however, deeper level modulations—which is the goal of STIR-ing—lead to altered goals, objectives, and assumptions of a project due to an enhanced awareness to the societal implications of their practice.

4 STIR AND LEARNING ANALYTICS

Social scientists can use socio-technical integration research (STIR) to uncover the societal and ethical dimensions of learning analytics practitioners as they build systems, develop data-based artifacts, and deploy analytic strategies (e.g., algorithms, models); see Figure 1.

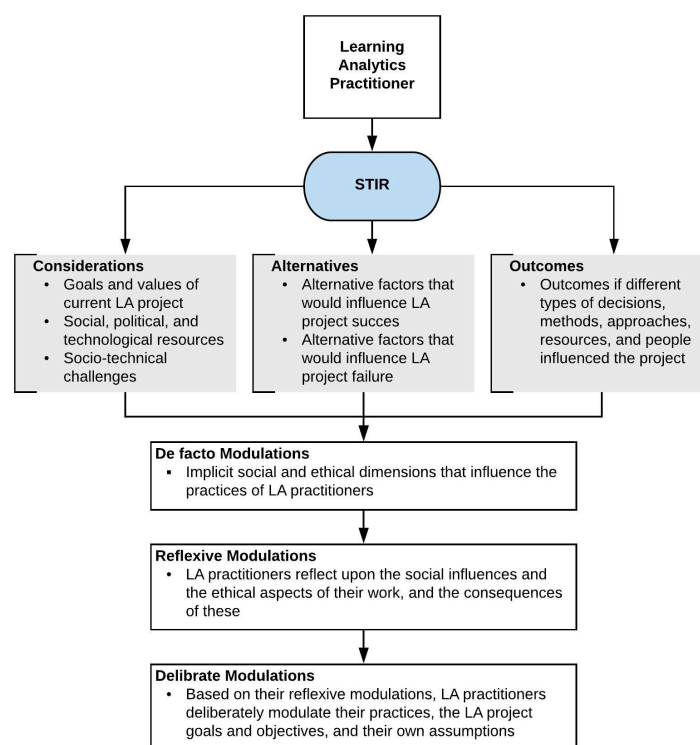


Figure 1. A model showing how a STIR of a learning analytics practitioner could lead to de facto, reflexive, and deliberate modulations.

But given that learning analytics is often embedded in and supportive of complex institutional bureaucracies (e.g., higher education), defining who is a learning analytics practitioner can be challenging. Unlike original STIR studies where it was quite obvious that the laboratory was the context and bench scientists were the participants, with learning analytics it can be difficult to make these methodological choices. Below, we make some recommendations for STIR-ing learning analytics practitioners.

Academics. In the spirit of original STIR studies, STIR-ing learning analytics could be done with research teams building learning analytics artifacts. Such individuals represent academic

departments and cross-institutional collaborations. Findings may reveal varying degrees of ethical sensitivity among different types of researchers (e.g., doctoral students, postdocs, tenured faculty).

Mathematical and Computational Scientists. Individuals responsible for programmatic and algorithmic code effectively write some of the rules of individual behaviors and determine the information they use to evaluate themselves (van Dijk & Poell, 2013). STIR-ing these practitioners could help them better understand how they embed their values and that of the institutions for whom they are designing learning analytics systems.

Interface and User Experience Technologists and Instructional Designers. Practitioners focused on human-computer interaction processes are steeped in affective and persuasive computing methods, which are often used to elicit particular user responses using design strategies and messaging campaigns (e.g., nudging). A STIR study of these practitioners could surface the ethical justifications designers make to, say, limit choice sets or educate students about predictive scores.

Educational Technologists and Instructional Designers. Technologists and designers in educational institutions are in unique positions to educate instructors on how to use learning analytics tools. STIR-ing these individuals could raise their awareness about student privacy issues, among other things.

Institutional Researchers, Registrars, and Other Information Professionals. In higher education, the deployment and successful diffusion of learning analytics tools and practices are impacted by various information professionals who access, steward, and analyze sensitive institutional information. STIR-ing these practitioners could develop interesting findings regarding their decision making around information disclosure and institutional politics.

Apropos to the last category of learning analytics practitioners above, in the next section, we will discuss a longitudinal STIR we conducted on a single institutional researcher engaged in developing learning analytics data artifacts for their institution's administration. The study will be explicated further in a forthcoming publication, but for this workshop paper, we will briefly discuss our preliminary findings.

5 STIR-ING A LEARNING ANALYTICS PRACTITIONER

To understand how STIR can help to uncover learning analytics ethics in praxis, we conducted a STIR study of a single institutional researcher at a mid-sized public university. The participant's responsibilities entail, among other things, conducting statistical analyses on important administrative metrics, such as retention, recruitment, and enrollment, and providing this information to their institution's administration. The STIR focused on assessing the potential value of the approach for uncovering and better understanding this practitioner's upstream privacy practices.

Over four months, we conducted 12 in-person and virtual interviews with the participant. We developed a STIR interview protocol based on elements in Figure 1 to guide the participant to reflect on their privacy practices and those of their staff within their office. The interviews sought to elicit

from the participant the considerations, alternatives, and outcomes of their work, and to uncover the three types of modulations and instances where their practices were modified to more explicitly consider privacy or, at the least, brought about ideas for future privacy-focused initiatives. Furthermore, during the interviews, the participant often shared data artifacts, such as an ongoing project on enrollment projections and trends, while discussing the data practices associated with their everyday work.

The participant's de facto modulations revealed that they value privacy in their work in regards to ensuring that they and their staff follow privacy policies set by FERPA and their institution. However, the participant's reflexive modulations uncovered that these guiding policies insufficiently address privacy issues in practice. The participant became aware that many of their office's and institution's actual privacy practices are not addressed in the policies, particularly, for example, in regards to how identifiable student data should be distributed throughout the institution, and who should be allowed access to sensitive student information. The participant reflected on how this policy lacuna has led to data access and distribution practices that differ between them and their colleagues throughout their institution.

The learning analytics practitioner's reflexive modulations gave rise to deliberate modulations. Here, not only did the participant become aware of the need to have more formal institutional and departmental policies to guide privacy in praxis, but they began the process of documenting their privacy practices. By working with other institutional actors engaged in learning analytics, conversations within in the practitioner's institution and within their office have begun around creating explicit institutional and departmental documents on with whom and how data should be shared within their institution. Furthermore, the participant stated the planned to establish opportunities, such as at an office retreat or during team meetings, for their staff to document their privacy practices.

6 CONCLUSION

In this paper, we have argued that socio-technical integration research (STIR) presents new opportunities to investigate how learning analytics practitioners define, justify, and act on their moral perspectives—their ethics in praxis. Our study suggestions and the summary of our forthcoming research provide insights into what STIR may accomplish. Yet, the learning analytics community may benefit from more structure to begin STIR studies of their own and, more importantly, adopt a reflective perspective about their ethics in praxis.

To increase the adoption of the STIR method, we see an opportunity to develop a multi-faceted research and training agenda. First, social scientists addressing ethical issues associated with learning analytics could develop a research agenda to further explore STIR and related intervention methods, as well as plan strategic STIR studies. Such an agenda could be developed at a pre-conference workshop or special research retreat, among other things. Should this agenda gain traction, learning analytics and STIR experts could develop training materials for non-STIR experts. While STIR is a rigorous method, we believe that it does not take advanced qualitative research training to learn its intricacies and apply its techniques. Non-research learning analytics practitioners could learn how to STIR and conduct STIR evaluations at their place of work.

The ethical issues associated with learning analytics are many and consequential. From privacy to discrimination, bias to fairness, and many others, these concerns deserve serious attention to ensure that learning analytics technologies are designed and deployed in ways that further the educational mission of higher education and protect its primary stakeholder group—students—and others from harm. Scholarly efforts to date have cataloged many of these issues, and in so doing they have recommended sound policy principles. While there is still more work to do on this front, it is arguably time to shift efforts to focus on how such ethical concerns materialize and are accounted for in everyday practice.

REFERENCES

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. doi: 10.1177/1461444816676645
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, 1–33. Retrieved from <https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf>
- Fisher, D. E. (n.d.). Causing a STIR. Retrieved from <https://sciencepolicy.colorado.edu/news/fisher.pdf>
- Fisher, E., Mahajan, R. L., & Mitcham, C. (2016). Midstream modulation of technology: governance from within. *Bulletin of Science, Technology & Society*, 26(6), 485–496. doi: 10.1177/0270467606295402
- Jones, K. M. L., & McCoy, C. (2018). Reconsidering data in learning analytics: Opportunities for critical research. *Learning, Media and Technology*. doi: 10.1080/17439884.2018.1556216
- Oudshoorn, N., & Pinch, T. (Eds.) (2003). *How users matter: The co-construction of users and technology*. Cambridge, MA: MIT Press.
- Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450. doi: 10.1111/bjet.12152
- Prinsloo, P., & Slade, S. (2015). Student privacy self-management: Implications for learning analytics (pp. 83–92). *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. doi: 10.1145/2723576.2723585
- Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. *Big Data and Learning Analytics in Higher Education*, 109–124. doi: 10.1007/978-3-319-06520-5_8
- Rubel, A., & Jones, K. M. L. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, 32(2), 143–159. doi: 10.1080/01972243.2016.1130502
- van Dijk, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, 1(1), 2–14. doi: 10.17645/mac.v1i1.70
- West, D., Huijser, H., & Heath, D. (2016). Putting an ethical lens on learning analytics. *Educational Technology Research and Development*, 64(5), 903–922. doi: 10.1007/s11423-016-9464-3

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction - CSCW*. doi: 10.1145/3274463

Early-adopter Iteration Bias and Research-praxis Bias in the Learning Analytics Ecosystem

Michael J. Meaney

University of Cambridge
Arizona State University
mjm234@cam.ac.uk

Tom Fikes, Ph.D.

Arizona State University
tgfikes@asu.edu

ABSTRACT: By devising a conceptual framework of the learning analytics ecosystem, we identify two types of bias that may stymie the efforts of leveraging learning analytics to produce fair and equitable virtual learning environments. First, Early-adopter Iteration Bias may lead learning analytics to derive insights about optimal course design based on preferences and behavior patterns of more prepared, lower need learners. Second, Research-praxis Bias prevents practitioners from properly utilizing insights derived from learning analytics and research.

Keywords: Educational equity, Human computer interaction, Interaction design, Design bias

1 INTRODUCTION

In the context of open-scale courses (including Massive Open Online Courses, or MOOCs), the learning analytics ecosystem has the potential to provide valuable insights into teaching and learning for virtual learning environments (VLEs) (Nguyen et al., 2017). This may be especially valuable for scaling low-barrier, individualized learning experiences that can reach traditionally underrepresented populations or other high-need students (Aguilar, 2018). The broader educational systems in which learning analytics are embedded, however, give rise to multiple sources of bias that may stymie the efforts to develop these courses into fair and equitable VLEs. First, an Early-adopter Iteration Bias may unintentionally lead to design recommendations that serve already well-educated and well-represented learners (Meaney and Fikes, 2018). Because analytics and research inform practice, if the data are not adequately disaggregated, and heterogeneous effects considered, conclusions will be biased toward the majority and drive the innovation and optimization of the courses to further favor these students, potentially disadvantaging underrepresented learners, or other high-need students. Second, Research-praxis Bias, whereby the producers of VLEs do not properly benefit from learning analytics and research insights into VLEs, might further prevent VLEs from meeting the needs of underrepresented or other high-needs learners (Meaney, 2018). A depiction of the learning analytics ecosystem that highlights these sources of design bias is illustrated in Figure 1.

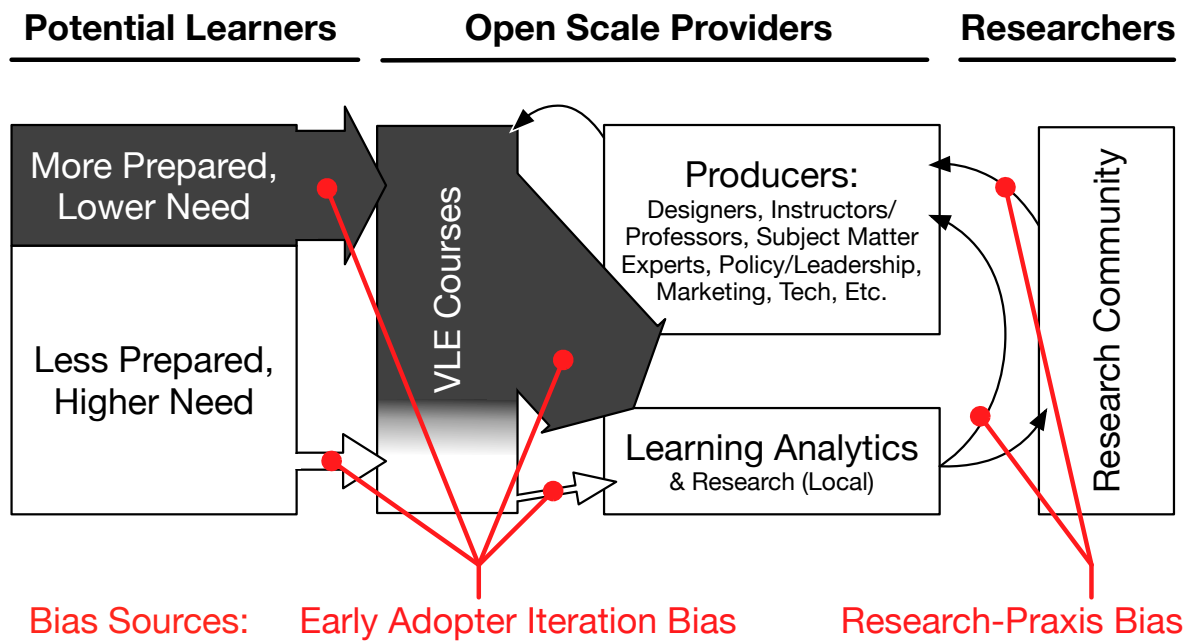


Figure 1: A model of the learning analytics ecosystem illustrating two sources of bias. The universe of students who could benefit from VLEs contains a high proportion of less prepared, higher needs students. *Early-adopter Iteration Bias* describes the situation in which courses designed for traditional higher-education students lead students from more prepared, lower need backgrounds to disproportionately enter VLEs and then succeed at higher rates. The data corpus produced by VLEs reflects the population of more prepared, lower need learners; and learning analytics and research conducted on this corpus produces results biased toward the majority. *Research-praxis Bias* describes the situation in which producers of VLEs receive insights from learning analytics and the research community that is driven by the more prepared, lower need majority, leading to innovation and optimization of VLE design that is even further away from the needs of less prepared, higher needs students. This is further complicated by the general disconnect between the research and practice communities.

2 EARLY-ADOPTER ITERATION BIAS

Early-adopter Iteration Bias is a conceptual model we are introducing to account for a series of processes and constraints that optimize open-scale course production for more prepared, lower need learners. The intuition is grounded in Rogers' (2010) notion that early adopters of technology will often have population characteristics different to that technology's later users, which may be the actual target population. Learning analytics of massive data sets have focused on behavior patterns of the average student, who are (we suggest) early-adopters who are more likely to be already well-educated (Rohs and Ganz, 2015; van de Oudeweetering and Agirdag, 2018). This leads optimization and design recommendations to be driven by insights derived from users less likely to need help. If future open-scale course iterations continue to be optimized based on present usage patterns of early-adopters, and if these usage patterns continue to reflect the needs and behaviors of more prepared, lower need learners, this could further exacerbate educational inequity by disadvantaging less prepared, higher need learners. Early-adopter Iteration Bias is illustrated in Figure 2.

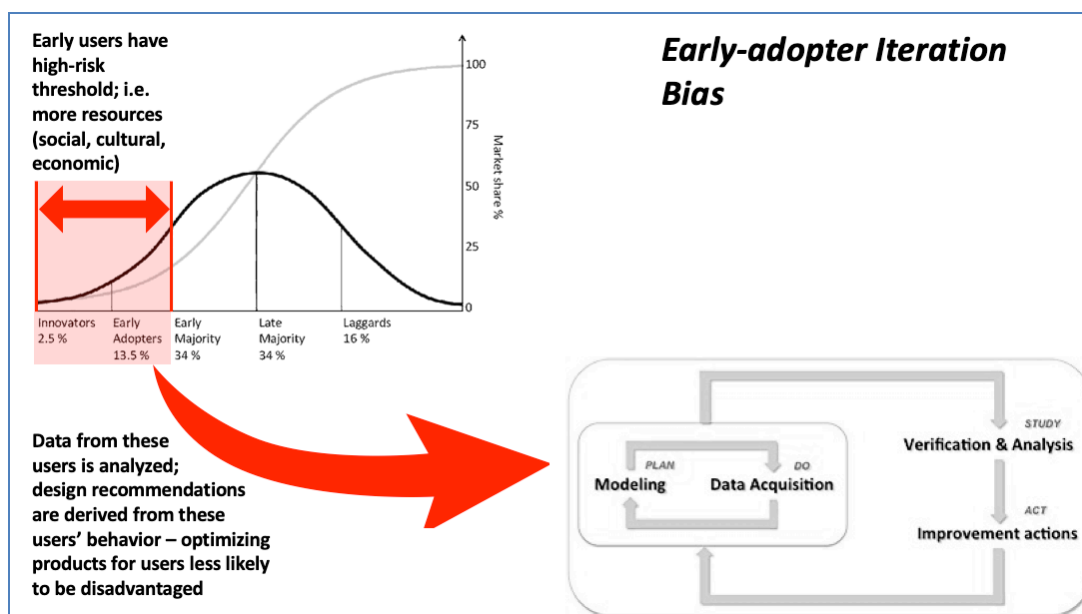


Figure 2: The diffusion of innovations is a concept developed by Rogers (2010). The theory suggests that innovations diffuse across society along different segments of the population, sequentially: innovators, early adopters, early majority, late majority, and laggards. Rogers notes that early adopters of new technologies will more likely be well-educated and wealthier. These users have access to more and better information, coupled with a higher tolerance of risk for new products. Early adopters are also likely to have disposable income and are a more attractive target market toward which to design new products. Innovations are iterated and optimized based on data available from early adopters.

Given the disproportionate rate of already well-educated learners using open scale courses and other low barrier VLEs, it is possible that Early-adopter Iteration Bias has already entered the learning analytics ecosystem. We created a graphic highlighting the educational attainment of users studied in eight learning analytics papers over the past few years. Nearly 80% of users already held a college degree, as illustrated in Figure 3 (data cited from: Robinson et al., 2015; Dillahunt et al., 2015; Christensen et al., 2013; van de Oudeweetering and Agirdag, 2018; Ho et al., 2015; Wang et al., 2018).

Arizona State University's Global Freshman Academy (ASU GFA) stands out for attracting a higher proportion of less prepared, higher need students. These courses offer university credit eligibility and earned-admission to ASU Online, and are intentionally designed to attract non-traditional learners without a post-secondary degree. Even still, more than half of learners in this VLE are more prepared, lower need learners.

Scaling low-barrier, individualized learning experiences that can reach traditionally underrepresented populations or other high-need students requires not only new marketing strategies, but also VLE content and pedagogy to suit the needs of these students. Analyzing data and deriving insights biased toward the majority of existing users may actually undermine this aim.

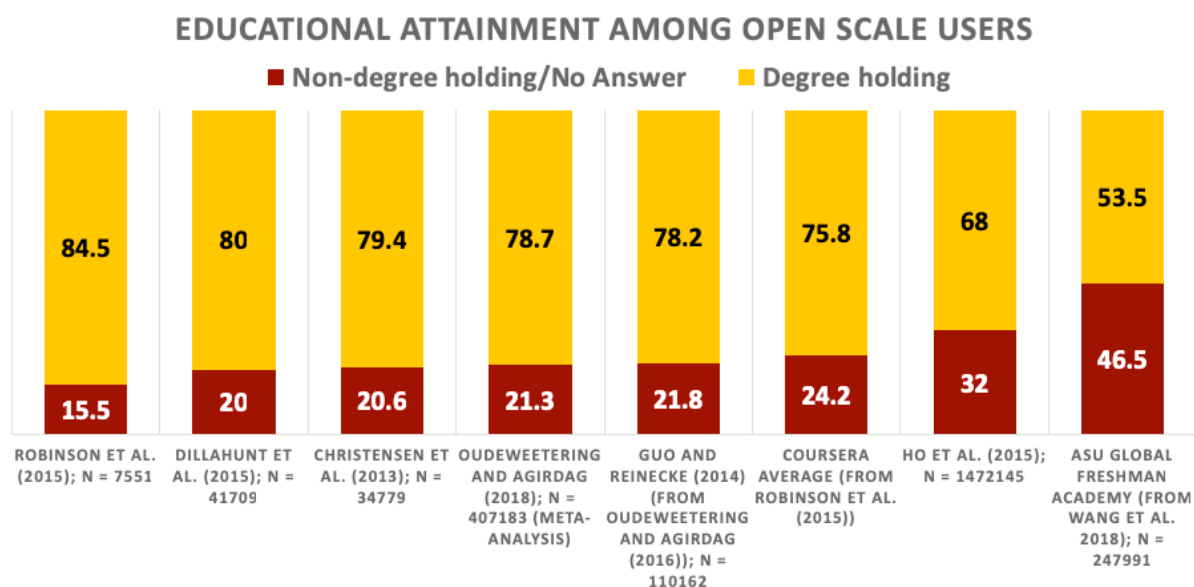


Figure 3: More prepared, lower need learners make up the majority of users in data analyzed by the learning analytics research community. This data drives innovation and optimization recommendations to course design, which might unintentionally lead to courses less suited for less prepared, higher need learners.

3 RESEARCH-PRAXIS BIAS

Research-praxis Bias compounds the potential problems from Early-adopter Iteration Bias, in two separate but interrelated ways. The first source of Research-praxis Bias is straightforward: practitioners who utilize the research and insights of the learning analytics community potentially embed into the design of the courses recommendations and conclusions derived from skewed data privileging behavior patterns of more prepared, lower need students.

The second source of Research-praxis Bias is more nuanced and complex. It was unveiled through recent qualitative work, a pilot study interviewing practitioners producing VLEs at ASU (Meaney, 2018), and builds off of the noted “chasm” between research and practice in the development of VLEs (Price et al., 2016; Bakharia et al., 2016). The qualitative study notes three important insights that may be worth further consideration and investigation and that, indeed, contribute to a Research-praxis Bias in the learning analytics ecosystem that may hinder aims of inclusion and equity.

First, it was discovered that relatively little is known about the production process of VLEs. Little research has been conducted to examine how the particular mindsets and processes of practitioners producing VLEs may impact design and thus, student outcomes.

Attempting to partly rectify this gap by interviewing producers of VLEs, a second, somewhat simple, but noteworthy insight was made: that it is important to not treat the practitioner community as a homogenous block. There are professors who create content; there are learning designers mediating the construction of the VLEs; and there are program managers charged with recruiting students and making the program sustainable, amongst others. These different subgroups of practitioners bring significantly different work and educational backgrounds, differing definitions of the ideal end user,

and different pedagogic paradigms to their design and production processes. These differences contribute to visions and goals for the product that are not always in alignment.

Some practitioners, for example, might take student self-regulation as a pre-requisite for successful completion of courses in a VLE; this can yield design choices less concerned with trying to equip students with study habits and time management strategies. There is some evidence that highly self-regulated users are more likely to complete courses, more likely to be older, and more likely to have a graduate degree (Kizilcec et al., 2017). These design orientations play significant roles in the production processes of these courses, and will have impact on the subsequent outcomes for heterogenous populations of learners. The learning analytics and research communities should take such differences and the resulting design dynamics into account when analyzing whether VLE designs promote educational equity.

Third, there is a noticeable variance among the practitioners' access and utilization of theory and academic research as a guide to their work. Some practitioners have a background in critical theory and disability studies, along with other theories from their post-graduate studies, and bring these to bear as theoretical lenses to their work. Others rely on a more quantitative, behaviorist view. The academic literature and discourse about open scale courses and MOOCs is often not visible or accessible to these practitioners. This represents a challenge and opportunity for the learning analytics research community to consider how to better disseminate their findings to a practitioner audience.

Determining how to better disseminate research insights in a constructive and actionable way to the practitioner community would be a worthy goal for learning analytics research community moving forward. Additionally, it seems that the learning analytics research community might consider some of the perspectives of practitioners themselves and create a more reciprocal work arrangement. The critical theory and disabilities studies referenced by practitioners might help guide learning analytics researchers to more thoughtfully sub-group and disaggregate data in order to pay more attention to groups who might be marginalized. This approach might help ensure that specific learning needs of certain populations of users are not obscured by the generalized and averaged insights produced by big data.

There are a number of vexing challenges to bridging the divide between research and practice (Prieto et al., 2018) that are beyond the scope of this paper. We do note, however, that the divide cuts both ways: the learning analytics and research community has much to offer the practitioner community in terms of specific insights and observations regarding student behavior derived from data, and the practitioner community has much to offer in terms of knowledge of learning theory, technology development, and differentiated teaching strategies for sub-groups of learners, among other insights. These insights should influence and build off of each other, hopefully resulting in a more informed, deliberate, careful, and, ultimately, more fair and equitable, construction of courses for learners.

4 CONCLUSION

MOOCs and other open-scale VLEs were intended to broaden access to high quality post-secondary education (Agarwal, 2013). Research has shown that, instead, most users are from more prepared, lower need backgrounds (Rohs and Ganz, 2015; van de Oudeweetering and Agirdag, 2018).

Diagnosing the sources of this dissonance is of paramount importance, especially as MOOCs and open scale courses approach an inflection point. Some members of the learning analytics and research community observe an imminent shift in strategy, in part resultant from the failure to make MOOCs and open scale courses more fair and equitable. A recent article in *Science* summarized the past few years of research on these VLEs, noting that the courses “disproportionately drew their learners from affluent countries and neighborhoods, and markers of socioeconomic status were correlated with greater persistence and certification,” (Reich and Ruipérez-Valiente, 2019). The researchers assert that universities may be doubling-down on this model: after hoping to reorient higher education toward providing access to a broadly defined conception of traditionally underserved learners, “we see the field instead coalescing around a different, much older business model: helping universities outsource their online master's degrees for professionals,” (Reich and Ruipérez-Valiente, 2019).

The research and practitioner communities may have inadvertently played a role in accelerating this shift. Our learning analytics ecosystem model hypothesizes that, despite the good intentions and noble efforts of researchers and practitioners, certain biases have unintentionally made the challenge of serving less prepared, higher needs learners more difficult. Early-adopter Iteration Bias may skew learning analytics and research toward recommendations that optimize course design for more prepared, lower need learners. Research-praxis Bias prevents the broader VLE producing community from fully utilizing the insights derived from learning analytics and research properly. We should note that, while closing the chasm between research and practice could greatly improve the design of MOOCs and open scale courses, this in itself would be insufficient; the insights derived from learning analytics and research may already be skewed as a result of Early-adopter Iteration Bias. Seeking to resolve these challenges requires a simultaneous approach.

We invite members of the learning analytics, research, and practitioner communities to reflect on this learning analytics ecosystem model and its implications with us, in the hope that doing so might help identify strategies to rectify these biases and the fairness and equity problems they may be exacerbating.

REFERENCES

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. doi: 10.1177/1461444816676645
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, 1–33. Retrieved from

- <https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf>
- Fisher, D. E. (n.d.). Causing a STIR. Retrieved from <https://sciencepolicy.colorado.edu/news/fisher.pdf>
- Fisher, E., Mahajan, R. L., & Mitcham, C. (2016). Midstream modulation of technology: governance from within: *Bulletin of Science, Technology & Society*, 26(6), 485–496. doi: 10.1177/0270467606295402
- Jones, K. M. L., & McCoy, C. (2018). Reconsidering data in learning analytics: Opportunities for critical research. *Learning, Media and Technology*. doi: 10.1080/17439884.2018.1556216
- Oudshoorn, N., & Pinch, T. (Eds.) (2003). *How users matter: The co-construction of users and technology*. Cambridge, MA: MIT Press.
- Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450. doi: 10.1111/bjet.12152
- Prinsloo, P., & Slade, S. (2015). Student privacy self-management: Implications for learning analytics (pp. 83–92). *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. doi: 10.1145/2723576.2723585
- Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. *Big Data and Learning Analytics in Higher Education*, 109–124. doi: 10.1007/978-3-319-06520-5_8
- Rubel, A., & Jones, K. M. L. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, 32(2), 143–159. doi: 10.1080/01972243.2016.1130502
- van Dijk, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, 1(1), 2–14. doi: 10.17645/mac.v1i1.70
- West, D., Huijser, H., & Heath, D. (2016). Putting an ethical lens on learning analytics. *Educational Technology Research and Development*, 64(5), 903–922. doi: 10.1007/s11423-016-9464-3
- Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction - CSCW*. doi: 10.1145/3274463

STRATEGIC OMISSION AND RISK AVERSION: A BIAS-RELIABILITY TRADEOFF

David Lang

Stanford University

david.nathan.lang@stanford.edu

ABSTRACT: Whether high-stakes exams such as the SAT or College Board AP exams should penalize incorrect answers is a controversial question. In this paper, we document that penalty functions can have differential effects depending on a student's risk tolerance. Moreover, literature shows that risk aversion tends to vary along other areas of concern such as race, gender, nationality, and socioeconomic status. In this article, we simulate Item Response Theory (IRT) data with and without a wrong answer penalty. In the presence of mild risk aversion, we find that students omit 12% more items than risk neutral individuals with identical ability. This translates into a nearly 2% difference in sum scores between the risk neutral and risk averse groups. We also find that penalty functions result in noisier estimates of student ability. These findings suggest that random guessing penalties should not be used in most circumstances, particularly for learning platforms.

Keywords: learning analytics, item response theory, risk aversion, differential item function, differential test function, simulation

1 MOTIVATION

In the past decade there have been notable shifts in the decision to penalize wrong answers in high-stakes testing. In 2010, the College Board removed its wrong answer penalty for the AP exams. The SAT has also removed this penalty from its exams in recent years.

In this paper, we explore whether learning platforms should follow suit. Many platforms implicitly or explicitly penalize guessing through either gamification mechanisms such as point systems or through hint generation. These designs often are associated with increased user engagement or performance but they may have downstream impacts on certain types of users (O'Rourke, Haimovitz, & Ballweber, 2014). Simulation may help us understand how these design features influence student behavior.

2 LITERATURE REVIEW

While much of the literature surrounding high-stakes testing has focused on bias in terms of gender and race/ethnicity, relatively little focus has been put forth into the effects of how random guessing penalties may mediate this bias. Past work points out that most exams with a penalty function are still designed so that a person who tries to maximize their average score will be indifferent to always guessing (Budescu & Bar-Hillel, 1993). Moreover, they point out that this penalty function introduces systematic biases for students. If students have a different objective (e.g. get a passing grade or get the top grade in the class), then these incentives may not hold. Other work found that there were substantial differences by gender in willingness to guess in the face of a penalty function (Baldiga, 2013). To date, there has been even less focus on how risk aversion affects the psychometric properties of these assessments.

2.1 Risk Aversion

There are three broad classifications of risk tolerance: risk-aversion, risk-preferring, and risk-neutrality. To understand these distinctions, consider a coin-flip bet where a person wins a dollar if the coin lands heads and loses a dollar if the coin lands tails. A risk averse person will never take a bet with an average payoff of zero. A risk-preferring person will always take this bet. The risk neutral person will be indifferent between taking this bet and not taking this bet.

In this paper, we model risk aversion using an exponential utility function:

$$U(points, risk_{tolerance}) = \frac{1 - e^{-points * risk_{tolerance}}}{risk_{tolerance}}$$

The components of the function are points (the number of points awarded or lost) and risk tolerance. Positive risk-tolerance parameters correspond to risk-aversion. Negative risk-tolerance parameters correspond to a risk-preferring behavior. In a testing framework, if the utility of attempting a question is positive, the examinee will attempt it. Otherwise, the examinee will omit it. This function exhibits several useful properties. First, it exhibits a constant coefficient of relative risk aversion. In decision analysis literature, this property is also known as the ‘delta property’ (Kirkwood, 1997). This property assures that an individual will have the same preferences regardless of their current wealth endowment. In a testing framework, this means that an individual’s decision to omit a particular item will not depend on one’s current score. This assumption is fairly reasonable for small scale decisions, such as one question on a forty-question exam. Additional benefits of this assumption are that it eliminates concerns with respect to item ordering effects interacting with risk aversion, and unlike other potential utility functions, this function can be transformed into a risk-averse/risk-preferring function simply by assigning a positive/negative risk tolerance value.

In terms of understanding what risk aversion looks like in the real world, most estimates suggest that individuals have positive risk tolerance and that a risk tolerance parameter of one is not unreasonable (Gandelman & Hernández-Murillo, 2014). Figure 1 shows that point estimates of risk aversion in the United States is around 1.5 . The most extreme countries are the Netherland with a risk tolerance of less than a quarter and Taiwan with a risk tolerance of nearly 2.5.

3 MODEL

To assess the question of omission on exams, we simulate a forty-question exam. The exam data is modeled as Rasch data such that each individual’s true ability estimate is known to us. The probability that a student will answer an item correct can be expressed by the following formula where θ_i corresponds to the ability of student i and b_j corresponds to the difficulty of item j :

$$\frac{1}{1 + e^{(\theta_i - b_j)}}$$

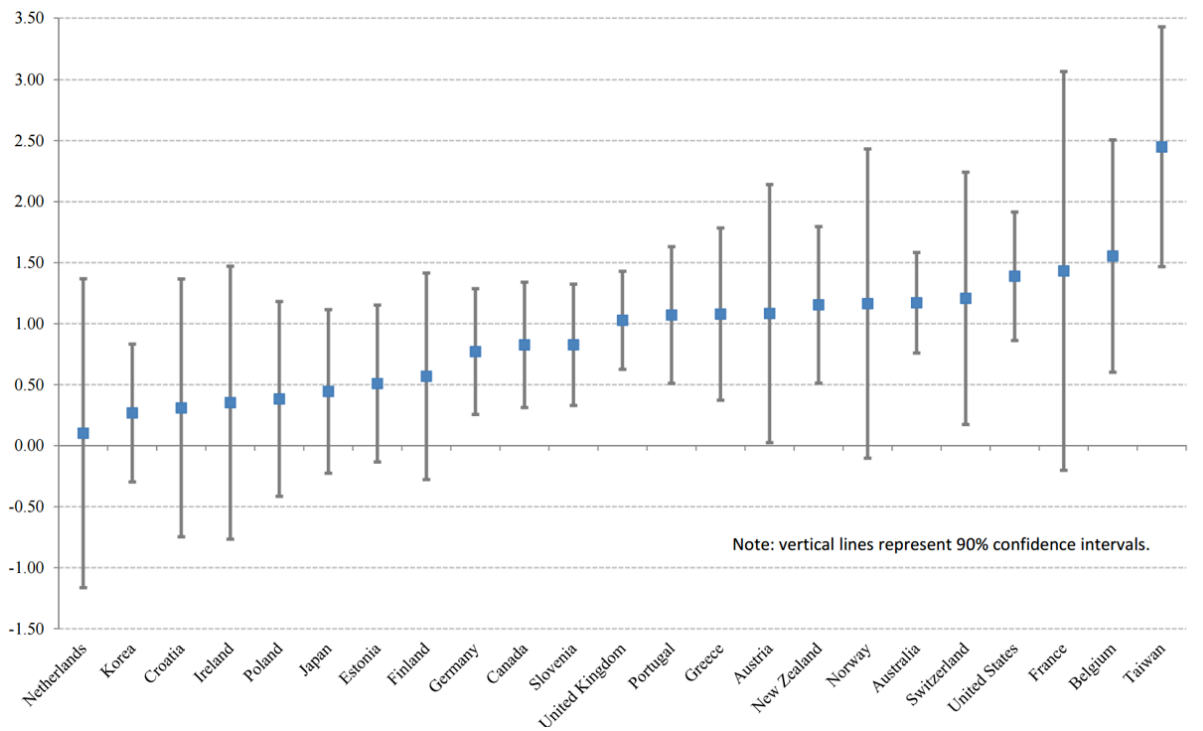


Figure 1 Relative Risk Aversion in Developed Countries (Source:St Louis Fed)

We further assume that students are aware of their ability and item difficulty but are uncertain whether or not they get the specific item correct. We also assume that they are aware of a one-quarter point penalty if they answer a question incorrectly. In this case, the students will respond to an item only if the expression below holds:

$$\Pr(\text{Correct}|\theta_i, b_j) * U(1, risk_{tolerance}) + (1 - \Pr(\text{Correct}|\theta_i, b_j)) * U\left(-\frac{1}{4}, risk_{tolerance}\right) \geq 0$$

We then re-estimate a person's ability based on their responses under three separate scenarios: (1) no penalty, (2) risk-neutrality, (3) risk-aversion with a risk tolerance of 1. We then repeatedly estimate the difference between these three groups and our true ability measures to assess whether or not this biases estimates of test performances. The underlying data generation process assumes both ability and item difficulty follow the standard normal distribution.

Figure 2 illustrates the utility of responding to a question in which the student is aware of the probability they will get the question right. The horizontal line at zero identifies the locations at which students of varying risk tolerances will be indifferent to answering the question and omitting their response. Points above the zero line correspond to attempting the item. Points below the line correspond to omitting the item. The dashed-line corresponds to a risk neutral student. For risk-preferring students, students with a risk preference of three will "guess" if their probability of getting the question right is at least 3%. The most risk averse student would not respond unless they had at least a 55% chance of getting the question correct.

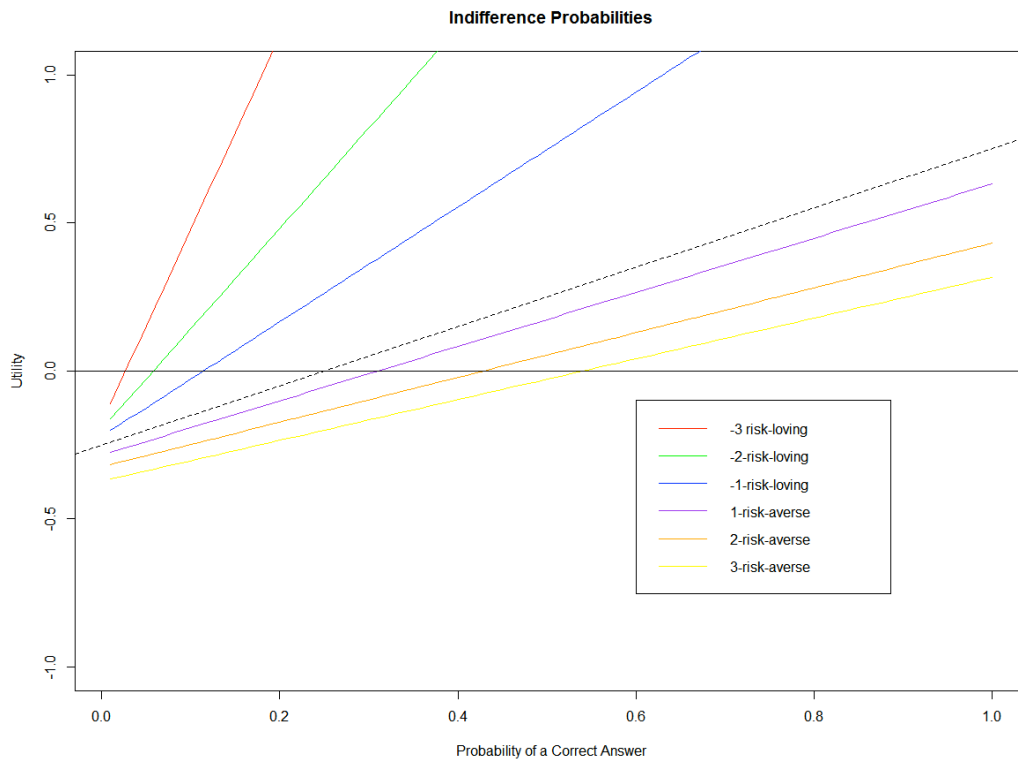


Figure 2 Indifference Probabilities and Utility

4 SIMULATIONS

A hundred bootstrapped simulations were run to better estimate the effects of strategic omission. Repeated simulations yields the omission rates plots below. On average, a risk-neutral simulation yields an omissions rate of 18%. In the risk-averse case, this omission rate jumps up to approximately 30%. Sum scores change relatively little with only a two percentage point difference in exam performance (Figure 3).

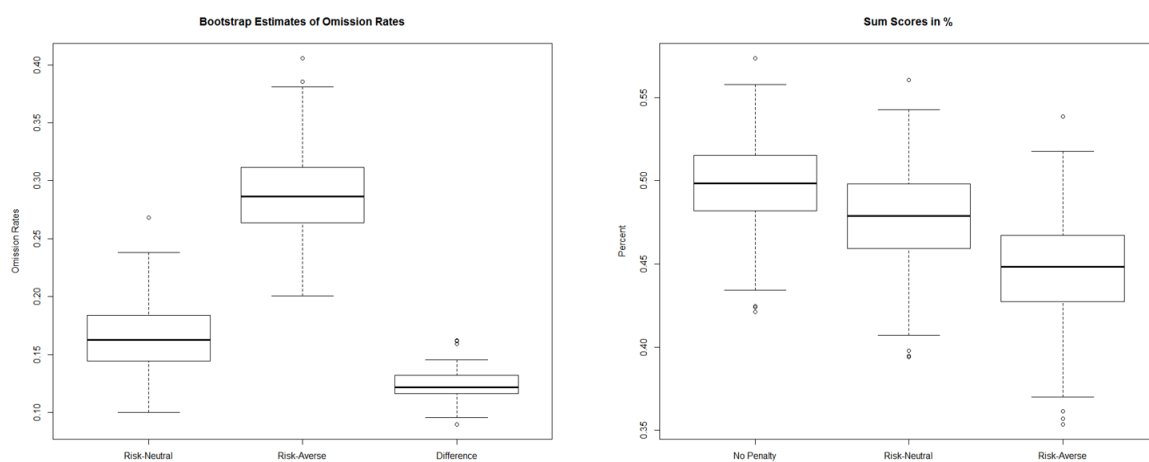


Figure 3 Bootstrapped Estimates of Sum Scores

4.1 Ability Measurement Error

By introducing a penalty, it introduces a large region where low ability individuals will not attempt certain items. This makes distinguishing between low ability people and very low ability people extremely difficult. From a maximum likelihood estimation perspective, this means that for each item there is a portion of the information curve where the estimate is completely flat. An illustration of that fact can be seen in Figure 4.

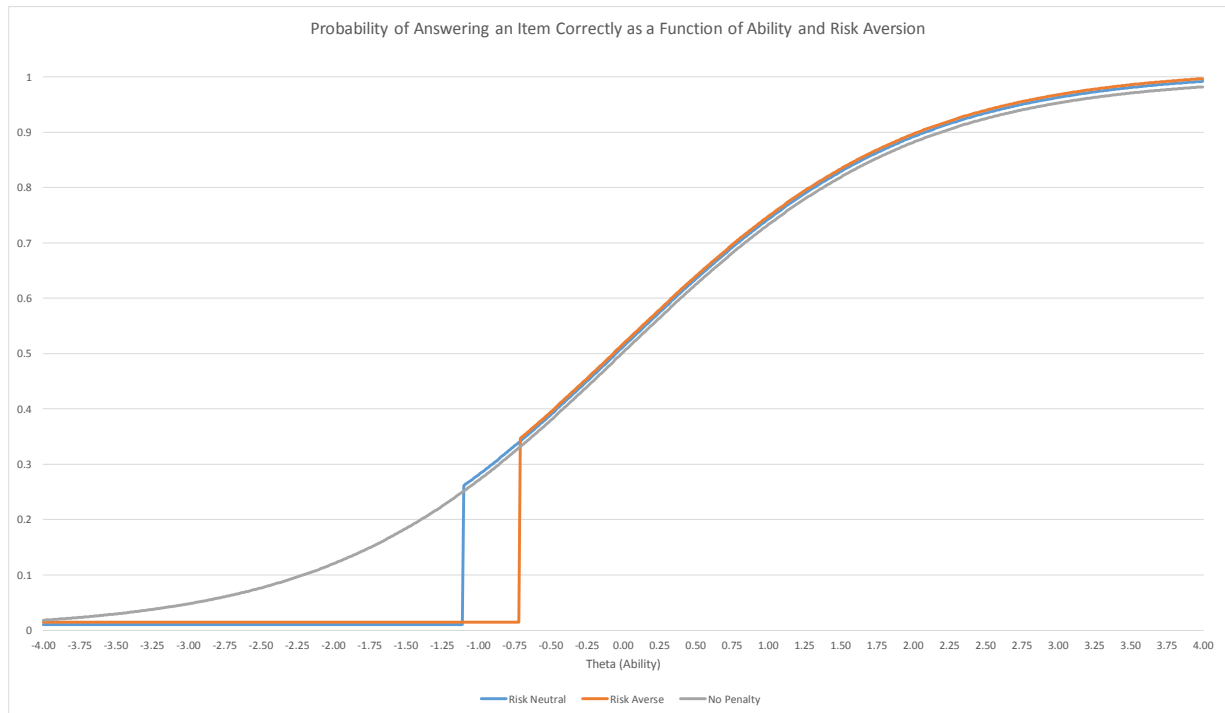


Figure 4 Probability of Answering an Item Correctly as a Function of Ability and Risk Aversion

We also recover individual ability estimates using a Rasch model and maximum likelihood. Estimates of these data yield unbiased estimates of an individual's ability (See Figure 3). The mean absolute deviations of theta increases as the penalty function is introduced and as the risk-aversion increases. As such, the amount of error in ability measurements is nearly twice as large for a risk-averse population than if there were no penalties enacted on the same population of students (See Figure 5).

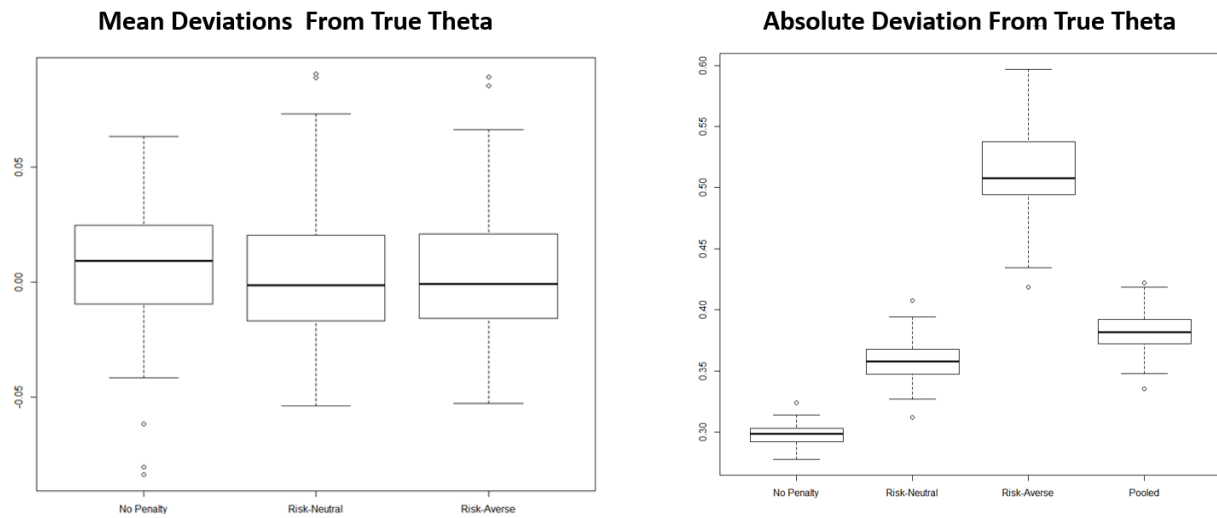


Figure 5 Absolute Mean Deviations of Ability

4.2 Reliability

So the fundamental question is why are these penalty functions used if it increases non-response rates and seems to introduce these potential claims of bias. One possible explanation is that improves measures of reliability. We compute the reliability of the generated exams using Cronbach's alpha (Cronbach, 1951). The boxplots below show that reliability increases if students are given an incentive to omit incorrect answers. This effect still holds even if one assumes heterogeneity of risk tolerance amongst users (See Figure 3). In effect, what happens is that users who have relatively low likelihood of getting an item correct through random guessing gets their answer compressed to zero in response to a penalty. This omission, in turn, increases the reliability of an exam.

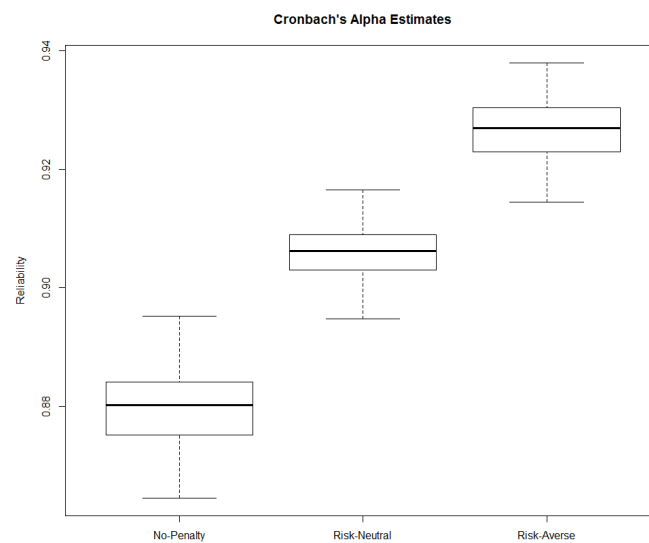


Figure 6 Cronbach's Alpha (Reliability)

5 DISCUSSION

From a reliability perspective, penalizing exams has some benefits. Introducing penalties tends to increase the reliability of the exams. This increase in reliability comes at the cost of certain measures becoming noisy. Further, if there's heterogeneity of risk aversion, it's possible that the rank ordering of students could jump noticeably when an exam switches from a penalty function to an exam without a penalty function. Strategic omission makes generating distinctions between the bottom-half of the distribution very difficult. To the extent that an exam is concerned with generating a precise estimate of ability, utilizing a penalty function is ill-advised.

The only cases where a guessing penalty could make sense are when risk tolerance is a parameter that is also being trained. For instance this type of penalty function could be useful when training actuaries, financial investors, or stockbrokers. The rationale for this is that their score would be both a composition of their true ability and their risk tolerance.

5.1 Implications for Learning Analytics and Platform Design

This work suggests that penalties should not be used for assessment purposes. If individuals are penalized for wrong answers, then risk-averse users will strategically omit more responses than risk-tolerant users. In turn, this means that learning platforms would direct risk-averse users into more remedial content than similar ability students who are risk-neutral. To the extent that these populations are underserved groups (females, underrepresented minorities, and low socioeconomic status), embedding penalties for random guessing could deter these groups from interacting with the platform and replicate existing inequalities. Further, our simulations suggest that guessing penalties may make it more difficult for learning platforms to distinguish between users in the lower end of the ability distribution. These are often the groups that are of focal interest to learning analytics researchers and policy makers.

Many learning platforms reward users with points or badges for engaging with the platform and penalize users for using built-in hint generation features. Removing penalties from these contexts seem like a natural decision. Generally, these penalties should be removed when items are being used as part of a formative assessment.

If random guessing penalties are to be used in a summative assessment, there are approaches that mitigate the performance bias between risk-averse and risk neutral users. One of the design choices is to allow students to respond to multiple items before submitting a response for grading. This will allow rational agents to hedge their responses and makes risk-averse users more likely to respond so long as their knowledge is truly better than random guessing.

6 ACKNOWLEDGEMENTS

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant (#R305B140009). We also acknowledge thoughtful reviews from the conference organizers.

REFERENCES

- Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*. Retrieved from <http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2013.1776>
- Budescu, D., & Bar-Hillel, M. (1993). To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. *Journal of Educational Measurement*, 30(4), 277–291. <https://doi.org/10.1111/j.1745-3984.1993.tb00427.x>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Gandelman, N., & Hernández-Murillo, R. (2014). Risk Aversion at the Country Level. *FRB of St. Louis Working* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2396103
- Kirkwood, C. (1997). Notes on attitude toward risk taking and the exponential utility function. *Department of Management, Arizona State University* Retrieved from <http://www.public.asu.edu/~kirkwood/DASstuff/refs/risk.pdf>
- O'Rourke, E., Haimovitz, K., & Ballweber, C. (2014). Brain points: a growth mindset incentive structure boosts persistence in an educational game. *Proceedings of The*. Retrieved from <http://dl.acm.org/citation.cfm?id=2557157>

Analytics as a Team Sport: Using Cloud-Based Tools to Support Data-Intensive Research-Practice Partnerships

Andrew Krumm
Digital Promise
akrumm@digitalpromise.org

Jeremy Roschelle
Digital Promise
jroschelle@digitalpromise.org

Particia Schank
Digital Promise
pschank@digitalpromise.org

ABSTRACT: This workshop will provide participants with the opportunity to develop skills needed to lead and support data-intensive research-practice partnerships. Using insights gleaned from multiple cases, workshop participants will engage in whole- and small-group activities around setting up a partnership and conducting collaborative data analyses using cloud-based tools that have been integrated into a free and open source set of services referred to as TeamSpace. The combination of a grounded data-intensive improvement process and an easy-to-launch set of analysis tools, will put participants on a path toward organizing more and more of their learning analytics work as a “team sport.”

Keywords: Research-practice partnerships, cloud-based analytical tools, data-intensive improvement

1 BACKGROUND

To improve teaching and learning in schools and universities, educators and staff often need to combine data sources and conduct analyses that can be technically challenging to accomplish on their own. For example, analyses that go beyond attendance and gradebook data to include traces of students’ learning events from digital learning environments have been shown to be particularly beneficial to helping practitioners understand learning processes and develop accompanying change ideas (e.g., Bowers et al., 2016). A key resource in helping practitioners achieve new insights into learning and develop better change ideas are structured collaborations with researchers (e.g., Penuel, & Gallagher, 2017).

Specialized expertise is often needed (1) to make sense of data from digital learning environments alongside other school records (e.g., test scores and grades) and (2) to interpret an analysis in terms of relevant learning theories. The expertise of educators is critical to interpreting an analysis with the constraints of a local context as well as understanding possibilities for improving local learning opportunities. Research-practice partnerships offer a way to bring the capabilities of both researchers and practitioners together to solve challenging problems. However, simply bringing researchers and practitioners together will not automatically lead to desired benefits. To support effective collaboration, partnerships between researchers and practitioners need (1) explicit processes to follow and (2) tools to support common workflows (Author, 2018). Therefore, in order to make the most of data collected by digital learning environments to solve pressing problems, researchers and practitioners need to work together using common tools and follow effective processes (e.g., improvement science).

The purpose of this workshop is to introduce participants to ways of developing and supporting data-intensive research-practice partnerships that are geared toward solving local problems while also supporting broader knowledge building (Penuel & Gallagher, 2017). Participants will work through the phases of an overall process referred to as collaborative data-intensive improvement (CDI), which came about as the result of a multi-year, multi-project research agenda that investigated how researchers and practitioners can work together to use data-intensive research techniques to support continuous improvement efforts (Author, 2018). CDI involves five phases: Phase I is about setting up a partnership, and includes jointly clarifying the problems to be solved and defining aims for the partnership. Phase II involves developing an overarching theory for how the partnership will reach its aims. Phase III is where a partnership engages in exploratory data-intensive analyses—combining data from digital learning environments and administrative data systems. To support partnerships engage in Phase III, we developed TeamSpace—an integrated set of cloud-based services that facilitate researchers, practitioners, and technology vendors in collaboratively analyzing data (see Figure 1). Phase IV is where insights from data-intensive analyses get translated into change ideas through iterative, collaborative design. Lastly, Phase V is where members of a partnership test out change ideas in real learning environments and improve upon the change ideas over time.

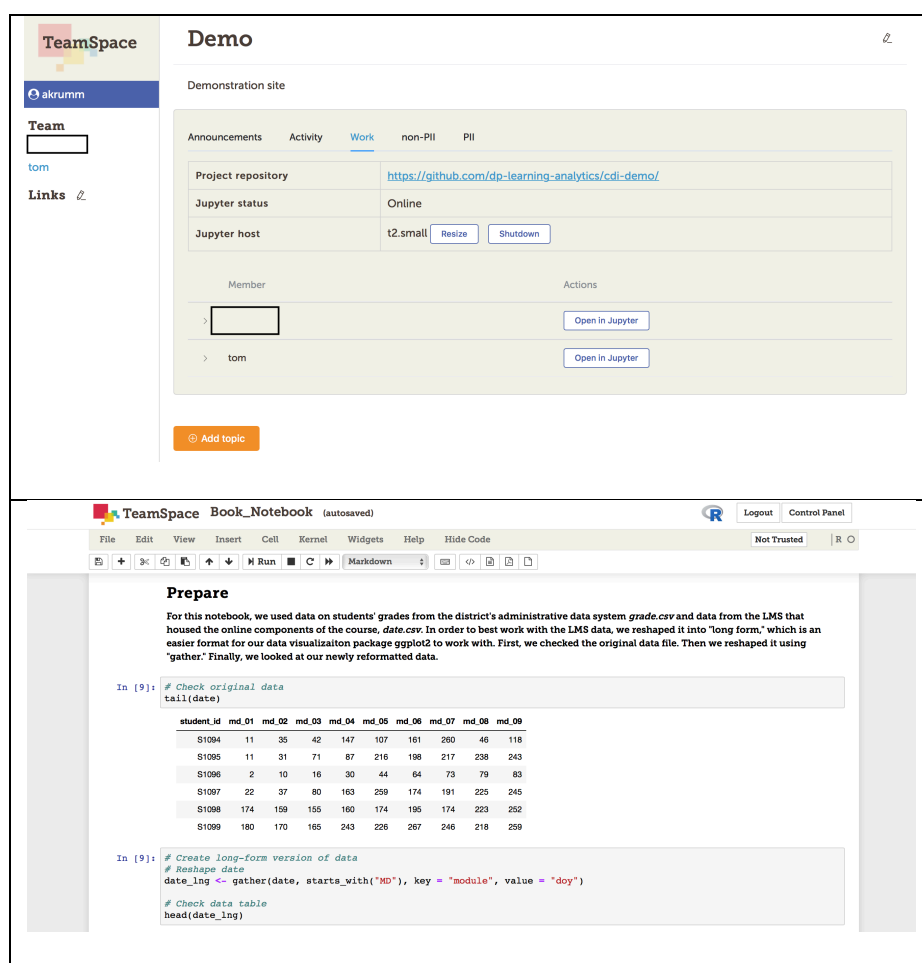


Figure 1: TeamSpace

In developing TeamSpace, we reflected on our own partnerships and explored what the best data science teams from multiple industries do. We identified that the best data science teams: (1)

provide password protected, role-based, and auditable access to data; (2) reduce the movement of data between partners and analysts; (3) move the data analysis engine as close to the data as possible; (4) use version control software and jointly scrutinize data analysis scripts. We created TeamSpace to address meet these four factors using widely-available, open source software (e.g., Jupyter Notebooks running R and Python kernels). Individual TeamSpaces are easy to set up and take down—they can be set up and destroyed in minutes. TeamSpaces are intended to democratize access to data science tools by integrating best-of-breed tools within a lightweight user interface.

2 ORGANIZATION

This event will be organized as a half-day workshop for 15-20 participants. Based on the tools that we want to use with participants (e.g., TeamSpace), we will ask participants to pre-apply so that we can set up individual spaces. Having participants pre-apply will also allow us to better understand the problems of practice that participants are working to solve. We will administer a short survey to participants after they have applied and been accepted so that we can understand the skills and interests of participants in addition to the potential problems of practice that they want to solve. Understanding skills, interests, and problems will help us tailor workshop activities.

The core elements of this workshop are collaborative data-intensive improvement processes, TeamSpace instances, and a common dataset. Each of the above CDI Phases has accompanying templates and worksheets that will be provided to participants. The workshop will begin by having participants work in small groups to discuss individual problems of practice that they would like to explore at their home institutions in line with CDI Phase I. Following this, workshop organizers will provide participants with examples of Phase II outcomes and processes that they can use in their own partnerships. To provide participants with hands-on experience conducting collaborative data analyses, we will use the Open University Learning Analytics dataset (https://analyse.kmi.open.ac.uk/open_dataset) and have small groups engage in scaffolded exploratory data analyses.

Table 1: Proposed schedule.

Time	Workshop activity
8:30	Introduction and workshop overview
9:00	Whole group: Overview of CDI
9:25	Small group: Clarify problems of practice (CDI Phase I)
9:50	Whole group: Phase II examples and introduce TeamSpace
10:15	Break
10:30	Small group: Collaborative analyses in TeamSpace (Phase III)
11:30	Whole group: Share out analyses and group brainstorm potential change ideas for own partnerships (Phases IV & V)
12:15-12:30	Closing

After engaging in exploratory data analyses, participants will share out what they learned through the process as well as any data products that they created. Workshop organizers will then provide

an overview of the follow up CDI Phases around co-designing and testing change ideas, i.e., Phases IV and V, respectively.

We will create a webpage at to advertise the workshop along with how to apply to participate. Participants will need individual laptops and workshop organizers will need an LCD projector to share our screens and present slides.

3 OUTCOMES

Though this workshop, participants will better understand the processes and necessary tools for engaging in collaborative data analyses. In working toward this objective, participants will learn about collaborative data-intensive improvement processes and tools as well as engage in collaborative analyses using TeamSpace. As outlined above, participants will engage in hands-on activities that will help them make connections between collaborative data intensive improvement processes and tools and their own context. Outcomes of the workshop will be disseminated in collaboration with our communications department, which has an extensive track record in using multiple digital content approaches to share stories at the intersection of technology and learning. Along with using our communications team, we will promote and disseminate the outcomes of the workshop through our social networks using Twitter and Facebook.

REFERENCES

Author (2018)

- Bowers, A.J., Krumm, A. E., Feng, & M., Podkul, T. (April 2016). *Building a data analytics partnership to inform school leadership evidence-based improvement cycles*. Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.
- Penuel, W. R., & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Model-Based Analysis of Gaze Data During Video Lectures

Hiroaki Kawashima, Kousuke Ueki, Kei Shimonishi

Kyoto University, Japan

{kawashima, shimonishi}@i.kyoto-u.ac.jp, kesuuko6123@gmail.com

ABSTRACT: Learners change their behavior-mode dynamically during video lectures, for example, “follow a lecturer’s guide (speech and pointers),” “look ahead of spoken parts and actively check slide content,” and “roughly browse a slide.” We propose a model-based analysis to decompose viewers’ gaze-behavior patterns into such modes during video lectures. The method assumes three modes of viewers and formulate their gaze behavior as a probabilistic generative model of the three-mode mixture. Our experiments demonstrate that the method is able to decompose gaze patterns into component distributions, which highlight important time/order-dependent regions and would enable personalized feedback.

Keywords: Eye-tracking, Video lectures, Probabilistic gaze-behavior model

1. INTRODUCTION

Learners behavior analysis during lectures at scale has a great potential to enhance learning experience. Once the detailed states are estimated, a variety of personalized feedback can be designed such as visualizing learners' own states with some indicators, generating follow-up questions, and summarizing content. Clickstream on MOOCs and e-book systems have therefore been actively studied to analyze learners' behavior during lectures, for example, how much students follow a lecture and when they tend to dropout (Kim et al., 2014; Shimada, Taniguchi, Okubo, Konomi, & Ogata, 2018).

However, since clickstream logs only contain users' intentional activities (e.g., page jumps, marks, and comments), it is not possible to know the details of how each user followed content in a slide, reacted to the lecture's speech and actions (e.g., pointing), and affected by the slide design (e.g., saliency of layout). To overcome this limitation, the analysis of eye-tracking logs has recently attracted attention of learning analytics researchers (Mangaroska, Sharma, Giannakos, Tr  tteberg, & Dillenbourg, 2018; Sharma, Jermann, & Dillenbourg, 2014). Thanks to low-price eye trackers (Nguyen & Liu, 2016; Rodrigue, Son, Giesbrecht, Turk, & H  llerer, 2015) and video-camera based gaze estimation using computer vision techniques (Zhang, Sugano, Fritz, & Bulling, 2015), the assumption that a large number of gaze data can be collected in lectures and exercises is becoming realistic.

The analysis of gaze data in video lectures provide a deeper insight into which consisting regions attracted/confused viewers (Nguyen & Liu, 2016) and how much learners followed a lecturer's speech (Sharma et al., 2014) even in one slide. In addition, spatio-temporal gaze patterns are expected to be useful cues to predict learners' states in multiple aspects: attention levels, knowledge/performance, and general attitude toward lectures.

In spite of the importance of gaze patterns in a lecture, it is often difficult to infer learners' internal states due to a large variety of learners' attentional modes. For example, learners do not always follow

lecturer's spoken words and pointers but often scan lecture content by themselves or just roughly browse the content to grasp keywords. Due to the mixed nature of these variety of viewing styles during video lectures, observed gaze data have large variance and cannot be easily interpreted.

To address the interpretation problem of gaze behavior so as to infer learners' states in video lectures, this paper proposes a method to decompose viewers' gaze patterns into three *modes (components)*, namely “roughly grasp slide content,” “actively follow slide content,” and “follow a lecturer's guide (speech and pointers).” Once a collection of gaze data is given for one slide page, the model finds parameters corresponding to each of the three modes through a machine learning technique. Since this can be seen as a clustering method, a submodel of each mode is expected to contain information of a specific aspect, which would enable inference of gaze behavior.

2. MULTI-MODE GAZE-BEHAVIOR MODEL

We consider one model for each slide and assume a model is trained by a collection of gaze data obtained in the period that the slide appears on a screen. Since relative timestamps from the beginning of a video can be obtained, gaze data of multiple viewers can be obtained asynchronously and later merged on a single time axis.

2.1 Gaze Data

Each of gaze data is assumed to be a sequence of regions on a screen including slide content. Assuming that raw data of an eye-tracker are densely-sampled temporal sequences of x-y coordinate points on a screen, each sequence is firstly segmented into saccadic movements and fixations based on the velocity of eye movements (Salvucci & Goldberg, 2000). Then, one region ID is assigned to each of fixation segments. Through these steps, a raw gaze sequence is converted into a sequence of area-of-interest (AOI); we here use the term *gaze regions* for the AOIs.

Suppose that R_1, \dots, R_N are regions in one slide. We use notation $r_k^{(v)} \in \{R_1, \dots, R_N\}$ for a gaze region (AOI) at time step k of viewer v . Note that k is an index of AOI switches, i.e., r_k is the k -th region that the viewer v fixated his/her gaze, and $r_k \neq r_{k+1}$. The number of AOI switches is different for each viewer. Therefore, a gaze sequence of viewer v can be written as $r_{\text{seq}}^{(v)} = \{r_0^{(v)}, \dots, r_{K_v}^{(v)}\}$ using the length of sequence, K_v , indexed by v . In addition, we use $t_k^{(v)}$ to denote the physical timestamp at time step k of viewer v . Unless otherwise noted, we use term *time* or *time step* for AOI switches $k = 1, \dots, K_v$ and *media time* for physical timestamps whose origin is the beginning of the slide.

2.2 Viewer's Modes and Submodels

Since we aim at constructing a minimal model, the following three modes (components) are assumed.

Mode 0: Attracted by salient regions such as high contrast or important terms.

Mode 1: Follow slide content by considering the meaning of content information.

Mode 2: Follow a lecturer's guide such as spoken words and pointers.

Corresponding to the viewer's modes introduced above, we consider submodels of gaze behavior. Each submodel describes the probability distribution of regions on which gaze fixation is attracted. Let

$m_k^{(v)}$ be the mode of viewer v at time step k . Remind that the model is viewer independent; therefore, the viewer ID, v , is dropped in what follows. Then each region distribution is formulated as

Mode 0 (base distribution): $P(r_k = R_i | m_k = 0) = a_i$,

Mode 1 (region order): $P(r_k = R_j | r_{k-1} = R_i, m_k = 1) = b_{ij}$,

Mode 2 (lecture's guide): $P(r_k = R_i | t_k = t, m_k = 2) = c_{it}$.

The first submodel (mode $m = 0$) describes the “base” distribution of regions (this is the reason number zero is used for this mode). While classical saliency models utilize the degree of contrast of each pixel to surrounding regions, we here consider a data-driven saliency model.

The submodel for mode $m = 1$ is expected to describe meaningful connection between regions. This model considers the first-order Markov property in each gaze sequence; that is, the probability distribution of gaze region r_k is assumed to be dependent on the previous gaze region r_{k-1} . Since viewers' scan path may depend on longer context, we focus on this simplest submodel to investigate the basic property of the mixture of submodels.

The concept of mode $m = 2$ is related to the *with-me-ness* proposed in Sharma et al. (2014) in terms that the corresponding submodel tries to describe specific spatio-temporal points that a lecturer attracts gaze of learners. In this submodel, with the dependency on media time t_k , regions that correspond to a lecturers' spoken word and pointed parts are expected to have higher probability. Note that media time t is discretized by giving a time-grid size (one second is used in our experiments).

2.3 Proposed Model

Viewer's gaze sequence is assumed to be affected by the three component submodels described in Section 2.2, where the influence ratio of the submodels may change dynamically:

$$P(r_k | \cdot) = P(m_k = 0 | \cdot)P(r_k | m_k = 0) + P(m_k = 1 | \cdot)P(r_k | r_{k-1}, m_k = 1) + P(m_k = 2 | \cdot)P(r_k | t_k, m_k = 2),$$

where (\cdot) denotes the previous gaze regions $r_{k-1}, r_{k-2}, \dots, r_1$ and timestamps t_k, t_{k-1}, \dots, t_1 . Note that conditional independence is assumed in each of submodels. The probability $P(m_k = m | \cdot)$ describes the influence ratio of submodel $m \in \{0, 1, 2\}$, whose sum is $\sum_{m=0}^2 P(m_k = m | \cdot) = 1$.

An example of typical situations described by this model is as follows: When a viewer follows the lecturer's guide (spoken words and pointers) at time k , the value of $P(m_k = 2 | \cdot)$ becomes higher than other modes. Meanwhile, when a viewer follows the slide content actively by ignoring the lecturer's guide, $P(m_k = 1 | \cdot)$ becomes higher. In a time period of mind wandering, $P(m_k = 0 | \cdot)$ get higher due to not following either the content or the lecturer.

3 EXPERIMENTAL RESULT

3.1 Experimental setting

11 university students were recruited to conduct a lab-setting research to address the question: (Q) *What kind of information can be extracted using the model-based decomposition analysis?*

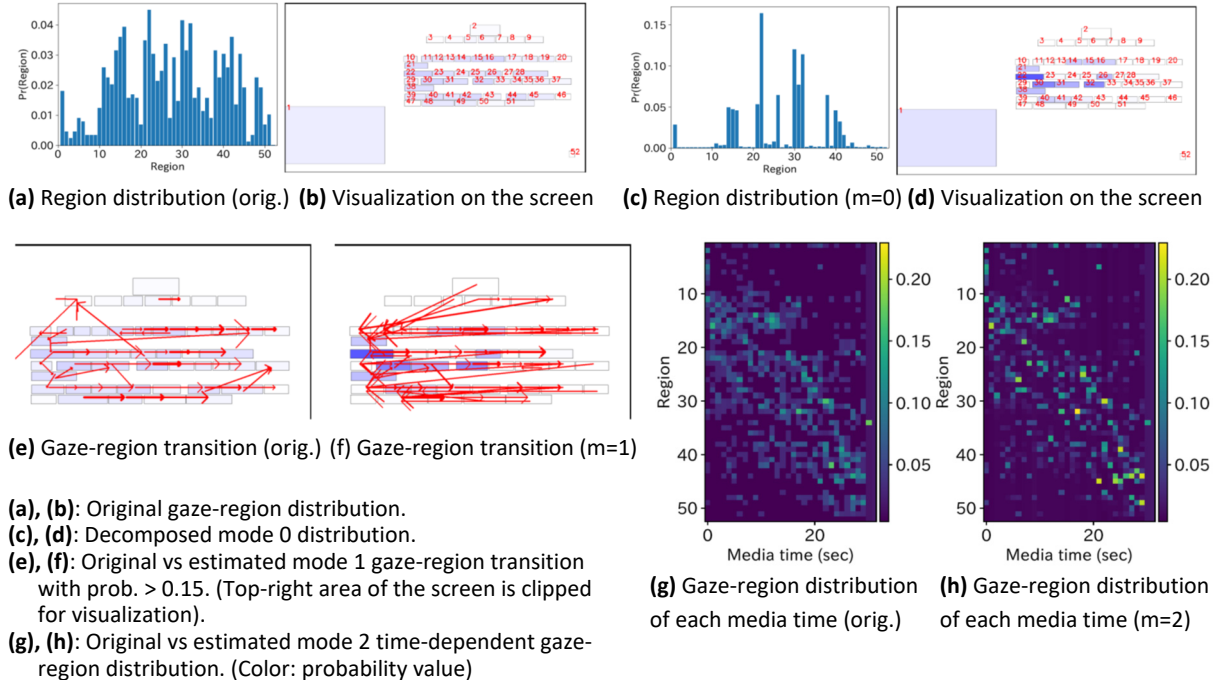


Figure 1: Estimated submodel parameters of gaze

Video-viewing and post-test tasks were assigned to each of the participants. The video used in this study was about “Design of Mind.” The length of the video was about 10 minutes which consists of 20 slides. The screen of the video had three regions: Lecturer (bottom left), additional information (top left), and a slide (right). Post-test questions were prepared to make each participant concentrate enough on the video content. Tobii X120 eye tracker was used to measure participants’ gaze points on a screen with 60 Hz sampling. Each participant was asked to sit in front of the screen where chin rest was used to reduce measurement noise as much as possible.

Using the obtained gaze data, the proposed model introduced in Section 2 was trained via the expectation-maximization (EM) algorithm similar to probabilistic latent semantic analysis (pLSA) (Hofmann, 1999), also known as topic models. While both “topic” in pLSA and “mode” in our model are both discrete latent variables and have some similar structure in their training step, modes in our model are order- or time-dependent and the number of latent variables are smaller than pLSA (i.e., the number of modes is three in our model). Therefore, the training step is not strongly affected by initial values compared to pLSA, which usually requires annealing techniques to avoid local minimum, and thus a standard EM algorithm was enough for our model training.

3.2 Estimated model parameters

To address the question (Q), we visualize the estimated parameters and compare them with original distributions before decomposition. To see the basic characteristics of the estimated parameters, in this qualitative evaluation, we pick only a specific period when a summary slide (#20) was displayed. Nevertheless, we have found that most of the results here also apply to other slides and lecture videos. Figure 1 shows the original gaze-region distribution and estimated parameters of the submodels.

Mode 0 (base distribution): Figure 1 (a) and (b) show the normalized frequency distribution of gaze regions in the slide #20 and its visualization on the screen, respectively (higher probability region is

colored by darker blue)¹; meanwhile, Figure 1 (c) and (d) show the estimated mode 0 distribution (i.e., a_i). Compared to the original distribution, only a limited number of regions have higher probability in the estimated mode 0 distribution. It is interesting to note that, in addition to the lecturer region (R_1 : the bottom left area), some important technical terms, e.g., “cultural psychology” (R_{15}, R_{16}), “ERP study” (R_{22}, R_{30}), and “fMRI” (R_{32}), in the slide are highlighted.

Mode 1 (region order): To visualize the estimated mode 1 parameter b_{ij} (transition probability from region R_i to R_j), it is not informative to directly show the transition probability matrix because of the large number of regions. Instead, we here depict transition probability higher than 0.15 as red arrows on the slide as shown in Figure 1 (e) and (f). These arrows show which regions will be high likely to be looked at after the gaze at each region.

In the both figures, horizontal left-to-right transition patterns are dominant as can be seen in standard book reading. However, the number of the arrows are much larger and back-track patterns are found more in the estimated mode 1 parameters (Figure 1 (f)) compared to the original transition probabilities (Figure 1 (e)). This indicates that the structure of gaze-region transition is decomposed from other structures (mode 0 and mode 2) and can be highlighted in the proposed model, while this transition structure is buried in other transition patterns in the original data without decomposition. The extracted transition information may also be useful for designing/optimizing content itself (e.g., too many back-track patterns may suggest the complexity of the content).

Mode 2 (media-time-dependent structure): Figure (g) shows the gaze-region distributions of each media time whose origin is the beginning time of the slide #20. The distribution is normalized in each media-time grid (one second). It can be seen that higher probability regions change over time from smaller to larger number regions, which mostly corresponds to top-to-bottom horizontal scan of the slide. Figure 1 (h) shows the result of the estimated mode 2 (i.e., c_{it}). This is also a media-time-dependent distribution of gaze regions similar to Figure (g) but is considered as a decomposed version from other modes.

At each time grid, the estimated mode 2 distribution has higher peaks in Figure 1 (h) while they are smoothed in Figure 1 (g). This can be considered that the participants were not only focus on regions related to the lecturer’s guide but also scans other regions independent of the lecturer’s behavior. On the other hand, the estimated mode-2 parameters emphasize which regions are related to the lecturer’s guide (i.e., spoken words). In fact, some most important keywords in this lecture such as “mind” (R_{44}) and “design” (R_{45}) can be detected clearly in the last part, after media time 20 (sec), in Figure 1 (h). This indicate that, once a large number of gaze data are collected, important regions at each media-time point can be estimated by the proposed model thanks to its clustering property.

¹ The original content is not displayed in this paper and only region boundaries are depicted.

4 DISCUSSIONS

4.1 Analysis of inter-subject variability using inferred modes

Each learner has different strategy and switches mode at a different pace. To address such inter-subject variability, the proposed model provides a probabilistic inference framework. Given an observed gaze data $r_{\text{seq}}^{(v)}$ of viewer v , the *mode posterior* $P(m_k = m | r_{\text{seq}}^{(v)}) = \gamma_k^{(v)}(m)$ ($m = 0, 1, 2; k = 0, 1, \dots, K_v$) can be obtained with a similar procedure as the E-step in the training phase. Note here that gaze data $r_{\text{seq}}^{(v)}$ can be either one of training data or newly obtained data after the training phase; new viewer may also be accepted if the model is trained by a large dataset and has been generalized enough.

A sequence of inferred posterior can be considered as the pattern of the influence of each submodel. For example, the value of $\gamma_k(2)$ is expected to be large for time k when the viewer focuses on the lecturer's talk, as explained in the last paragraph of Section 2. Therefore, one interesting future research is to analyze subject variability based on the pattern of mode-posterior sequences $\gamma_0^{(v)}, \dots, \gamma_{K_v}^{(v)}$ from each viewer v to see the relation between his/her performance.

4.2 How to decide the number of modes?

In addition to the base model, $m = 0$, we considered two submodels, $m = 1, 2$, each of which is conditioned by the previous gaze region and current media time, respectively. While we believe this is a minimal model design to decompose viewers' behaviour, the appropriate number of modes should be validated with larger dataset in future. In fact, the number of modes is not necessarily be "three" as there is no strong theoretical background behind this decision. We here introduce two possible approaches to decide the number of modes: data-driven approach and design-based approach.

Data-driven approach considers a trade-off between prediction accuracy and the size of model parameters using information criteria (e.g., AIC and BIC) or cross-validation with a large dataset. In the design-based approach, we can introduce new modes corresponding to available modalities or features extracted from multimodal data. Design-based approach is expected to be suitable to obtain more interpretable modes compared to data-driven approach.

4.3 The use of other modalities

To infer learners' state, predict their performance, and design appropriate feedback, we can use two types of multimodal information with gaze data: modalities observed by a learner (input modalities to a learner) and modalities observable from a learner (output modalities from a learner).

Content information available from video is a typical example of input modalities to a learner. While our model considers the influence of video context implicitly by the dependency on media time t_k in the submodel 2, detailed content information (e.g., lecturers' spoken words, face movements, pointing actions) can also be used explicitly. For example, we found that there is a strong synchronization between a lecturer's head direction and viewers' gaze, which would contribute to predict learners' performance.

Output modalities from a learner, such as clickstreams and playback speed, can also be analyzed together with gaze data by considering the granularity of time. When click events happen inside the period of a slide, this information can be analyzed together with gaze data within the period the slide is presented. Meanwhile, observed gaze (or inferred mode) information should be summarized for each of slides to be integrated with inter-slide information such as page jumps between slides; in this case, each slide transition is considered as a time step.

Our model can be extended to utilize additional input and output modalities since our generative model presented in Section 2.2 basically has the form $P(\text{output}|\text{input}, \text{state})$. By adding input modalities as “input” conditions, and output modalities as “output” observations, more accurate inference of learners’ internal “states” (e.g., modes) would be possible. When the interpretation of modes is necessary, one can introduce modality-specific submodels, which depends on one of input modalities (see also Section 4.2).

4.4 Toward real-course settings

While a lab-setting research was conducted as an early-stage experiment in this paper, it is now realistic to introduce gaze measurements in real-course settings thanks to recent advances in eye tracking technology. As explained in the introduction, not only less-expensive eye trackers but webcam-based methods are available, and they are improving year by year using computer vision and deep learning techniques (Zhang et al., 2015). Note that our analysis is based on a probabilistic clustering technique (Hofmann, 1999), which are widely used large online data (e.g., recommender systems in e-commerce). Therefore, the proposed method has a potential to be applied to large-scale gaze data obtained through online e-learning platforms and to extract common and personal key features useful for designing personalized feedbacks.

4.5 Design of personal feedback

Several types of personalized feedback can be designed based on the proposed analysis. For example, (1) visualization of the summary of gaze data or estimated internal states can be given to learners to support their self-reflection; and (2) content design itself can be adapted to each of learners.

Once the model is trained with large data, we may obtain a set of “standard” viewers’ behaviour patterns. When a newly observed gaze behaviour does not fall into one of such standard patterns, it would be able to detect an anomaly period of the learner. This information can be used for both the design (1) and (2). For example, by detecting regions or slides that the viewer had anomaly behaviour, corresponding technical terms or topics can be listed as content that the viewer possibly had some trouble. We may also find relation between gaze behaviour and performance, which gives us some insight to improve learning materials, such as saliency, layout, and content density both in time and space in a video.

5 CONCLUSION

This paper proposed a model-based gaze analysis to decompose viewer's gaze behavior into three modes by introducing a novel probabilistic mixture model. The effectiveness of the model was demonstrated by visualizing parameters obtained through each of the submodels. While the results show a potential to interpret gaze behavior of learners during video lectures and to design a variety

of feedbacks (e.g., keywords and regions) to the learners, the method and the results are still in a preliminary stage and have many limitations. For example, duration of gaze fixation is not exploited, and the situation of the experiments was artificial and needs to be evaluated in actual courses with performance tests. Validation of the proposed model such as the number of submodels (modes) and their degree of freedoms also need to be investigated with a larger size of data. Our future work includes a comparative study of the effectiveness of the proposed method by summarizing content based on the highlighted region information and by introducing personal feedbacks.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP16H06304 and JST PRESTO Grant Number JPMJPR14D1.

REFERENCES

- Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Paper presented at the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 50–57. <http://dx.doi.org/10.1145/312624.312649>
- Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller., R. C. (2014). Understanding In-Video Dropouts and Interaction Peaks in Online Lecture Videos. *Paper presented at the International Conference on Learning @ Scale (L@S)*, 31–40. <http://dx.doi.org/10.1145/2556325.2566237>
- Mangaroska, K., Sharma, K., Giannakos, M., Trætteberg, H. & Dillenbourg, P. (2018). Gaze Insights into Debugging Behavior Using Learner-Centred Analysis. *Paper presented at the ACM International Conference on Learning Analytics & Knowledge (LAK)*, 350–359. <http://dx.doi.org/10.1145/3170358.3170386>
- Nguyen, C. & Liu, F. (2016). Gaze-based Notetaking for Learning from Lecture Videos. *Paper presented at CHI Conference on Human Factors in Computing Systems*, 2093–2097. <http://dx.doi.org/10.1145/2858036.2858137>
- Rodrigue, M., Son, J., Giesbrecht, B., Turk, M., & Höllerer, T. (2015). Spatio-Temporal Detection of Divided Attention in Reading Applications Using EEG and Eye Tracking. *Paper presented at the International Conference on Intelligent User Interfaces (IUI)*, 121–125. <http://dx.doi.org/10.1145/2678025.2701382>
- Salvucci, D. D. & Goldberg, J. H. (2000). Identifying Fixations and Saccades in Eye-Tracking Protocols. *Paper presented at the Symposium on Eye Tracking Research & Applications (ETRA)*, 71–78. <http://dx.doi.org/10.1145/355017.355028>
- Sharma, K., Jermann, P., & Dillenbourg, P. (2014). “With-me-ness”: A Gaze-Measure for Students’ Attention in MOOCs. *Paper presented at International Conference of the Learning Sciences (ICLS)*, 1017–1022.
- Shimada, A., Taniguchi, Y., Okubo, F., Konomi, S., & Ogata, H. (2018). Online Change Detection for Monitoring Individual Student Behavior via Clickstream Data on E-book System. *Paper presented at the ACM International Conference on Learning Analytics & Knowledge (LAK)*, 446–450. <http://dx.doi.org/10.1145/3170358.3170412>
- Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. *Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520. <http://dx.doi.org/10.1109/CVPR.2015.7299081>

An Integrated Framework for Content Analysis of Learning Videos and Learner-Generated Artefacts

Yassin Taskin, Tobias Hecking, H. Ulrich Hoppe

University of Duisburg-Essen, Germany

{taskin, hecking, hoppe}@collide.info

ABSTRACT: This paper presents the implementation of an integrated framework for content analysis of learning videos and an evaluation of the framework through the analysis of the content wise relation of learning videos and learner-generated wiki articles. The goal of the video analysis is the transcription and temporal segmentation of the video content. Temporal cut points identified in the videos should correspond to changes in presentation from the instructor and indicate thematic changes. The transcribed content of the video segments generated via these cut points can then be compared to learner-generated artefacts, to determine relevance between segments and learner-generated content. For this purpose, a segmentation method based on the encoding of video files was developed. For the evaluated dataset the relation of the extracted text segments and learner-generated wiki articles was analyzed using Network-Text-Analysis. The different measures were evaluated to compare which one is best suited to identify video segments that were important for the formulation of the wiki articles.

Keywords: Learning Analytics, Video-based Learning, Video Segmentation, Network-Text-Analysis

1 INTRODUCTION

Video-based learning has gained much popularity over the recent years (Yousef, Chatti, & Schroeder, 2014). Today videos are an essential part of large-scale learning environment like MOOCs (Guo, Kim, & Rubin, 2014) or Khan Academy¹, as well as in flipped classroom scenarios (Kurtz, Tsimerman, & Steiner-Lavi, 2014) and informal learning. These platforms experienced an increasing influx of participants, and due to its potential to support informal and self-directed learning the research interest for video-based learning is steadily increasing (Giannakos, 2013). In the context of these environments, analyzing the activities of learners centered around video learning resources is one approach to better understand how video-based learning takes place and how it can be improved. On the other hand, it is important to also relate learner-generated content to the video content provided by instructors. To achieve this, the gap between the different modalities of the teacher performance captured in visual and audio information and the learner-generated artefacts which are often digital text must be bridged. This requires the extraction of usable information about content and its structure from the raw video and audio data and transformation of video content and learner-generated content into forms that are suitable for comparison. Analyzing the relation of content from videos and learner-generated artefacts is still an underexplored area in Learning Analytics. Instead, previous work has focused on the interaction with videos themselves in the form of click-stream analyses to perform ex-post analyses on how students navigate through videos (Giannakos, Chorianopoulos, & Chrisochoides, 2015) or to give immediate visual feedback (Chatti et al., 2016). Analyses of learner-generated content in video-based learning settings has used specific

ontologies to determine students' conceptual understanding of specific topics (Daems, Erkens, Malzahn, & Hoppe, 2014). Similarly, Hecking, Dimitrova, Mitrovic, & Hoppe, (2017) investigated learner engagement through the analysis of learner-generated comments using a manually created taxonomy of domain keywords. They applied Network-Text-Analysis (NTA) (Carley, Columbus, & Landwehr, 2013) to extract learner-keyword networks. Both approaches make use of external resources given by experts to represent the video content, which can be an obstacle for scalable and real-time analyses that aim at comparing learner-generated texts and teacher provided video content.

In our approach, we combine transcription and extraction of thematically coherent segments of the video streams with analysis of learner-generated artefacts. To this end, a method for video segmentation was developed and different relevance measures were explored that describe how well learner-generated content overlaps with the content of each video segment. This approach has the advantage that no external ontologies and taxonomies describing the video content are needed and that specific parts of videos can be identified that are highly relevant in the sense that the conceptual knowledge they aim to convey is strongly reflected in the artefacts created by video watchers. There are several possible applications ranging from feedback for the instructors to nudging of learners to (re-)watch certain parts of a video that are not well reflected in student-generated content.

For the evaluation of the framework learner-generated artifacts in the form of wiki articles were analyzed in relation to teaching videos based on screen-capturing with audio. The wiki articles are well suited for evaluation since compared to other artefacts like video comments they are long enough such that techniques like NTA can be used for comparing video segments and wiki content.

2 APPROACH

In the sequel, we give an overview of the architecture of the framework and used methods. To make video content and student-generated wiki articles comparable, videos are transcribed to text using offline methods, i.e. CMU Sphinx (Lamere et al., 2003), or cloud based Speech-to-Text services from Microsoft, Google, or IBM. Next, the videos are split into thematically meaningful segments. To this end, there are different approaches. Visual segmentation aims to divide the video into scenes and shots, whereby scenes are story units and a collection of shots which are a spatial and temporal continuous segment of a video (Ngo, Pong, & ZHANG, 2001). The approach developed for this framework falls into this category. It is based on the video encoding of the MP4/ H.264 format developed by the Moving Picture Experts Group (MPEG) and Video Coding Experts Group (VCEG) which can be used for low- or high-quality videos and achieves compression of video data by removing spatial and temporal redundancy (Sullivan, Topiwala, & Luthra, 2004). This bitrate data can be extracted from the video and forms a graph with peaks at different locations. These peaks are detected using a simple peak detection algorithm¹. The assumption is that these peaks represent

¹ <http://billauer.co.il/peakdet.html>

borders between thematic segments signified through events like scene switches in camera guided videos or slide changes in screen presentations. The peak detection can be tuned through 3 parameters. These are the difference in height for surrounding points of a peak, a maximum distance between peaks to exclude close double peaks and a minimum threshold for the size of the peaks. Fig. 1 shows an example graph from a course about *Fundamentals of Interactive Teaching/Learning Systems* (FITLS).

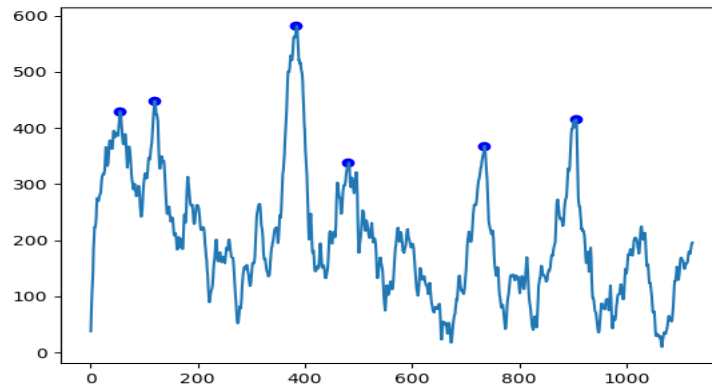


Figure 1 .: Example Bitrate graph in Kbits per second with detected peaks from a video in the course FITLS

The next step in the framework is to calculate relevance measures between content of video segments and learner-generated artefacts. For this purpose, a transformation into a network representation is performed using NTA. NTA has originally been conceived to reveal social structures from texts (Diesner & Carley, 2005). It has also been applied for the visual analysis of collaborative writing products in learning settings (Hecking & Hoppe, 2015). Conversion of text to a network uses a sliding window of a certain size over the text. Two words are connected if they appear together in the window. Different graph measures can then be applied to these networks based on the learner-generated artefacts and video segment transcripts.

3 EVALUATION

The dataset used for evaluation in this section consists of learning videos and wiki articles from the aforementioned course FITLS. The provided videos were lecture recordings of screen presentations using slides for different topics for which student groups wrote wiki articles. Wiki articles are rich artefacts containing mostly text for which the application of NTA is suitable. The analysis of wiki articles, compared to the corresponding video segments, can give insight into which part of the video content was well understood and taken up to write the wiki article. This information could be useful for teachers/tutors to identify problems in understanding or for the automatic generation of individualized hints to students. The chosen videos and corresponding wiki articles used for evaluation were for the following topics: (1) Cognitive Architectures and Cognitive Complexity Theory (CCT), (2) Direct Manipulation, and (3) Media Theories. For the first two videos a wiki article with the same topic was chosen and for the last video a wiki article which was about a subset of the topic was chosen. For evaluation of the video segmentation, ground-truth split points were determined by human judgement based on the criteria that there is a slide change after that a new

subtopic is introduced. Compared to these manual annotations, our automatic bitrate-based segmentation method achieved a recall value of 0.68 and precision of 0.44.

3.1 Relevance measures between video segments and wiki articles

Relevance measure for each segment can give us insight into which segments were important for the formulation of wiki articles. The measures used for calculating relevance were the following:

- “Degree centrality cosine similarity” (DCCS): degree centrality describes the number of edges connected to a node. Two vectors with the degree centralities of each word as entries are compared using cosine similarity.
- “Node intersection”: Number of words in common between the two compared artefacts.
- “Edge intersection”: Number of edges (defined by the two endpoints) shared between the two text networks. This measure incorporates the relation between words and can show whether students sufficiently connected different concepts
- “Word Mover Distance” (WMD): Calculates a distance measure between documents (Kusner, Sun, Kolkin, & Weinberger, 2015) based on learned semantic word vector representations (Joulin, Grave, Bojanowski, & Mikolov, 2016).

Measures listed above were calculated for the entire videos and all wiki articles, as well as for the individual segments of each video. For the entire videos each video and wiki article were compared pairwise. All measures correlated strongly with each other and showed the corresponding videos and wiki articles as strongly relevant. Yet, the distinction between relevant and non-relevant videos was weakest for the video “Media Theories” in which the relevant wiki article only covered part of the topics in the video.

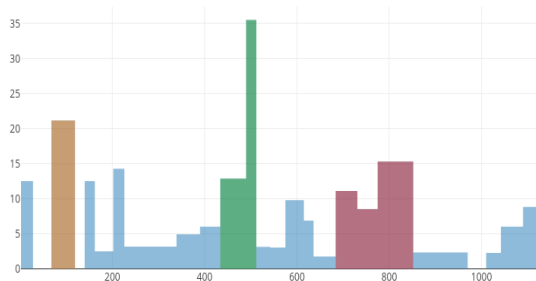


Figure 2: Edge Intersection for video segments of the CCT video and wiki article. X-axis – time in seconds, y-axis number of overlapping edges.

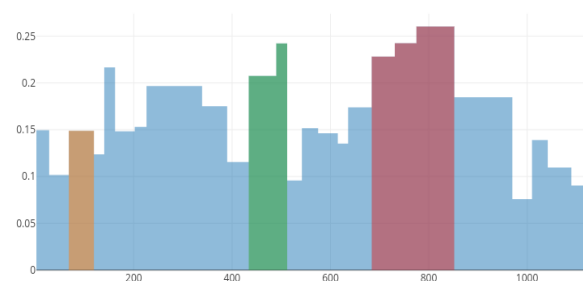


Figure 3: Cosine similarity of degree centrality for video segments of the CCT video and wiki article. X-axis – time in seconds, y-axis cosine similarity of segment and wiki article.

For the analysis of video segments, a relevance measure is needed that clearly maps segments into the categories of relevant and non-relevant similar to a “characteristic function” and therefore identifies connected coherent sequences of segments that are important for creation of the wiki articles. For the comparison of differences between measures for video segments, the video and the wiki article on the topic of Cognitive Complexity Theory (CCT) are used as an example case and the measures DCCS and edge intersection are presented in detail. The relevance measures are visualized

in Figures 2 and 3. Three parts of coherent segments were judged as important through a qualitative analysis done by the authors judging each segment as relevant or not-relevant. These are indicated by yellow, red, and green colors. Edge intersection showed high relevance scores for these 3 segment sequences and the following discusses how these scores can be explained. The first single segment is a description of CCT in the historic context, relating it to other technologies. In the second segment sequence consisting of 2 segments the workings of a CCT architecture are described in a practical way and an explanation of the consequences of activating a production rule is given. Many domain specific concepts are interlinked with each other and these same relations are reflected in the wiki article. The third segment sequence consisting of 3 segments covers the topic of a cognitive interpretation of production rule systems. Parts of the system are mapped to parts of a user that the system simulates. The wiki article also covers the relation of these concepts, which is an important piece in understanding CCT. The average score for relevant segments is 3.30 times higher than that of non-relevant segments allowing for a clear mapping.

For the DCCS, these sections do not differ substantially from other segments. There is no strong separation between relevant and not relevant segments. Indeed, the entire graph is quite homogenous. Average score for relevant segments was only 1.56 times higher than for non-relevant segments and the first relevant segment was scored lower than the average. There were also similar results for node intersection and WMD which are not displayed here.

In summary, even though the measures are equally viable to calculate relevance for entire videos, only edge intersection can clearly identify highly relevant segment sections of the video which had a big influence on writing the wiki article. This seems to be due to the incorporation of relations between relevant domain concepts.

4 DISCUSSION AND CONCLUSION

In this paper we presented a framework for content analysis of learning videos and calculation of relevance measures between the content of learning videos and learner-generated artefacts. The analysis of the videos included an approach to segment the video stream data into thematically coherent segments, representing the subtopics covered in the video. We also evaluated different relevance measures for comparing learner-generated content and learning videos. The evaluation showed that segmentation of screen-presentation videos is possible using bitrate changes of the video stream. Furthermore, different measures based on Network-Text-Analysis for the comparison of student-generated texts and content of video segments were evaluated on the dataset of lecture videos and wiki articles described in section 3. The measure of edge intersection appeared to be much better suited for identifying highly relevant segments of a video for the students' production of wiki articles compared to other measures. The identification of such segments can be used for further research. Identified segments could be used as an indicator of the understanding of concepts and their relations that students have. Low relevance scores for video segments could be used to trigger personal interventions, especially in comparison to relevance scores for wiki articles that historically had good grades or expert solutions. A second possible application is the analysis of revisions of wiki articles. This would allow for an understanding of how learning content was understood and incorporated over time.

5 REFERENCES

- Carley, K. M., Columbus, D., & Landwehr, P. (2013). Automap User's Guide. *Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR*, 13–105.
- Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Yousef, A. M. F., & Schroeder, U. (2016). Video annotation and analytics in CourseMapper. *Smart Learning Environments*, 3(1), 10.
- Daems, O., Erkens, M., Malzahn, N., & Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. *Journal of Computers in Education*, 1(2–3), 113–131. <https://doi.org/10.1007/s40692-014-0013-y>
- Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for research in information technology* (pp. 81–108). IGI Global.
- Giannakos, M. N. (2013). Exploring the video-based learning research: A review of the literature. *British Journal of Educational Technology*, 44(6), 191–195. <https://doi.org/10.1111/bjet.12070>
- Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *International Review of Research in Open and Distance Learning*. <https://doi.org/10.19173/irrodl.v16i1.1976>
- Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement. *Proceedings of the First ACM Conference on Learning @ Scale Conference - L@S '14*, 41–50. <https://doi.org/10.1145/2556325.2566239>
- Hecking, T., Dimitrova, V., Mitrovic, A., & Hoppe, H. U. (2017). Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching. *25th International Conference on Computers in Education*, 326–335.
- Hecking, T., & Hoppe, H. U. (2015). A network based approach for the visualization and analysis of collaboratively edited texts. *CEUR Workshop Proceedings*, 1518, 19–23.
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. <https://doi.org/10.1111/09249v1>
- Kurtz, G., Tsimerman, A., & Steiner-Lavi, O. (2014). The Flipped-Classroom Approach: The Answer to Future Learning? *European Journal of Open, Distance and E-Learning*, 17(2). <https://doi.org/10.2478/eurodl-2014-0027>
- Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning* (pp. 957–966).
- Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong (Vol. 1, pp. 2–5)*.
- Ngo, C.-W., Pong, T.-C., & ZHANG, H.-J. (2001). Recent advances in content-based video analysis. *International Journal of Image and Graphics*, 1(03), 445–468.
- Sullivan, G. J., Topiwala, P. N., & Luthra, A. (2004). The H.264/AVC advanced video coding standard: overview and introduction to the fidelity range extensions, 454. <https://doi.org/10.1117/12.564457>
- Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). The state of video-based learning: A review and future perspectives. *International Journal on Advances in Life Sciences*, 6(3–4), 122–135. [https://doi.org/10.1016/S0049-3848\(13\)70151-8](https://doi.org/10.1016/S0049-3848(13)70151-8)

Multimodal Learning Analytics Runtime Framework

Jan Schneider

DIPF | Leibniz Institute for Research and Information in Education
Schneider.Jan@dipf.de

Daniele Di Mitri

Open University of the Netherlands
Daniele.Dimitri@ou.nl

Hendrik Drachsler

DIPF | Leibniz Institute for Research and Information in Education
drachsler@dipf.de

Marcus Specht

Open University of the Netherlands
Marcus.Specht@ou.nl

ABSTRACT: Multimodal Learning Analytics (MMLA) applications allow learners to take advantages of the digital world while performing any type of learning activity without being constrained by the direct interaction of a traditional computer interface. In recent years MMLA applications designed to support learners in specific learning activities have been developed and studied. Developing this type of applications and sharing knowledge among them is a difficult process. To address this issue, we propose the development of a generic MMLA runtime framework. In this paper, we describe our proposed requirements for the development of this framework.

Keywords: Multimodal Learning Analytics, System Architecture, Feedback

1 INTRODUCTION

Learning Analytics (LA) applications should be created for ultimately help the learners in achieving their goals. Traditional LA applications are fed with data that comes from direct mouse and keyboard interactions between learners and learning management systems (LMSs). Learning, however, is not constrained to these types of direct interactions. Learning happens in multiple scenarios including the deliberate practice of a skill, face to face discussions, contemplating a phenomenon or reflecting on past events. With sensors, it is possible to unobtrusively capture the behavior and environment of learners, and support them throughout their learning process (Schneider et al., 2015), hence expanding the horizons of LA.

The direct mouse and keyboard inputs are straightforward to interpret for computational systems, thus making use of them is also straightforward. In the case of sensor-data, its interpretation becomes more complex, interpretations must be inferred. Usually, sensor-data from multiple sources is required to do obtain “good enough” interpretations. Multimodal Learning Analytics (MMLA) is the use of multiple data sources in order to

understand and support the learning process. MMLA applications should provide learners with learning interventions in order to help learners to achieve their goals. Feedback is one of the most important interventions in learning (Hattie, & Timperley, 2007). MMLA applications enable to feedback the results of the recorded and analyzed data to learners. Keeping learners in the “feedback loop” is highly relevant for testing the actual relevance of MMLA applications.

MMLA as a field of research appeared in 2013 (Blikstein, 2013), MMLA applications are still relatively complex and costly to develop, most of the research on MMLA is focused on tailored-made applications (Di Mitri et al., 2018). This tailored-made approach makes it difficult to share best practices and general knowledge among people working in MMLA, thus hindering the progression of the field. The use of generic approaches and frameworks can facilitate the sharing of best practices and knowledge, reduce the development costs of MMLA applications and contribute to the field. In this article, we present and discuss the requirements, constraints, and characteristics of a generic MMLA runtime framework.

2 BACKGROUND

MMLA real-time feedback applications have already been studied for a wide variety of learning scenarios including playing the violin (Van Der Linden et al. 2013), practicing snowboarding (Spelmezan et al. 2009), public speaking (Barmaki & Hughes, 2018; Schneider et al., 2016), etc. Most of these sensor-based applications support learners with the practice of their skills while receiving feedback regarding their performance. The feedback for these type of applications needs to be carefully designed so that the learner is able to interpret it correctly while conducting a practice session. For example, using dashboard interfaces to present real-time feedback to learners has shown to be overwhelming for learners (Schneider et al. 2015), in contrast to MMLA applications that at maximum display one feedback instruction at a given time and have shown to significantly help to improve the learner’s performance (Schneider et al., 2016).

An example of the feedback mechanism used by MMLA applications can be seen with the Presentation Trainer (PT), which is a research prototype designed to support the development of nonverbal communication skills for public speaking. The PT used multimodal data captured by a depth camera and an array of microphones. With the depth-camera is possible to infer the current posture of the learner and with the microphones, it is possible to infer the current volume of her voice and hence whether she is speaking or not. The PT uses a rule-based system that triggers some feedback instructions once some data values are identified. For example, if the volume data happens to be above a certain threshold for a certain period of time the PT triggers a “speak softer” feedback instruction.

The methods to integrate and make sense of the data coming from the depth camera and microphones and the feedback rules used by the PT are hardcoded. This is what we mean by tailored made for the specific application. Adding new sensors and feedback rules to the PT so that it could support the development of different skills is impractical.

3 TOWARDS A GENERIC MMLA RUNTIME FRAMEWORK

3.1 Step 1 Multimodal data collection and integration: The Multimodal Learning Hub

The Multimodal Learning Hub (*LearningHub*) (Schneider et al., 2018) is a system that focuses on the data collection and data storing of multimodal learning experiences. It uses the concept of Meaningful Learning Task (MLT) and introduces a new data format (MLT session file) for data storing and exchange. The *LearningHub* implements a set of specifications that shape it for certain types of learning activities. It was created to be compatible primarily with commercial devices (e.g. Microsoft Kinect, Leap Motion, Myo Armband) and other sensors with drivers running in operating systems that allow UDP and TCP communication protocols. It focuses on short and meaningful learning activities (10 minutes) and uses a distributed, client-server architecture with a master node controlling and receiving updates from multiple data-provider applications. It also handles video and audio recordings with the main purpose to support the human annotation process. The expected output of the *LearningHub* is one (or multiple) MLT session files including 1) one-to-n multimodal, time-synchronized sensor recordings; 2) a video/audio file providing evidence for retrospective annotations.

3.2 Step 2 Keep me in the Loop: Proposed approach

Our proposed approach is to create a generic MMLA runtime framework (MMRunTime) designed to support the learner in context. For that, the plan is to develop a feedback engine that makes use of the multimodal data collection and integration services provided by the *LearningHub*.

The proposed architecture of the MMRunTime is shown in Figure 1. It features multimodal sensor interfaces consisting in both sensors applications (input devices) and feedback actuators (output devices) or a combination of the two. The MMRunTime consists of the *LearningHub* and a *Feedback Engine*, which is able to send activations to multiple feedback actuators based on certain feedback rules. The feedback rules are both dependent on Expert rules and the state of the learning environment which is acquired with the *Runtime Interpreter*.

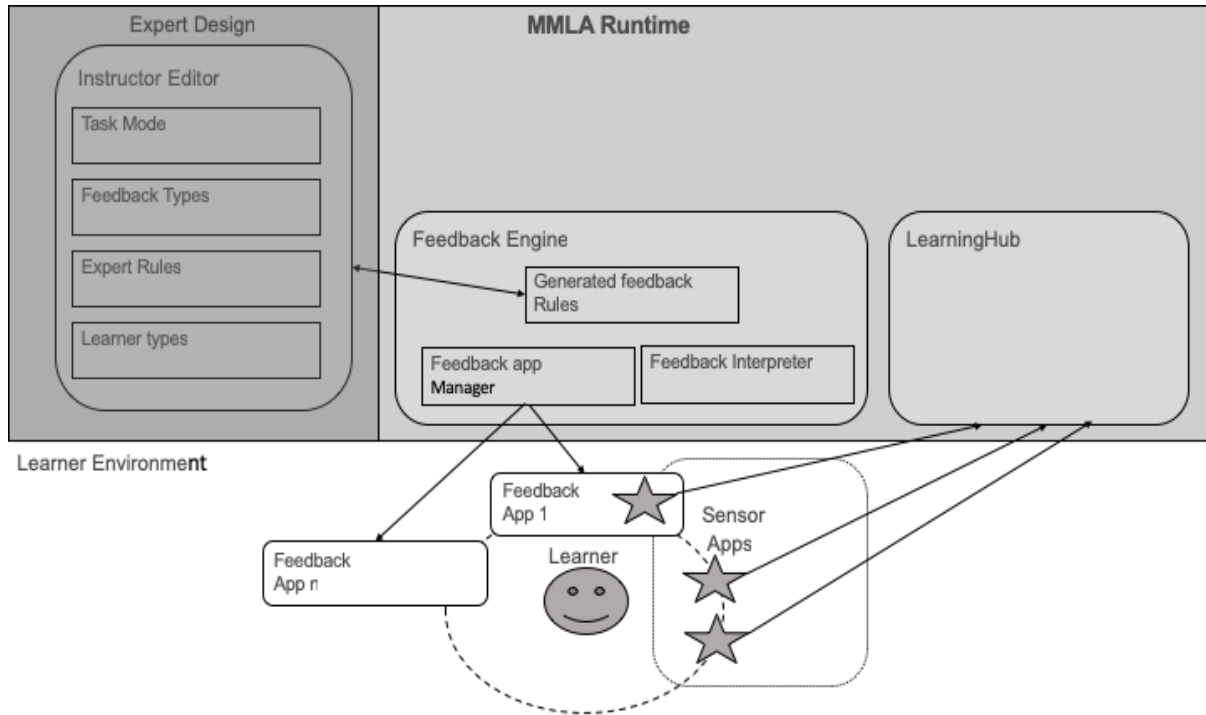


Figure 1. MMLA RunTime Architecture

The *Expert Rules* are defined in the *Instruction Editor*. In this editor, the Expert is able to define the type of task that the learner is executing, the kind of feedback which can be prompted and the learner types, which are based on the kinds of errors that the learner makes. Considering these three dimensions, the expert is able to define certain rules which are generated at runtime in the *Feedback Engine*.

3.2.1 Feedback Rules

Feedback depends on the assessment of the learning task performance. The assessment can be seen as the comparison of the learning performance according to the function which defines the threshold of (non-)optimal learning performance. This function of (non-)optimal can be handcrafted based on heuristics, known by the learning domain experts and expressed as forms of conditions (or constraints). In the domain of presentation, relevant conditions can, for example, be “if voice is too low then speak louder” or “if arms are crossed then release arms” and so on.

The feedback can use Constraint-Based Modelling (Kodaganallur, 2005): the rules can be expressed in form of constraints like tuple $\langle Cr, Cs \rangle$, where Cr is the Relevant Condition and the Cs is the Satisfaction Condition (Cs). If Cs is not satisfied then we prompt a Feedback error. An example constraint in the case of the Presentation Trainer: if the learner is in “presentation mode” and the microphone is active, if the volume of the microphone is below 10, then prompt the feedback to a feedback device. This example is shown in Listing 1.

```
// Not speaking Loud enough (Myo feedback)
IF (presentationMode==True) && micActive==True THEN // Cr
```

```

IF mic.Volumne<10 THEN // Cs
    feedbackDevice = 'Myo';
    label = 'VibrateMedium'
    message = 'Speak Louder'
END IF

```

Listing 1. Example feedback rule when speaking not enough

3.2.2 Limitations Beyond expert rules

The proposed MMRunTime based on rule-based models, presents some shortcomings. In the field of MMLA rule-based models, have some shortcomings. Modalities like physiological responses can be quite complex and counter-intuitive. It could be difficult, for instance, to define rules like the one shown in Listing 1 with brain waves collected with an EEG. An alternative consists using computerized algorithms can compute these rules automatically and approximate the function of (non-)optimal learning “a posteriori”, analyzing the collected data and associating it with expert evaluations. This is typically defined as the machine learning approach, described in the MLeAM model (Di Mitri et al., 2018).

The vision is such that multiple machine learning models are stacked and reused so that future applications don’t need to refer to the single sensor value but rather to an aggregated interpretation. This option goes into an *Activity recognition module*.

Along with activity recognition, another important dimension in learning is the emotional sphere. Research has shown certain types of emotions play an important role in learning together. For this reason, we place an *Emotion recognition module*.

Special attention should be paid to the co-located collaborative tasks. In these cases, the ITS can facilitate the learning of multiple learners. In addition, the *LearningHub* could be collecting data from multiple learners. Distinguish “Who did what” or “Who said that” can be not easy (Martinez-Maldonado, 2011). We propose to add the *User recognition module* to the *LearningHub*. Using the MLeAM, we aim to create learner-specific models which are trained to recognize the movements, voice intonation, skin *color*, gestures and faces of individual learners in the group. A softmax classification algorithm should be added on top of this model to correctly classify a particular moment in the session to the active learner. Consequently, feedback can be prompted to one specific learner or to the group.

We are still analyzing reliable options for the integration of these modules to the proposed MMRunTime.

4 CONCLUSIONS

While the proposed MMRunTime has some limitations, we foresee that the development of a generic framework like it, will greatly contribute to the field of MMLA by speeding up the process of the development of MMLA research prototypes. Therefore, as future work, our plan is to continue with this endeavor.

REFERENCES

Barmaki, R., & Hughes, C. E. (2018). Embodiment analytics of practicing teachers in a virtual immersive environment. *Journal of Computer Assisted Learning*.

- Blikstein, P. (2013, April). Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 102-106). ACM.
- Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4), 338-349.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.
- Kodaganallur, V., Weitz, R. R., & Rosenthal, D. (2005). A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal of Artificial Intelligence in Education*, 15(2), 117-144.
- Martínez, R., Collins, A., Kay, J., & Yacef, K. (2011, November). Who did what? Who said that?: Collaid: an environment for capturing traces of collaborative learning at the tabletop. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (pp. 172-181). ACM.
- Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2015). Augmenting the senses: a review on sensor-based learning support. *Sensors*, 15(2), 4097-4133.
- Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2015). Stand tall and raise your voice! a study on the presentation trainer. In *Design for teaching and learning in a networked world* (pp. 311-324). Springer, Cham.
- Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2016). Can you help me with my pitch? Studying a tool for real-time automated feedback. *IEEE Transactions on Learning Technologies*, 9(4), 318-327.
- Schneider, J., Di Mitri, D., Limbu, B., & Drachsler, H. (2018, September). Multimodal learning hub: a tool for capturing customizable multimodal learning experiences. In *European Conference on Technology Enhanced Learning* (pp. 45-58). Springer, Cham.
- Spelmezan, D., Jacobs, M., Hilgers, A., & Borchers, J. (2009, April). Tactile motion instructions for physical activities. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2243-2252). ACM.
- Van Der Linden, J., Johnson, R., Bird, J., Rogers, Y., & Schoonderwaldt, E. (2011, May). Buzzing to play: lessons learned from an in the wild study of real-time vibrotactile feedback. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 533-542). ACM.

Learning objectives and curriculum standards as multimodal analytics in learning platforms

Morten Misfeldt¹, Benjamin Brink Allsopp², Jonas Dreyøe³ and Andreas Lindenskov Tamborg⁴,

Department of Learning and Philosophy, Aalborg University

1: misfeldt@learning.aau.dk, 2: ben@learning.aau.dk, 3: jmd@learning.aau.dk, 4: alt@learning.aau.dk

Daniel Spikol

Internet of Things and People, Malmö University

daniel.spikol@mau.se

ABSTRACT: In this position paper, we describe how learning platforms that support teachers' documentation work provide multimodal data about teachers' planning and documentation. The paper describes a new type of learning platform that are being used in the Danish school system, where there is a well-established relation between national curriculum standards and teachers' planning and documentation work. We describe a software prototype in which teachers register their planned lessons and examples of analytics from data produced in experiments with teachers using the prototype. Based on this, we discuss the advantages of viewing this type of learning platform as a site for multimodal analytics.

Keywords: Multimodal learning analytics, Epistemic Network Analysis, Mathematical Competencies, Learning Platforms.

1 INTRODUCTION: OUTPUT ORIENTED CURRICULUM DOCUMENTATION OF TEACHING AND LEARNING

Learning platforms, as they are being implemented in several countries are caught between the potentials that learning analytics brings and concerns raised in more established disciplines about teaching and learning. One important concern is how the data generated mediates between on one side, the curriculum and standards, which are typically politically decided, and on the other side the teaching situation as it occurs in the classroom. **Hence, we need to study examples of ecosystems consisting of curriculum, tools for supporting teacher work and actual teaching situations.**

The new Danish curriculum (Undervisningsministeriet, 2014) is organized around competencies which are broken down into pairs of knowledge and skills to be learned by the students and formulated as something that the students should be able to do after the teaching has ended. The curriculum can be presented in several graphical modes, e.g. in a matrix or in a hypertext structure. After the reform, the curriculum for Mathematics now consists of four overall areas (1) Mathematical competencies, (2) Numbers and Algebra, (3) Geometry and Measures, and (4) Statistics and Probability. These four areas are then broken down into themes and further into

several pairs; consisting of knowledge and skills showing the envisioned progression in the theme. The structure is shown in figure 1.

Matematik

Færdigheds- og vidensmål (efter 3. klassetrin)

See fig. 2.



Kompetenceområde	Kompetencemål	Faser	Færdigheds- og vidensmål									
Matematiske kompetencer	Eleven kan handle hensigtsmæssigt i situationer med matematik		Problemsbehandling		Modelisering	Ræsonnement og tankegang	Repræsentation og symbolsbehandling		Kommunikation		Hjælpemidler	
		1.	Eleven kan bidrage til løsning af enkelte matematiske problemer	Eleven har viden om kendskabet ved udvalgte matematiske problemer	Eleven kan undersøge enkelte hverdagsituationer ved brug af matematik	Eleven kan stille og besvare matematiske spørgsmål og svar	Eleven kan anvende konkrete, visuelle og enkelte symbolske repræsentationer	Eleven har viden om konkrete, visuelle og enkelte symbolske repræsentationer	Eleven kan deltage i mundtlig og visuel kommunikation med og om matematik	Eleven har viden om enkelte mundtlige og visuelle kommunikationsformer, herunder digitale værktøjer	Eleven kan anvende enkelte hjælpemidler til tegning, beregning og undersøgelse	Eleven har viden om enkelte materialer og redskaber
		2.	Eleven kan løse enkelte matematiske problemer	Eleven har viden om enkelte strategier til matematisk problemløsning	Eleven kan tolke matematiske resultater i forhold til enkelte hverdagsituationer	Eleven kan give og følge skriftlige matematiske forklaringer	Eleven har viden om enkelte matematiske forklaringer	Eleven kan anvende matematiske tænkning med uformelle skriftlige noter og tegninger	Eleven har viden om forskellige former for uformelle skriftlige noter og tegninger	Eleven kan anvende enkelte fagord og begreber mundtligt og skriftligt	Eleven har viden om digitale værktøjer til undersøgelse, beregning og tegning	Eleven har viden om enkelte digitale værktøjer til undersøgelse, beregning og tegning
Tal og algebra	Eleven kan udvikle metoder til beregninger med naturlige tal		Tal ①		Regnearter ②	Algebra						
		1.	Eleven kan anvende naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om enkelte naturlige tal	Eleven kan foretage enkelte beregninger med naturlige tal	Eleven har viden om strategier til enkelte beregninger med naturlige tal	Eleven kan opdagde systemer i figur- og talnumre	Eleven har viden om enkelte figur- og talnumre	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge
		2.	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven kan anvende forskellige naturlige tal til at beskrive antal og rækkefølge	Eleven har viden om forskellige naturlige tal til at beskrive antal og rækkefølge
Geometri og måling	Eleven kan anvende geometriske begreber og måle		Geometriske egenskaber og sammenhænge		Geometrisk tegning	Placeringer og flytninger	Måling ③					
		1.	Eleven kan kendetegne figurer	Eleven har viden om egenskaber ved figurer	Eleven kan beskrive egne tegninger af omgivelser med geometriske begreber	Eleven kan beskrive objekters placering i forhold til hinanden	Eleven har viden om enkel længde, tid og vægt	Eleven kan beskrive enkel længde, tid og vægt	Eleven kan kendetegne figurer	Eleven har viden om egenskaber ved figurer	Eleven kan beskrive egne tegninger af omgivelser med geometriske begreber	Eleven kan beskrive objekters placering i forhold til hinanden
		2.	Eleven kan kendetegne figurer efter geometriske egenskaber	Eleven har viden om egenskaber ved figurer	Eleven kan tegne enkelte plane figurer ud fra givne betingelser og plane figurer, der gengiver enkelte træk fra omverden	Eleven har viden om metoder til at tegne enkelte plane figurer, herunder med et dynamisk geometriprogram	Eleven kan beskrive og fremstille figurer og mønstre med spejlings-, symmetri- og rotationsmetoder	Eleven har viden om metoder til at måle længde, tid og vægt samt om analoge og digitale måleinstrumenter	Eleven kan anvende geometriske begreber og måleinstrumenter	Eleven har viden om egenskaber ved figurer	Eleven kan anvende geometriske begreber og måleinstrumenter	Eleven kan anvende geometriske begreber og måleinstrumenter
Statistik og sandsynlighed	Eleven kan udføre enkelte statistiske undersøgelser og udtrykke resultaterne		Statistik		Sandsynlighed							
		1.	Eleven kan anvende tabeller og enkelte diagrammer til at præsentere resultater af optællinger	Eleven har viden om enkelte tabeller og diagrammer	Eleven kan udtrykke enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte chancestørrelser i hverdagsituationer og enkelte spil
		2.	Eleven kan genkende enkelte statistiske undersøgelser med enkelte data	Eleven har viden om enkelte metoder til at samle, ordne og beskrive enkelte data	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte metoder til at samle, ordne og beskrive enkelte data	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte metoder til at samle, ordne og beskrive enkelte data	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte metoder til at samle, ordne og beskrive enkelte data	Eleven kan anvende enkelte chancestørrelser i hverdagsituationer og enkelte spil	Eleven har viden om enkelte chancestørrelser i hverdagsituationer og enkelte spil

① Se bilag for opmærksomhedspunkter

Fig. 1. the curriculum for grade 1-3 in mathematics

The many 'atomized' learning objectives together with an increased focus on documental work of teachers has led to the development of several new platforms that supports documentation, preparation and the actual conduction of teaching (see Misfeldt et al. accepted for further description).

Competence area	Competence goal	Stages	Goals for skill and knowledge	
Mathematical competencies	The student can act appropriate in situations involving mathematics		Problem solving	
		1	The student can contribute to the solution of simple mathematical problems	The student knows characteristics of investigative work
		2		
		3	The student can solve simple mathematical problems	The student knows simple problem solving strategies
Numbers and algebra	The student can develop methods for calculation with natural numbers		Numbers	
		1	The student can use natural numbers to describe amount and order	The student knows of simple natural numbers
		2	The student can use natural numbers with multiple digits to describe amount and order	The student knows of the natural numbers construction in the decimal number system
		3	The student can recognise simple decimal numbers and fractions in everyday-situations	The student knows of simple decimal numbers and fractions.

Fig. 2. Translation of part of the curriculum for grades 1-3 in mathematics (see fig. 1.)

A review of the literature regarding learning and teaching platforms (Tamborg et al. submitted), address the question of what themes, potentials and pitfalls that are discussed in the international literature on learning platforms, and to what extent this knowledge can qualify as research and practice of learning platforms in a Danish context. The review identifies 21 studies that fall into three categories in their themes (1) Digital learning platforms support of pupil learning, (2) Implementation of digital learning platforms, (3) Skills development of pedagogical staff about the use of digital learning platforms.

2 ENCODING LEARNING OBJECTIVES ‘THE GOAL ARROW’

This is a derivative of a design-based research project involving the development of a digital tool that distinguishes the National standards, from the day to day objectives in the classroom. *The Goal Arrow* is a tool that allows teachers to express their own learning goals for their students and evaluate their progress in relation to those goals (Misfeldt, Bundsgaard, Slot, Hansen, Jespersen, 2015). One of the promises of the tool is to move the assessment practice away from single situations (e.g. tests) and towards a more integrated, ubiquitous and ongoing part of the teaching practice.

The Goal Arrow supports teachers in describing lesson plans, expressing associated situated learning goals and relating these to the National Curriculum. Each learning goal is specified into three objectives, which can be identified in students’ actions or products. The descriptions, goals and objectives can be used when the teacher communicates with the students about the goals of the course. Each goal is also related to the national curriculum by clicking on the various elements in a matrix similar to the one in figure 1.

The objectives are then used to measure individual progress by teachers and students, alike. Data is collected in relation to several local learning goals and presented as shown in figure 3.

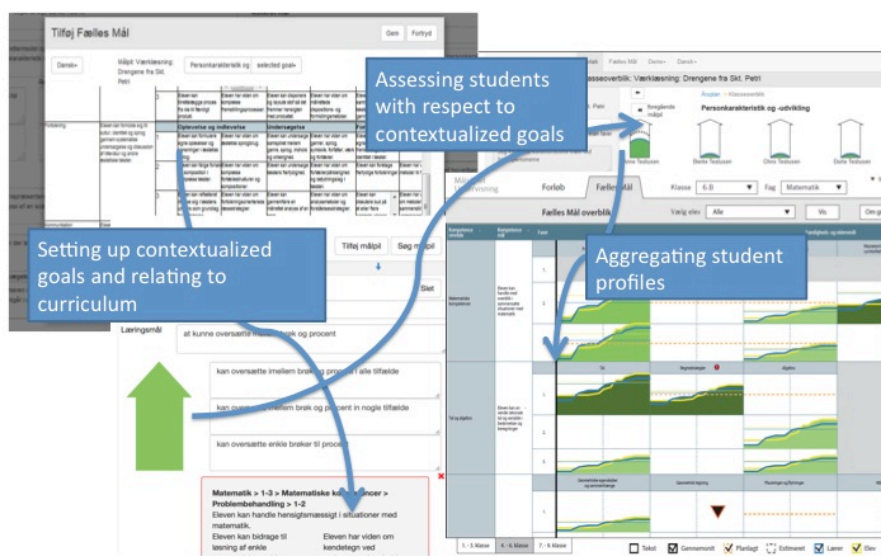


Fig. 3. The functionality of the Goal Arrow

The initial results from our work with the technology suggests that the Goal Arrow is a well-functioning tool for setting up learning goals, relating them to the National curriculum and using them as marker points in the class activities. However, we still need to see the benefits of the Goal Arrow as a mean for developing student profiles and thus becoming a real alternative to event based assessment.

The tool was tested over three months with approximately 100 teachers in 10 schools. We are currently finishing a report describing the intervention. Hence, we have results from both qualitative and quantitative investigations that we can bring to the symposium.

3 A MULTIMODAL LENS

Multimodal data is used differently among researchers within the field of Learning Analytics. Worsley et al. (2016) argue that the essence of multimodal learning analytics is the utilization and triangulation of “non-traditional as well as traditional forms of data in order to characterize or model student learning(...)” and a recognition that “teaching and learning are enacted through multiple modalities” (Worsley, Abrahamson, Blikstein, Schneider, Grover, & Tissenbaum, 2016, p. 1). Student learning in school contexts is often directly dependent on teaching activities. The Goal Arrow collects data about teachers’ documentation work that consists of the curriculum standard(s) chosen by the teacher, the teachers’ interpretation of this/these standards (including a taxonomy dividing the standard into three sub-levels), and a specification of the resources, tasks and activities enabling students to meet the standard(s). Together, these data sources contain representations for teacher activities in distinct modalities, which together enable different entry points for understanding and interpreting teachers’ planning of student learning.

4 ANALYSIS: TYPOLOGY OF MATHEMATICAL COMPETENCIES

In our preliminary analysis, we used the data generated in the Goal Arrow when teachers chose one or more standard from the curriculum that the particular lesson addressed. We extracted data from Goal Arrow into a spreadsheet in which every row represented a lesson planned by the teacher, and each column had data about the lesson plan. This data contains information about the curriculum standard(s) chosen by the teacher and a description of the objective of the lesson by the teacher. Moreover, we included data about the taxonomy of the objective, a description of activities and resources in the lessons, and how the lessons were to be evaluated. The spreadsheet also had background variables, an ID of the teacher, the school they work at, grade level and subject. For the analysis presented here, we clustered the teacher-chosen curriculum standards according to what overall competencies they belong. This clustering allowed us to investigate how teachers combined different competencies in their lesson plans. This matrix is used as input into the ENA tool, which by using default settings performs a singular value decomposition based on co-occurrence of competencies in lesson plans (for more on the method, see Shaffer & Ruis, 2017). We then configured three views on the data corresponding to the lesson plans for students in the grade ranges 1st to 3rd, 4th to 6th and 7th to 9th. This operation resulted in the visualizations below (Allsopp et al 2017) .

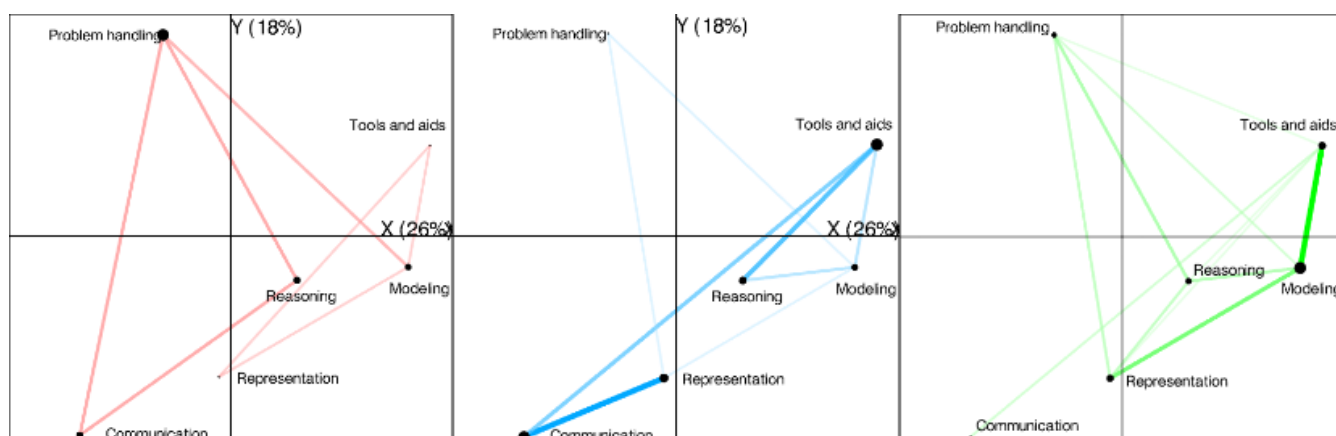


Figure 4: Three networks of learning objectives – for different grade levels

Looking at the visualizations we see a strong focus on communication reasoning and problem-solving in the lower grades (1-3 to the left), but in the middle school (grade 4-6 in the middle), problem tackling is almost out of the picture. The visualization produced for grade 7 to 9 (to the right), suggests a teaching that is heavily oriented towards mathematical modelling and real-world problems.

5 MULTIMODAL PLATFORM ANALYSIS

From the previous sections, it is apparent that our initial analysis of the platform data relies on data in a single modality; the curriculum standards chosen by the individual teacher. Hitherto, this data has enabled us to explore differences in how teachers connect standards across different grade level. Integrating the other data available in the platform allows us to gain a deeper understanding of the reasons behind these differences. For example, a semantic analysis of the specific goals the

teachers articulate, sorted by the different connections, could support an analysis of the how's and why's of the different connections of competencies.

Moreover, the relation between evaluation forms, teaching activities and curriculum standards could also be investigated to either explore differences and/or similarities in assessment forms across grade levels as well as competency connections. This could support a better understanding of teacher practices and inform teachers of trends in their assessment practices that could be improved. Relations between resources and curriculum standards – are certain areas of the curriculum typically covered by using particular resources, and are there good reasons for this?

Furthermore, the platforms contain data about student progress, which provide an obvious data source for investigating how the choices of activities, resources and objective of lessons affect student learning. Rather than providing a deterministic understanding of how teaching *should* be done, this holds the potential of giving teachers and researchers a better understanding of the relation between teaching and learning.

For the potential of supporting the understanding between teaching and learning, stronger theoretical foundations are needed (Wise & Shaffer, 2015) to understand the data and to investigate beyond single modalities. Our paper begins to explore how to create this foundation that would allow further development of tools and research into the relationship between teachers, the learning resources, and progress on specific curricular elements. Additionally we aim to explore the organisation of teaching (group work, workshops, and class) that have scales of progress. The development of school typologies and traces of the effects of specific initiatives (e.g. in-service training) on student learning is key challenge in Denmark and Sweden. Additionally, once we establish a stronger foundation, additional means of multi-modal data collection can be used that explore how teachers and students engage physically and digitally across learning materials and classroom space. All the potentials mentioned above are interesting research areas in themselves; however, more importantly, they can constitute a catalyst for pedagogical reflection and discussion among teachers and other educational professionals – individually and collectively.

REFERENCES

- Allsopp, B. B., Dreyøe, J., & Misfeldt, M. (2017). Using Epistemic Network Analysis to understand core topics as planned learning objectives. *Poster presented at Learning Analytics Summer Institute, Bergen, Norway.*
- Misfeldt, M., Bundsgaard, J., Slot, M. F., Hansen, T. I., & Jespersen, M. (2015). A Digital Tool Supporting Goal-Oriented Teaching in Classrooms. In A. Jefferies, & M. Cubric (Eds.), *Proceedings of 14th European Conference on e-Learning ECEL-2015* (pp. 388-395). Reading, UK: Academic Conferences and Publishing International.
- Misfeldt, M., Tamborg, A. L., Dreyøe, J., & Allsopp, B. B. (Accepted/In press). Tools, rules and teachers: The relationship between curriculum standards and resource systems when teaching mathematics. *International Journal of Educational Research.*
- Tamborg A., L., Bjerre A., R., Andreassen, L., B., Albrechtsen, T. R.S. & Misfeldt, M. (Submitted). Review over international forskningslitteratur om digitale læringsplatforme, submitted to *Learning Tech.*

- Shaffer, D. W., & Ruis, A. R. (2017). Epistemic Network Analysis: A Worked Example of Theory-Based Learning Analytics. *Handbook of Learning Analytics* 175-187. Wise, A. F., & Shaffer, D. W. (2015). Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics*, 2(2), 5–13. <http://doi.org/10.18608/jla.2015.22.2>
- Undervisningsministeriet (2014). Forenklede Fælles Mål available from [https://www.emu.dk/soegning?f\[0\]=field_omraade%3A5464&f\[1\]=field_tags%3A27735](https://www.emu.dk/soegning?f[0]=field_omraade%3A5464&f[1]=field_tags%3A27735)
- Wise, A. F., & Shaffer, D. W. (2015). Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics*, 2(2), 5–13. <http://doi.org/10.18608/jla.2015.22.2>
- Worsley, M., Abrahamson, D., Blikstein, P., Schneider, B., Grover, S., Tissenbaum, M. (2016) 12th International Conference of the Learning Sciences: Transforming Learning, Empowering Learners, ICLS 2016. Situating multimodal learning analytics. *Proceedings of International Conference of the Learning Sciences*, Icls, 2, 1346-1349.

Analyzing Affective States Alongside Qualitative Analysis

Kit Martin

Northwestern University

kitmartin@northwestern.edu

Emily Wang

Northwestern University

eqwang@u.northwestern.edu

Connor Bain

Northwestern University

connorbain2015@u.northwestern.edu

Marcelo Worsley

Northwestern University

marcelo.worsley@northwestern.edu

ABSTRACT: We build on work in museums to apply multimodal learning analytics to investigate interaction with a digital exhibit, Ant Adaptation. We use emotion data to track affective state of participants with the exhibit. We then examine how cross examining qualitatively coded data of the interaction with affective state sheds light on moments of learning in the interactions. In this paper we first show how information retrieval techniques can be used on facial expression features to show emotional variation during key moments of the interaction. Second, we connect these features to moments of learning identified by other qualitative methods. Finally, we present an initial pilot using these methods in concert to identify key moments in multiple modalities.

Keywords: Emotion Tracking, Multimodal Learning Analytics, Informal Learning Environments

1 INTRODUCTION

Learning is everywhere, and learning environments can provide rich educational experiences without the need for an instructor or a classroom (National Research Council, 2009). While these sorts of environments are becoming more common, and museums have long used learning environment design to engage their visitors, new analytics is adding to their evaluation (D’Mello, Dieterle, Duckworth, 2017). As a community of museum educationalists, we need to understand how participants learn in these designed, but more open learning environments. It is often difficult to track learning through these environments (Ochoa & Worsley, 2016). In this workshop, we will present our developing method to track learning in open ended learning environments using multimodal learning techniques, such as body tracking, affect tracking, and knowledge mapping. This effort builds from current theories of learning and their ways of understanding.

Currently, research in museum exhibits uses ethnography. We either take field notes and write them up as research memos, or we analyze video of the interactions back at the lab. In this workshop, we

show how we have been moving toward multimodal observation through instrumentation in addition to traditional ethnography to capture multiple kinds of interaction. And while our methods may lose some of the depth of focused qualitative methods (Miles, Huberman, & Saldana, 2014), we see promise in sharing our current work with the Cross MMLA community.

1.1 Prior Work

Prior work in neurobiology, ethnography, and artificial intelligence, as applied in learning research, informs our work. We situate this work in the broader growing field of multimodal learning analytics (MMLA) (Ochoa & Worsley, 2016; Oviatt, Cohen, & Weibel, 2013; Schneider & Blikstein, 2015). In MMLA, recent work explored the relationship between cognition and emotions. Neurobiological research shows emotionally arousing stimuli increase the consolidation, preservation, and encoding of memory engrams (McGaugh, 2003; McGaugh, 2006). This process is associated with both negative affect, like confusion, and positive affect, like joy (McGaugh, 2004). D’Mello and Graesser (2012) advanced a model of affect dynamics that describes the complex interactions among different affective states and how these states afford learning. Specifically, they focused on how students transition into and out of moments of confusion. In their model, surprise and joy serve as proxy indicators of a student moving in or out of a moment of confusion. While still focused on short term cognitive gains, the model serves as an example of how examining affective states can improve learning research, as it points to key moments of change and encoding.

Worsley, Scherer, Morency, & Blikstein (2015) similarly leveraged facial expressions (affective state) to segment multimodal data streams. They use changes in facial expression as a proxy measure of changes in learners’ cognitive and behavioral states. Like with all affective state tracking, their approach uses a probabilistic detection approach and tracks the likely state. Underlying this proxy is the assumption that the detected state reflects the participant’s internal state and that it may shed light on complex learning processes. Of course, the performative nature of facial expressions means they may only be partial proxies of internal states (Howell, Chuang, De Kosnik, Niemeyer, & Ryokai, 2018). In this workshop, we will present our approach to affect tracking around a digital learning environment in a museum and discuss the limitations of the affective state tracking methods in these environments.

2 METHODS

We have collected data from 122 participants as they used a museum digital interactive exhibit called *Ant Adaptation*. Shown in Figure 1, *Ant Adaptation* is a game built from an agent-based model implemented in NetLogo (Wilensky, 1999).

The game teaches the role of animal adaptations, such as size and aggressiveness, on animal colony success in an ecosystem. The complex systems ideas of system feedback and food source proximity affect the colony success in an ecosystem and are emergent in the game from the manipulations of animal adaptations, such as size and aggressiveness. Two players participate in the game at a time, each controlling an ant colony. Game play is facilitated through digital sliders that affect the size and aggressiveness of the ants. Players can also touch the screen to place chemicals (pheromones) the ants follow towards food or the other player’s ants. Ants can also “fight” over resources, which sets up a feedback loop that drives the action of the complex system. Through gameplay players learn

about (a) ant colony behavior and (b) agents interactions in the complex system, (c), those agents properties.



Figure 1: A screenshot of the *Ant Adaptation* tabletop game. Each player can manipulate parameters that control ant properties (e.g., size, aggression level, etc.) .

We will show how we are using MMLA and qualitative methods to analyze users' interactions with *Ant Adaptation*. As shown in Figure 2, we used Social Signal Interpretation (SSI) (Wagner, Lingenfelser, Baur, Damian, Kistler, & Andre, 2013) to collect synchronized video, audio data. We processed this data into transcripts for Constructivist Dialogue Mapping and individual videos of participants for affective state tracking. We supplemented this data with field notes of user interactions and semi-structured interviews.

Analysis Pipeline

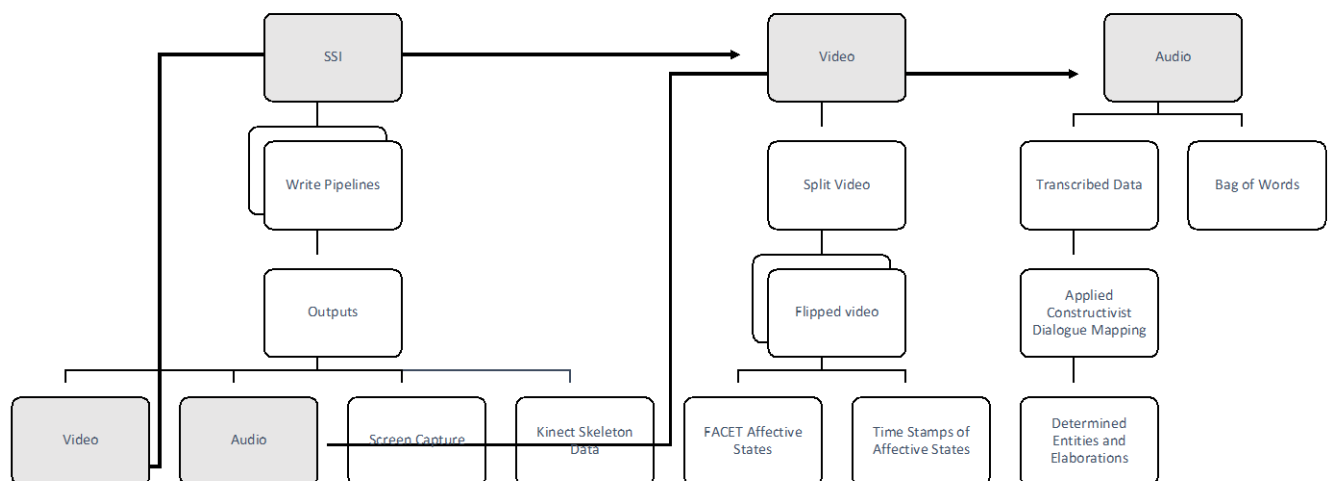


Figure 2: Data Analysis Pipeline for computer augmented ethnography

In this paper we present data collected from one dyad to highlight how this approach might be beneficial for analyzing this type of interaction. The dyad played the game side-by-side on a 52" 3M

touch screen display. A researcher interviewed the pair before and after play to understand their evolving understanding of the ant colony life cycle.

2.1 Data Analysis

Multiple coders followed constructivist dialogue mapping (Martin, 2018; Martin, Horn, & Wilensky, 2018) to analyze conversations during the pre-post interviews and gameplay conversations. We used affective state tracking to record participants' facial expressions through the interaction to guide further interaction. We use these data streams to describe and analyze the moments of high affective states as 'windows into learning' during the interaction.

2.2 Cognitive Mapping of Learning Concepts

In our data, we define learning as players' elaborations of ants, their life cycle, and functions their behavior serves. For example, if initially a participant says "ants walk," but after playing says "ants walk along paths other ants lay down towards food, pick it up and return it to the colony to feed themselves and nestmates," we would count this as an elaboration of their understanding. Each coder created a concept map for transcripts of the interviews before gameplay, during gameplay, and in the interview after gameplay. Each map consisted of entities, functions, and sub-functions discussed by the dyad. Entities included different agents and resources in the game, such as ants and flowers. Functions are the processes that engage entities, such as leaving a trail. Sub-functions are the motivations or results for functions, such as collecting food or directing the paths of other ants. Coders had a 96% inter-rater reliability on the cognitive maps they coded.

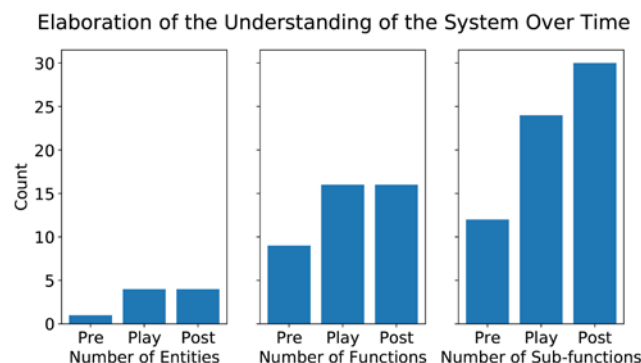


Figure 3: A cumulative plot of how many entities, functions, and sub-functions the dyad verbalized during the pre-gameplay interview, gameplay, and post-gameplay interview.

2.3 Affective State Tracking to Identify Moments of High Stimulation

We used FACET (Taggart, Dressler, Kumar, Khan and Coppola, 2016) to extract the strength of detected emotions. To identify potential emotional moments that might be associated with moments of learning (McGaugh, 2004), we used a 2.5th and 97.5th percentile threshold to reveal peaks and valleys of joy values. The percentile thresholds are visualized as horizontal line boundaries in Figure 4. This search for peaks and valleys is informed by McGaugh (2003)'s previous work, which argued high stimulus periods lead to higher memory encoding. We then extracted the dialogue from the transcript that occurred approximately ten seconds before each peak or valley to confirm and further analyze whether learning occurred at these time segments by cross-comparing to the

cognitive map data. For use these moments are potential moments to investigate, but we are cautious in categorizing them based only on the affective state, because performative acts such as facial expression can come about for many reasons.

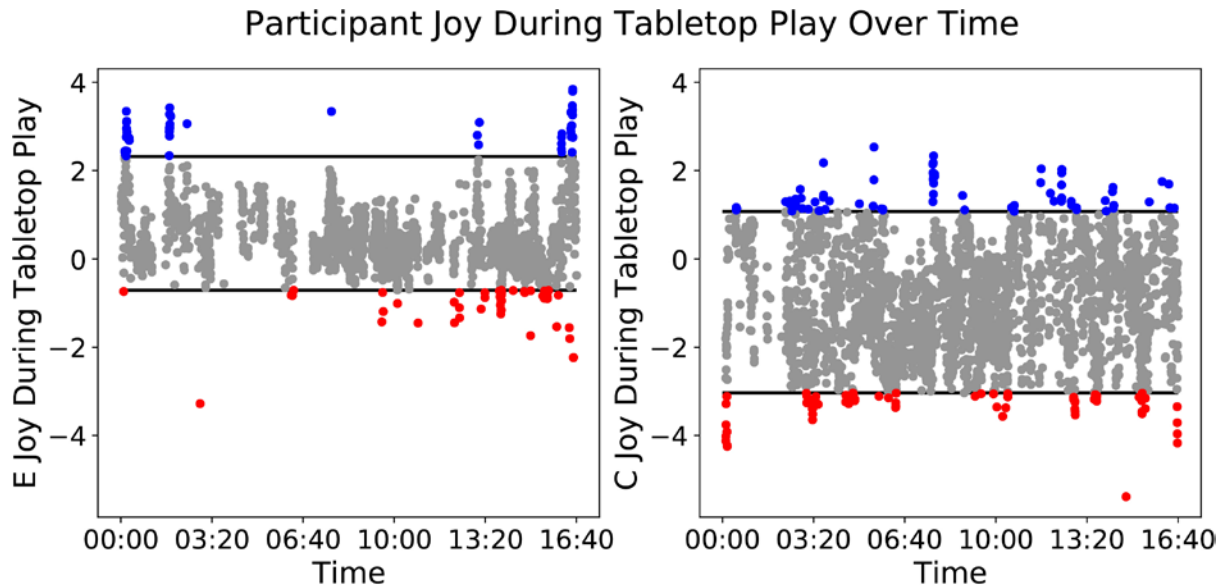


Figure 4: A visualization of participant joy over time. Emotional peaks and valleys outside of the middle 95% of the data are colored.

3 RESULTS

We present our current findings from using this method to analyze learning windows. First, we present our findings from the cognitive map coding and then describe two findings that affective state mapping reveals. We used the regions of high joy to find two parts of the conversation that we either did not initially identify as important in our transcripts and field notes, or that shed light on other learning processes than we were coding for during the cognitive mapping process. We see this interlacing of human coders and computational techniques as a useful process that both utilizes the power of computation, but also recognizes the deep and fruitful history of ethnography.

3.1 Cognitive Maps

As shown in Figure 3, participants elaborated their understanding of ants, their life cycle and the function of behaviors. Before the gameplay began, the focal dyad presented one entity, “ants”, with nine functions, such as “placing paths to food”, and twelve sub-functions, such as “following” pheromone trails. During and after gameplay, the participants elaborated their understandings. During gameplay, players mentioned three additional entities (flowers, queen ants, and a GUI element) as well as sixteen functions, such as ants hiding in their colony, or organizing society. They also discussed twenty-four sub-functions (e.g., food collection through leaving attractant pheromones). By the end of the intervention, the participants expanded to thirty sub-functions. In other words, most of the concepts learned (75% of the entities and 40% of the sub-functions) were emergently elaborated on during play. While cognitive mapping provides a structured set of codes about participant sensemaking, one key limitation is that it extracts insight solely from transcribed audio and does not necessarily capture evidence of learning in other modalities. Next, we share two

examples of how emotion tracking extended our analysis of participant knowledge evolving throughout the session.

3.2 Possible Learning Moments for Further Investigation

This example shows how emotion tracking revealed moments of learning that we did not previously identify during qualitative coding. In Figure 4, we noticed a peak joy value with participant E and a valley joy value with participant C at timestamp 00:04, showing that this is a high-stimulus moment for both participants. Revisiting the transcript, this moment corresponded with when the facilitator described how to play the game: "On your side of the board, you have three sliders. One of them changes the size of new ants formed, one of them changes the aggressiveness." This evidence of learning during instruction was unexpected because when coding interview transcripts, the analyst cannot confirm whether participants are learning when they are not verbalizing. Given our interest in informal learning that occurs in museums through games, as analysts we primarily focus on the spontaneous, emergent learning that occurs rather than learning that happens during instruction. As such, during the cognitive mapping analysis, we did not identify learning during the moments when the facilitator instructed participants about the game. Ultimately, emotion tracking enabled us to detect a possible learning moment when participants were silently listening to instructions on how to play the game. Given that peak joy values do not guarantee that learning occurs, evidence in another modality or a follow-up question in interviews is currently necessary to confirm that this was a learning moment. In future implementations, we could use this experience to design further probing questions or capture additional data streams at these moments to identify what sort of learning might be happening. We see the potential to augment qualitative coding beyond what can be detected with a single analytic frame and transcription techniques.

3.3 Joy During Previously Identified Learning Moments

In contrast to the previous example, the following example shows how emotion tracking provided additional insight on a learning moment already identified by human coders in the cognitive mapping analysis. Looking at Figure 4 at timestamp 7:00, E and C have peak joy values of 3.8 and 1.8-2.2, respectively. We revisited the video at this segment for further analysis. C and E verbalized that the flowers at close proximity to the ant hill increases their ants' population. C says, "Can I have more flowers?" E responds, "Yes. Ring of flowers." They place a ring of flowers around their ant hills and notice ants picking up the food. Both watch the tabletop intently to observe the resulting ant behavior and C says, "Ooh, now I've got lots of ants." The dyad discovered a powerful relationship they can use to manipulate the environment.

As the dyad continues tinkering with parameters, they laugh through their trial and error attempts, and elaborated on the concept that food close to the nest increased the population of ants. This moment was coded in the cognitive mappings as one of the flowers' primary functions, and our emotion analysis adds additional understanding to this moment. That is, this discovery led to a sense of joy or what might be interpreted as "satisfaction" which is important to cultivate in informal learning environments and gameplay. Though we selected many of the same moments using emotion logging as we did through manual cognitive mapping, emotion logging also drew more attention to specific moments. In this example, our multimodal data identified an opportunity to specifically evaluate design decisions that affected the learning of participants. In other words, the

approach both reinforced our prior units of analysis and added to our approach to analyzing the interaction. We aim to continue this back-and-forth between qualitative methods and computational techniques as we collect data on more dyads and extract insights from other modalities.

4 CONCLUSION

While we have cognitive mapping data from over 100 participants, we think it is worthwhile to share our methods with the community. Computational techniques have a great deal of promise in augmenting ethnographic practice. In this workshop we would like to share how we are integrating artificial intelligence with human analysis to understand learning in and around museum games. These methods are beginning to become useful tools to explore learning moments captured computationally and ethnographically.

5 ACKNOWLEDGEMENTS

We would like to thank the members of the Center for Connected Learning, the Technological Innovations for Inclusive Teaching and Learning lab, and the Inclusive Technology Lab for their thoughtful comments and support. Additionally, we would like to thank our advisors Uri Wilensky and Anne Marie Piper for their support. Lastly we would like to thank the generous support of the research by IEF, Multidisciplinary Program in Education Sciences (IES: Award # R305B090009).

REFERENCES

- D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, 52(2), 104-123.
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.
- Howell, N., Chuang, J., De Kosnik, A., Niemeyer, G., & Ryokai, K. (2018). Emotional Biosensing: Exploring Critical Alternatives. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 69.
- Martin, K. (2018). *Constructionist Dialogue Mapping: A means to assess informal learning*. Visitor Studies Association, Chicago, Illinois.
- Martin, K., Horn, M., & Wilensky, U. (2018). *Ant Adaptation: A complex interactive multitouch game about ants designed for museums*. Constructionism 2018 Conference, Vilnius, Lithuania.
- McGaugh, J. L. (2003). *Memory and emotion: The making of lasting memories*. Columbia University Press.
- McGaugh, J. L. (2006). Make mild moments memorable: add a little arousal. *Trends in cognitive sciences*, 10(8), 345-347.
- National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. National Academies Press.
- Ochoa, X., & Worsley, M. (2016). Augmenting Learning Analytics with Multimodal Sensory Data . *Journal of Learning Analytics*, 3(2), 213-219.

- Oviatt, S., Cohen, A., & Weibel, N. (2013, December). Multimodal learning analytics: description of math data corpus for ICMI grand challenge workshop. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 563-568). ACM.
- Schneider, B., & Blikstein, P. (2015). Unraveling students' interaction around a tangible interface using multimodal learning analytics. *Journal of Educational Data Mining*, 7(3), 89-116.
- Worsley, M., Scherer, S., Morency, L. P., & Blikstein, P. (2015, November). Exploring behavior representation for learning analytics. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 251-258). ACM.

Using Multimodal Analytics to Analyze Family Interactions in a “Making” Activity

Melissa Perez

Northwestern University
melissaperez2019@u.northwestern.edu

Kira Furuichi

Northwestern University
kirafuruichi2019@u.northwestern.edu

Stephanie Jones

Northwestern University
stephaniejones2023@u.northwestern.edu

Sarah Lee

Northwestern University
spl@northwestern.edu

Kya Suzuki

Northwestern University
kyasuzuki2019@u.northwestern.edu

Marcelo Worsley

Northwestern University
marcelo.worsley@northwestern.edu

ABSTRACT: “Making” refers to the process of tinkering, hacking, creating art, etc., through which the maker “creatively design[s] and build[s] projects for both playful and useful ends”, often including modern digital fabrication tools such as 3D printers, laser cutters, etc., in the process (Martin, 2015). In this study, families came together to collaboratively “make” their own board game based off of shared family interests or stories, using both traditional arts and crafts and digital fabrication tools. Utilizing indoor positioning, electrodermal activity, and video data, we perform initial network, location, and engagement analyses. While many challenges exist for using multimodal data collection for rigorous analysis, we find that it is still an appropriate methodology for identifying interesting making and collaborative moments.

Keywords: Multimodal Learning Analytics, Intergenerational Making, Network Analysis

1 INTRODUCTION

In this paper, we present the results of a multimodal analysis of a maker activity for families. The objective is to use techniques from Multimodal Learning Analytics (MMLA) to quantify learning as it

takes place in an informal learning environment, and to see what insights can be gained from an MMLA approach.

The analysis comes from the implementation of an activity that engaged families in making board games together. An example of significant prior work with families engaging in making is the Family Creative Learning program. During this program, families came together to build projects using MakeyMakey and Scratch, two common maker technologies (Roque, 2016). This program was meant to create an informal learning space in which families can collaboratively engage in maker activities that promote wider engagement amongst low-income families, driven by the maker technologies presented to them. Our program expands work in engaging families in making with an approach that is technology agnostic. Therefore, the focus is not on teaching technology, but rather facilitating engagement and relationships using tools that vary in technological complexity. To look at this, we designed a program with two parts. The first part utilized traditional arts and crafts tools while the second part incorporated digital fabrication tools. We were interested in the ways digital fabrication tools shift or alter family dynamics and the roles taken up during the activity.

In this paper, we focus on two primary modes of analysis: indoor positioning (IP) and electrodermal activation (EDA). IP has been used to explore how children interact within participatory simulations, specifically looking at paths and delineating the play space of children during an activity (Peppler et al., 2018). One technology that is commonly used for IP is ultrawide-band radio wave. This high frequency band ranges from approximately 3 to 10 GHz and has been validated to provide centimeter level accuracy (Karbownik et al., 2015). EDA refers to the electrical potential on the skin's surface. EDA measures fluctuations in the sympathetic nervous system driven by stress, physical or mental exertion, and more (Boucsein, 1992). Increases in the sympathetic nervous system ultimately drive perspiration production, motivating the use for EDA to be analyzed in situations of cognitive and emotional load (Boucsein, 1992). Furthermore, skin conductance has been tied to helping identify key events within team activities, allowing for a clearer understanding of team dynamics and teamwork (Ahonen, 2018).

2 FAM JAM!

The specific context that we are examining is a workshop we entitled "Fam Jam!". Fam Jam! took place on a Saturday morning in a university campus lab. The lab is outfitted with a variety of fabrication tools and machines typically found within makerspaces. The session lasted from 09:00 to 13:00. Breakfast and lunch were provided for participants. Participants were recruited from administrative staff in the computer science department at the university and consisted of four families. There were a total of eight children (ages ranging from 2-13) and five adults. During the session, the families participated in three different phases of interacting. In the first phase, the families played board games together. In the second phase, they brainstormed and began to make their own games using traditional arts and crafts materials (i.e., construction paper, pipe cleaners, etc.), and in the final phase, the families constructed their games using digital fabrication tools (3D pens, laser cutters, Chibitronics, etc.). Following the completion of their games, families ate lunch and presented their games to the other participants.

Before starting on the making part of the activity, families were immersed in family game play with a variety of popular American board games. After playing games for approximately thirty minutes, the research staff provided learning prompts intended to spur ideation, creativity and careful reflection.

3 DATA COLLECTION

We used four data streams for this study: audio, video, position, and physiological. In following with prior work in MMLA (Worsley et al., 2016), we used these different modalities to enable use to capture that various forms of engagement and interaction and participants might exhibit. Furthermore, data collection devices were distributed throughout the learning space, as to ensure that all participants contributions and interactions were successfully captured. The floor plan and physical data collection placements are shown in Figure 1.

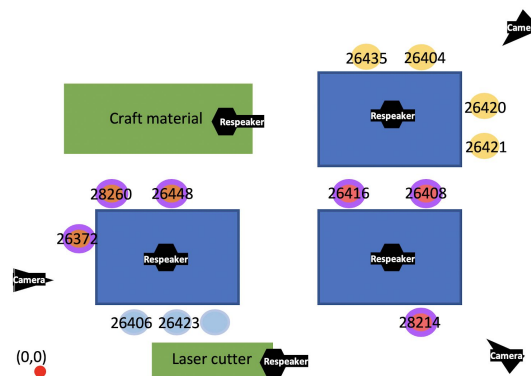


Figure 1: Room layout with data collection placement. Circles of the same color represent family members, and those with outlines are participants who are also wearing Empaticas

3.1 Indoor Positioning (IP)

Position data was recorded using the Pozyx system, which utilizes ultrawide-band radio wave positioning to get accurate location coordinates in a set coordinate grid system. Participants were given fanny packs with Pozyx tags to wear for the duration of the program. This allowed us to capture their location and movement throughout the space. The x, y, z coordinates at any given time of each participant was recorded to a file so that the data could later be analyzed. This data was validated by referencing the video recordings of the space and identifying when the positions being recorded matched the position of the people in the room. An example can be seen in Figure 2, with connections in Pozyx data (right) being based on proximity. In the left image of Figure 2 is the video data; green dots are mislocated compared to the estimation given by pozyx data (right). The red dots are in the general expected position, and blue dots are not detected by Pozyx.



Figure 2: Tag position validation.

3.2 Audio/Video

Audio was recorded using the Respeaker Core V2, a 6 microphone array that provides information about the direction of arrival. A Respeaker was placed at each table where a family unit was working, with an additional speaker at the supplies table where families picked up materials (Figure 1). Video cameras were placed above each family unit and captured both video and audio data during the duration of the session (Figure 1).

3.3 Physiological

Empatica E4 wristbands captured physiological (electrodermal activity; skin conductance [SC]; galvanic skin response, temperature, accelerometer, heart rate, and internal beat interval) data. Two of the four family units wore E4 wristbands for the duration of the program. Our primary motivation in collecting physiological data is to better understand how SC levels differentiate during the two periods of making with arts and crafts supplies and making with digital technology, as well as differentiating between the phasic and tonic components of participants' SC levels.

4 ANALYSIS

Preliminary results surface different analytic potentials for IP and EDA analysis to track interactions between family groups before and after the addition of digital fabrication tools. The analyses also support the examination of overall participant interactions within the space and with materials. Using the IP data, we found: 1) betweenness centrality and degree using two different distances, 2) total distances moved during different phases of the activity, and 3) a positional heatmap. The EDA data provided a means to consider the hard to detect physiological responses that participants exhibited during the course of the workshop. Concretely, this involved looking for changes in skin conductance response, and skin conductance level. The first refers to momentary spikes in EDA data, while the other looks at much more gradual increases in EDA.

4.1 Network Analysis

Using the Python networkx library, we performed a network analysis on the data. This particular analysis is based on all of the data in aggregate, and is included here to demonstrate a potential utility of IP. Networks are based off of the distance between participants during the session.

4.1.1 Betweenness Centrality and Degree

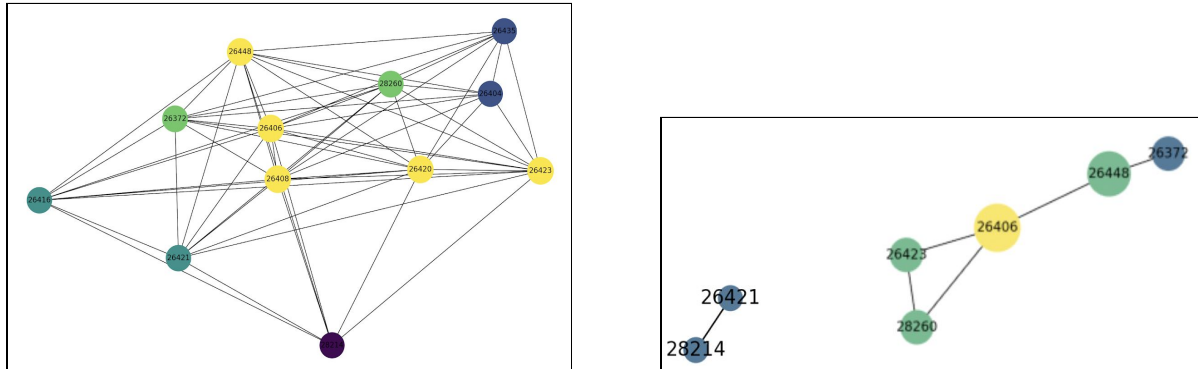


Figure 3: Participant Networks; left: distance= 2,000mm; right: distance=750mm

Due to data loss in certain tags, the networks were formed by using the tag with the most data points as a delimiter for updating the positions of each of the tags at any given time. After being collapsed into networks (i.e. at a given snapshot in time) connections were formed based off of the pairwise distance between participants. The distances used were determined first by when participants were within the same general area of the room (Figure 3 left), and then by when participants were in close proximity (Figure 3 right).

The resulting networks in Figure 3 demonstrate the betweenness centrality and degree of each of the participants based off of these distances. Yellow points have the highest degree and betweenness centrality, with the more purple dots being the least connected. Looking at the connectedness of each of the participants gives insight as to how participants are interacting across families and within.

4.1.2 Distances traveled during different activities

Using IP, we also look at the average distances traveled during the different phases of the program; playing games, traditional making, and digital fabrication based making. The data can be seen in Table 1.

Table 1: Average distance traveled for different activities

Activity	Avg Distance Traveled
Playing games	2943.61 mm
Traditional making	3463.48 mm

Digital fabrication

2707.44 mm

The data shows that each of the activities did have different levels of distance traveled, with the traditional making portion of the session having the highest distance traveled, and the digital fabrication portion having the lowest. The game play portion did not have the most movement, due to the nature of board games causing families to staying in place. The decrease in movement in the digital fabrication portion could be explained by the fact that the participants did not have to move around to reach the digital fabrication tools, as the facilitators brought over most materials. Additionally, participants would often crowd around the laser cutter or around a given group's game to ask questions, or simply be excited by the technology.

4.2 Position Histogram

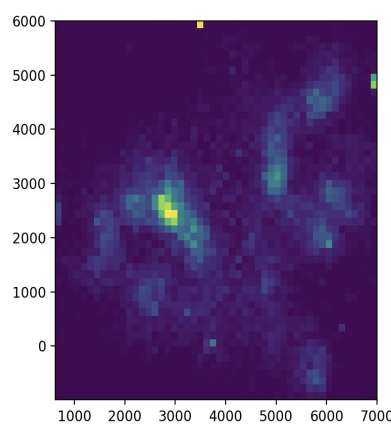


Figure 4: Location histogram of participants movement

Using just the x, y coordinates of each of the participants, the most highly trafficked spots in the room can be seen in Figure 4; the actual layout of the room can be seen in Figure 1. The data was first cleared of points with (0, 0) coordinates, as that position was tagged when there was no available data for a person. Then, using the Python numpy library two-dimensional histogram feature, each of the points was hashed to a bin (50 bins for Figure 4). The bins with the highest number of tallies appear yellow on the diagram.

From the histogram, the overall positions of people around the space can be seen. The movement around the tables emerges similarly to what we would expect given the constraints of the room, with the highest tracked area being close to where people moved to and from the craft table.

4.3 EDA

Six participants (two families) were provided with Empatica E4 sensors. Due to participants turning the Empatica E4s on and off during the session, only three of the six sensors captured data. From these three participants, initial peak and noise analyses were run using an automatic detection of artifacts analyzer (Taylor, 2015). In addition, from these three participants, only two captured data

during the entire duration of the session. Due to the limited dataset, we are unable to correlate EDA responses to the cognitive and emotional load from the session.

5 DISCUSSION

Preliminary analyses based on IP data provide a glimpse into the ways that individuals were physically engaged within this collaborative making project. For future Fam Jam! sessions, we intend to utilize orientation data from the Pozyx tags, which would enhance our understanding of how the participant is moving in the space (i.e., standing still vs. moving around). Moreover, as opposed to just knowing their location within the space, we will be able to better ascertain the direction they are facing. This is crucial in relatively small spaces, where two people might be located next to each other but facing the opposite direction. In the current analysis, two such individuals would be recorded as “collaborating” when, based on observation, this was not always the case.

In addition, future work will also look to correlate EDA with IP data, to better understand how the layout of “making” spaces differs by EDA levels. For example, are there specific areas within the physical space that correlate with high levels of electrodermal activation. Put differently, one might find that it is in the presence of certain individuals that other participants experience high electrodermal activation. Either of these situations could point to technologies or people that seem to spur increased engagement, cognitive load, or emotional response. Regardless, having such information would be beneficial for better understanding and supporting these types of learning environments.

The challenges of missing data is something else that we wish to address with future work. This is in regard to both data collection and data processing. Ideally, we would be more cognizant of data collection challenges in the moment, while also able to reconcile for this missing data in the analyses.

Lastly, this session provided insight into the challenges that remain for the data integration of multimodal analysis. While there are technologies and tools available for certain multimodal data sources, the final integration and synchronization of physiological, video, audio, location, and other data still remains a challenge.

6 CONCLUSION

In this paper we presented preliminary analyses based on indoor position tracking and electrodermal activation. These two modalities are becoming increasingly accessible to researchers and practitioners who wish to explore complex learning environments through multimodal data. The analyses presented provide a glimpse into what may be possible with these types of tools, as well as a few lessons learned and potential pitfalls for utilizing these data streams.

REFERENCES

Ahonen, L. (2018). Biosignals reflect pair-dynamics in collaborative work: EDA and ECG study of pair-programming in a classroom environment. *Scientific Reports*, 3138, 8.

- Boucsein, W. (1992). *Electrodermal Activity*, Plenum Series in Behavioral Psychophysiology and Medicine, Plenum Press.
- Karbownik P., Krukar G., Shaporova A., Franke N., & von der Grun N. (2015). Evaluation of Indoor Real Time Localization Systems on the UWB Based System Case. *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*
- Martin L. (2015). The Promise of the Maker Movement for Education. *Journal of Pre-College Engineering Education Research (J-PEER)*: Vol. 5: Iss. 1, Article 4.
<http://dx.doi.org/10.7771/2157-9288.1099>
- Peppler, K., Thompson, N., Danish, J., Mozek, A. & Han, S. (2018). Indoor positioning technology & enhanced engagement in early elementary systems thinking and science learning. In J. Kay & R. Luckin (Eds.), *Rethinking learning in the digital age: Making the Learning Sciences count: The International Conference of the Learning Sciences (ICLS) 2018* (Vol. 3, pg. 1077-1080). London, UK: International Society of the Learning Sciences. ISBN: 978-1-7324672-2-4
- Roque, R. (2016). Family Creative Learning. In Peppler, K., Kafai, Y., & Halverson, E. (Eds.) *Makeology: The maker movement and the future of learning*. New York, NY: Routeledge.
- Taylor, S., Jacques, N., Chen, W., S., Sano, A., and Picard, R. (2015). Automatic Identification of Artifacts in Electrodermal Activity. In *EMBC*.
- Worsley, M., Abrahamson, D., Blikstein, P., Grover, S., Schneider, B., & Tissenbaum, M. (2016). Situating multimodal learning analytics. In *12th International Conference of the Learning Sciences, ICLS 2016: Transforming Learning, Empowering Learners, Proceedings* (Vol. 2, pp. 1346-1349). International Society of the Learning Sciences (ISLS).

Exploring Different Software Platforms for Multimodal Learning Analytics

Daniel Spikol

Internet of Things and People
Malmö, University
daniel.spikol@mau.se

ABSTRACT: Multimodal Learning Analytics (MMLA) provides diverse challenges across technology and research design on how best to design research interventions, technical infrastructure, collection, and analysis of the diverse data. Just from the underlying software engineering perspective of maintainability, dependability and security, efficiency, and acceptability most MMLA systems do not meet professional standards. This paper aims to explore how do we define and begin to create a series of MMLA systems that begin to meet standards beyond one-off research projects. Also, the purpose of the tutorial is to discuss and demo several systems that show promise.

Keywords: Multimodal Learning Analytics, Software Platforms, Robotics, IoT

1 INTRODUCTION

Collaborative problem solving (CPS) is a fundamental skill in modern society, and it is gaining much and much attention across education systems around the globe. CPS is crucial in many constructivist learning approaches, such as problem-based learning, inquiry-based learning, project-based learning and practice-based learning. It is a prevailing opinion that constructivist teaching approaches can bolster some of the needed skill of modern society because learners closely collaborate to solve a specific task. For many years, this approach received strong appreciations (Barron, Walter, Martin, & Schatz, 2010) and Cukurova and colleagues (2016) presented an original framework to identify observable and objective differences in students Collaborative Problem Solving (CPS) behaviours in open-ended, practice-based learning environments.

One approach to providing these new skills is to create opportunities for learners to work in unison to solve complex problems in socially interactive rich learning environments. However, appropriate monitoring, support, and feedback for students in these learning environments are essential for their success (Spikol, Ruffaldi, & Cukurova, 2017). Providing appropriate support for each student in CPS is an extremely challenging task for teachers. However, new advances and methods in learning analytics research, particularly Multimodal Learning Analytics (MMLA), offer novel methods that generate characteristic information about what happens when students are engaged in these activities (Worsley, 2014).

However, MMLA provides diverse challenges across technology and research design on how best to design research interventions, technical infrastructure, collection, and analysis of the diverse data. Just from the fundamentals of software engineering perspective of maintainability, dependability and

security, efficiency, and acceptability most MMLA systems do not meet professional standards yet (Sommerville, 2016). This paper aims to explore how do we define and begin to create a series of MMLA systems that begin to meet standards beyond one-off research projects. Also, the purpose of the tutorial is to discuss and demo several systems that show promise.

2 BACKGROUND

These challenges are recognised in the MMLA community and work done by different research projects is ongoing. The Multimodal Learning Hub (MLH) is a notable project that addresses and investigates how to enhance learning in ubiquitous environments by collecting and integrating multimodal data from different data sources (Schneider, Di Mitri, Limbu, & Drachsler, 2018). Additionally, the work builds upon a conceptual model for MMLA (Di Mitri, Schneider, Specht, & Drachsler, 2018) that creates a working taxonomy to support the technical development. The MLH project is still under development.

Shankar and colleagues (2018) review the MMLA architectures and highlight the design tensions between the different architectures across research. They classify these tensions across three main issues, the role of learning specific constructs in data organisation, flexibility and extensibility of architectures, and the need for the simple of interfaces. Additionally, the article raises the point about the more widespread adaptation of these emerging software and the lack of standards about the use of xAPI and LRSs.

Worsley (2018) literature survey is complementary to Di Mitri's (2018) conceptual model of MMLA. Rather than present a taxonomy, past, present, and future features required for MMLA are explored. The paper challenges the community to consider what directions we need to investigate and develop to grow the field. A substantial amount of work over the last several years has gone into making MMLA accessible to researchers and practitioners, however, as a community, we still need to develop software systems and data standards at a larger scale. Additionally, we need to investigate other fields beyond LA that may offer solutions and inspiration.

3 SOFTWARE PLATFORMS

The following section presents three groupings of different approaches that seem warranted for further investigation of how to design and develop a scalable and robots MMLA system. We started with social signal and situated intelligence approaches that investigate human behaviour and how to support interaction between people and interactive systems. The second approach is to build upon the social robotics community, considering, in the end, the system needs to understand what happens between people through the use of different sensors. The third approach is through IoT systems that combine different sensors though cloud computing that executes functions in response to events. The following is not a systematic approach to different systems, but rather the first step for exploration what types of systems might be relevant for experimentation.

3.1 Social and Situated Systems

The Social Signal Interpretation (SSI)¹ framework offers tools to record, analyse and recognise human behaviour in real-time, such as gestures, mimics, head nods, and emotional speech (Wagner et al., 2013). Following a patch-based design, pipelines are set up from autonomic components and allow the parallel and synchronised processing of sensor data from multiple input devices. SSI supports the machine learning pipeline in its full length and offers a graphical interface that assists a user to collect their training corpora and obtain personalised models. In addition to a large set of built-in components, SSI also encourages developers to extend available tools with new functions. Human-centred Multimedia Group is developing the project at the Department of Computer Science at Augsburg University.

Recently Microsoft Research has begun to research how to develop a platform for Situated Intelligence (PSI)² an extensible framework intended to enable the rapid development, fielding and study of situated, integrative AI systems. They define the term situated for their framework to target systems that sense and act in the physical world, that includes a broad class of applications, including various cyber-physical systems such as interactive robots, drones, embodied conversational agents, personal assistants, interactive instrumented meeting rooms, software systems that mesh human and machine intelligence and so on. The platform provides an infrastructure, a set of tools and an ecosystem of reusable components that aim to mitigate some of the challenges that arise in the development of these systems. The primary goal is to speed up and simplify the development, debugging, analysis, maintenance and continuous evolution of integrative systems by empowering developer-in-the-loop scenarios and rapid iteration.

3.2 Robotics

The Robot Operating System (ROS)³ is a flexible framework for writing robot software that has supported the development of robotics systems over the last 10+ years. ROS is a collection of tools, libraries, and conventions that aim to simplify the task of creating complex and robust robot behaviour across a wide variety of robotic platforms. What is relevant about ROS is that it has a vibrant community and some extent can be seen to address some of the same issues as MMLA if we consider the learning environment what the robot needs to perceive. One of the software architectural strengths of ROS is the module approach allowing new sensors and components to easily added (Quigley et al., n.d.).

For instance, a relevant approach for MMLA could be to use the Online Multimodal Interactive Perception (OMIP)⁴ is a framework to exploit the interaction capabilities of a robot to reveal and perceive its environment. The information to perceive this environment is contained in the high

¹ <https://hcm-lab.de/projects/ssi/>

² <https://github.com/microsoft/psi>

³ <http://www.ros.org/>

⁴ <https://github.com/tu-rbo/omip>

dimensional, continuous, multimodal sensor stream of the robot. OMIP offers tools to interpret this stream based on prior knowledge encoded into recursive estimation loops. The prior knowledge used for perception includes physics laws (rigid body physics, kinematics, ...) and knowledge about the correlation between robot actions and responses of the environment that could be applied in an MMLA context.

3.3 IoT Frameworks

A different approach than the social, situated, robotics avenues is to consider IOT and how these systems deal with the notion of Emergent Configuration which consists of a dynamic set of things, with their functionalities and services, that cooperate temporarily to achieve a goal (Alkhabbas, Spalazzese, & Davidsson, 2017). MMLA can adopt this approach by exploring different IoT platforms like Things that Speak⁵, the prototype frameworks that allow rapid prototyping. Additionally, Apache OpenWhisk⁶ an open source, distributed serverless platform that executes functions (fx) in response to events at any scale would be a viable approach to exploring how IoT could manage the infrastructure, servers and scaling using Docker containers.

4 CONCLUSION

The current work on MLH (Schneider et al., 2018) begins to address some the software architecture of the social, situated, and the IoT approaches. However, different efforts need to go forward that explore how we can create several systems that could address specific learning contexts, different types of sensors, and data inputs while having some common standards for LA (for example xAPI) that allows ease of use for different research efforts and real-world interventions. This proposal aims to try to broaden the MMLA community's approach, begin to identify requirements that are relevant and to investigate different software platforms that could be used to somewhat rapidly prototype and test out different scenarios. However work needs to continue on different efforts, before the community can see any adoption and verification of our results.

Developing MMLA software platforms in-line with software engineering standards needs to be part of the priorities for our field. Currently, different approaches and development are projects centred resulting in a diverse approach to data standards, platforms, and sensors. This diversity is a good thing for research, however, we need to have a larger aim to create some guidelines for intra-operability. Unlike, the broader field of learning analytics which uses primarily clickstream data from LMS systems that have developed some basic standard, MMLA is finding its footing. If we start with the basic principles of software engineering that use specifications, development options, validation, and evolution as the foundation, we can begin to set a course for our projects to work towards a common goal. The first place is for our community to start that we can expand to different stakeholders that will benefit from our CrossMMLA approaches.

⁵ <https://thingspeak.com/>

⁶ <https://openwhisk.apache.org/>

REFERENCES

- Alkhabbas, F., Spalazzese, R., & Davidsson, P. (2017). Architecting Emergent Configurations in the Internet of Things. In *2017 IEEE International Conference on Software Architecture (ICSA)* (pp. 221–224). IEEE. <https://doi.org/10.1109/ICSA.2017.37>
- Barron, B., Walter, S. E., Martin, C. K., & Schatz, C. (2010). Predictors of creative computing participation and profiles of experience in two Silicon Valley middle schools. *Computers & Education*, 54(1), 178–189. <https://doi.org/10.1016/j.compedu.2009.07.017>
- Cukurova, M., Avramides, K., Spikol, D., Luckin, R., & Mavrikis, M. (2016). An analysis framework for collaborative problem solving in practice-based learning activities. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 84–88). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2883851.2883900>
- Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4), 338–349. <https://doi.org/10.1111/jcal.12288>
- Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., ... Ng, A. (n.d.). *ROS: an open-source Robot Operating System*. Retrieved from <http://stair.stanford.edu>
- Schneider, J., Di Mitri, D., Limbu, B., & Drachsler, H. (2018). Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences (pp. 45–58). Springer, Cham. https://doi.org/10.1007/978-3-319-98572-5_4
- Shankar, S. K., Prieto, L. P., Rodriguez-Triana, M. J., & Ruiz-Calleja, A. (2018). A Review of Multimodal Learning Analytics Architectures. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 212–214). IEEE. <https://doi.org/10.1109/ICALT.2018.00057>
- Sommerville, I. (2016). *Software engineering*. Boston: Pearson.
- Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning. Philadelphia, PA: International Society of the Learning Sciences.
- Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013). The Social Signal Interpretation (SSI) Framework Multimodal Signal Processing and Recognition in Real-Time. *Proceedings of the 21st ACM International Conference on Multimedia*. <https://doi.org/10.1145/2502081.2502223>
- Worsley, M. (2018). Multimodal Learning Analytics' Past , Present , and , Potential Futures (pp. 1–16). Retrieved from <http://ceur-ws.org/Vol-2163/paper5.pdf>
- Worsley, M. (2014). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. *3rd Multimodal Learning Analytics Workshop and Grand Challenges, MLA 2014*, 1–4. <https://doi.org/10.1145/2666633.2666634>

Workshop: Innovative problem solving assessment with learning analytics

Lishan Zhang

Beijing Normal University
lishan@bnu.edu.cn

Baoping Li

Beijing Normal University
libp@bnu.edu.cn

Yigal Rosen

ACT Inc.
Yigal.rosen@act.org

Kristin Stoeffler

ACT Inc.
Kristin.stoeffler@act.org

Shengquan Yu

Beijing Normal University
yusq@bnu.edu.cn

ABSTRACT: Solving dynamic and ill-structured problem is one of the most important skills for the 21st century. In addition, people often need to collaborate to solve problems together in real-life. Therefore, it is important to establish the assessments for evaluating both individual and collaborative problem solving ability for K12 students to ensure that the students are ready for dealing with real-life problems when they leave schools. To achieve this goal, learning scientists have designed various simulations to implement interactive and dynamic assessments. On the other hand, some techniques of learning analytics such as regression model, neural network, and hidden markov model have been used to analyze problem solving procedures. The workshop aims for further exploring how learning analytics could facilitate both of individual and collaborative problem-solving assessment through presentation, interactive event and roundtable discussion among the researchers with different backgrounds but the same interest.

Keywords: problem solving assessment, collaborative problem solving, simulation, log file analysis.

1 BACKGROUND

Regardless of their occupations, people need to handle and solve different types of problems every day. Problem solving is the process of finding a method to achieve a goal from an initial state. However, real-life problems are usually ill-structured without clear goals and givens, so cannot be solved in a routine manner. Knowing how to solve problems in real-life situations has become an

essential skill for the 21st century (Griffin, McGaw, & Care 2012; Greiff et al. 2014). The assessment of problem solving skill has some fundamental differences with the traditional assessment of curriculum content knowledge. Problem solving assessment has to be able to successfully assess students' abilities in dealing with ill-structured and dynamics environments. It requires the assessment environment should change upon students' behaviors and responses. Traditional static and paper-based assessment clearly fails to do so. The existing problem solving assessments usually provide students the dynamics situations by implementing a series of simulations (Zhang, Yu, Li, & Wang 2017; Schweizer, Wüstenberg, & Greiff 2013). Then problem solving performance is evaluated in terms of students' outputs in the simulation. Although students' behaviors, also called as process data, are usually logged and analyzed as well, the analysis on the process data is still very limited. Aggregated measures like time, number of clicks are often used to profile problem solving procedures. Few studies identified simple problem solving strategies from the log files (Zhang et al. 2014; Greiff, Niepel, Scherer, & Martin 2016).

Besides individual problem solving assessment, researchers started to emphasize collaborative problem solving assessment in the recent years. Collaboration is a long-standing practice in many environments, and people often needs to collaborate to solving real-life problems (Wilson, Gochyyev, & Scalise 2017). Because collaborative problem solving has to happen with at least two participants, a participant's collaborative problem solving performance is highly affected by the collaborators. The assessment of the skill faces reliability issue. Some researchers solved the issue by creating simulated agents, also called as avatars, which solve problems together with an individual (Rosen 2017). Because the behaviors of a simulated agent are well scripted in advance, only the individual, which is the test taker, can affect the collaboration. On the other hand, some researchers carefully created joint activities for multiple individuals, and managed to assess collaborative skills by analyzing their collaborative behaviors (Wilson, Gochyyev, & Scalise 2017).

Several different techniques of learning analytics have been used to analyze both individual and collaborative problem solving. In general, two types of approaches have been adopted in the analysis: (1) Aggregate problem solving behaviors into some indicators, and explore the correspondence between the aggregated indicators and problem solving outputs. Correlation analysis, supervised learning algorithms such as decision tree and neural network are used in this approach. (2) Directly analyze problem solving behaviors without aggregation. Algorithms such as hidden markov model, lag sequential analysis, association rules mining are used in this approach. Despite of the adopted analysis approaches, the problem solving actions in the log files have to be "reasonable" coded before feeding to the algorithms. The coding process can be treated as a kind of data pre-processing in typical data mining. "Reasonable" here means that the coded behaviors should be neither too specific nor too general. For example, a coded behavior is too specific if it records the specific pixel where an action occurs; a coded behavior is too general if it only records the existence of an action. Finite state machine is often used to auto code problem solving behaviors at an appropriate level of abstraction. The learning theory aligned with the problem solving assessment should guide the designs of behavior coding schemas. In this context, learning analytics can be seen as the method of transforming learning theory of problem solving into analysis results. Therefore, it is important to explore at the intersection of problem solving assessment and learning analytics.

Note that the design of the user interface for an assessment actually decides what problem solving unit actions go to the log files. Apparently, it is impossible to acquire how a student solves a problem if the student is only required to fill up the final answer of the problem. Theories from learning science and cognitive science should drive the design to ensure appropriate problem solving unit actions are recorded for the purpose of assessment. For example, the given of a problem is intended to be hidden after a series of interactions when knowledge acquiring skill needs to be assessed. The problem relevant documents are mixed with irrelevant documents in a virtual library if the skill of information identification needs to be assessed (Zhang, Yu, Li, & Wang 2017).

In summary, researchers in learning science and data analytics have to work together to develop high quality of problem solving assessment. The proposed workshop aims for inviting researchers who have interests in facilitating problem solving assessment with learning analytics from both of the two areas. The organizers will invite researchers who have previously conducted related studies to present their findings and lessons learned. Then all the workshop participants are invited to discuss together to learn from each other and explore any collaboration opportunities.

2 ORGANISATIONAL DETAILS

Type of event: Workshop

Proposed schedule, duration, type and activities: It ought to be a half-day open workshop with some invited presenters. The specific activities include presentation, interactive event and roundtable discussion.

For the presentation, the presenters should be able to cover some of the bullets below:

- Describe the problem solving skills they intend to assess, and the theoretical framework from the perspective of learning science or cognitive science.
- Explain the design of the assessment and expected behaviors of the test takers.
- How students' log files can be used for assessments.
- Discuss any case studies or experiments of problem solving assessment that have been conducted.
- Explain the techniques used for analyzing problem solving behaviors, including but not limited to neural network, dynamics Bayes network, Markov modeling, and finite state machine.
- Explain how learning scientist and data analysts can collaborate to improve problem solving assessment or understand problem solving procedures.
- How curriculum content may integrate with problem solving assessment to motivate school teachers.

Each presenter is given about 20 minutes for presentation followed by 10 minutes for Q & A.

For the interactive event, presenters are advised to prepare a demo and a poster, so that participants can learn the general picture from the poster and have practical experience with the demo. The interactive event will last 40 minutes.

For the roundtable discussion, the organizers will host the discussion. Each participant will introduce themselves at first, then explain their own thinking about the design, analysis technique or case study for the topic focusing on “how learning analytics could facilitate the assessment of collaboration and enhance collaboration”. The organizers hope that all the participants can shape the view of how collaboration can be analyzed computationally more clearly by joining the discussion.

The workshop will be advertised through Twitter, Weibo, Wechat, and email list.

Required equipment for the workshop: There is no required equipment but participants are advised to bring their laptops, so that they can access the demo when possible.

3 INTENDED OUTCOMES

The workshop aims for gathering together the researchers who are interested in problem solving assessment from both of the field of learning science and data analytics. By presentation, interactive event and roundtable discussion, we hope researchers with different backgrounds can inspire each other and even form some future collaboration after the workshop. We will also note how researchers from different backgrounds work together in the workshop and disseminate our summary notes via Weibo, Wechat, email list, Twitter, and organizers’ websites.

The outcomes of the workshop ought to be highly relevant to the special theme of LAK 2019, which is “learning analytics can be used to promote inclusion and success”. For “inclusion”, we try to extend the functionality of problem solving assessment, and make the assessment cover more skills in real-life with the help of learning analytics. For “success”, we try to improve problem solving assessment to ensure students success after school education.

REFERENCES

- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36-46
- Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C.,... Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, 13, 74-83
- Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*: Springer.
- Rosen, Y. (2017). Assessing Students in Human - to - Agent Settings to Inform Collaborative Problem - Solving Learning. *Journal of Educational Measurement*, 54(1), 36-53
- Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning & Individual Differences*, 24(2), 42-52
- Wilson, M., Gochyyev, P., & Scalise, K. (2017). Modeling data from collaborative Assessments: Learning in digital interactive social networks. *Journal of Educational Measurement*, 54(1), 85-102
- Zhang, L., VanLehn, K., Girard, S., Bursleson, W., Chavez-Echeagaray, M. E., Gonzalez-Sanchez, J.,... Hidalgo-Pontet, Y. (2014). Evaluation of a meta-tutor for constructing models of dynamic systems. *Computers & Education*, 75, 196-217
- Zhang, L., Yu, S., Li, B., & Wang, J. (2017). Can Students Identify the Relevant Information to Solve a

Problem? *Journal of Educational Technology & Society*, 20(4), 288-299

Mining LMS Data to Make Early Prediction of Learning Failure

Song Lai

Beijing Normal University
laisong@mail.bnu.edu.cn

Fati Wu

Beijing Normal University
wft@bnu.edu.cn

ABSTRACT: With the adoption of learning management systems, plenty of data about students has become available. Numerous researchers have exploited these e-learning data to predict student achievement by applying educational data mining approaches. The prediction outcomes can not only identify at-risk students and then give them learning warning but also make instructors master students' learning status, so that instructors provide timely interventions to help them in the learning process. The study employs educational data mining approaches to build a model for early identification of at-risk students by mining students' online interaction data. The experimental results indicate that deep belief networks algorithm gives a best accuracy of 0.84, which delivers a relatively better prediction effectiveness. The findings support the potential for early prediction of learning failure.

Keywords: e-learning, educational data mining, achievement prediction, early identification

1 INTRODUCTION

The success of BYOD (Bring Your Own Device) aligns with global trends toward mobility as more people, from children to adults, own mobile devices and are accessing the internet increasingly different environments for learning (Freeman, Adams Becker, & Cummins, 2017). With BYOD expanding in schools, students can keep on learning anytime and anywhere through interaction repeatedly with their own devices outside a traditional classroom. Each interaction action of students is supervised and stored in state-of-the-art learning management systems (LMSs), which are able to track and analyze students' online activities without the necessity of time-consuming data-collection (Conijn, Snijders, Kleingeld, & Matzat, 2017). These actions commendably describe students' online learning behaviors contributing to their learning results, whose analysis involves applying the techniques of learning analytics due to the instructors often becoming short of a comprehensive vision of students' learning information. Learning analytics is defined as "the measurement, collection, analysis and reporting of data about learners and their context, for the purpose of understanding and optimizing learning and the environments in which it occurs" (Long, Siemens, Conole, & Gasevic, 2011). It can be extensively applied to predict student achievement about whether students fail to pass a course or not by researchers from the field of computer science and education (Baker & Yacef, 2009; Hu, Lo, & Shih, 2014; Macfadyen & Dawson, 2010; Romero & Ventura, 2010). The final prediction outcome can provide feedback such as giving a warning about risk of learning failure for students in the learning process to promote self-regulated learning. Also, it will allow instructors tutor the corresponding students by providing appropriate

guidance in a relatively easy way. Hence, the more accurate outcome of prediction can facilitate the improvement of learning for students and teaching for instructors, resulting in preventing learning failure by at-risk students (Costa, Fonseca, Santana, de Araújo, & Rego, 2017; Jayaprakash, Moody, Lauría, Regan, & Baron, 2014).

The present study aims to make early prediction of learning failure by presenting a model, which is built by mining the LMS data concerning online activities during the blended learning process. A total of 662 senior high school students participated in experiments wherein they were asked to learn with mobile devices. To accurately predict learning failure, educational data mining (EDM) approaches are exploited to determine the best effective and predictive model. The experimental results show that deep belief networks (DBN) algorithm results in a best accuracy of 0.84 in achievement prediction. The findings would contribute to the possibility of early identification of students who are likely to become at-risk.

The rest of the study is organized as follows. Section 2 briefly introduces background information. Next, section 3 presents data sources and classification approaches employed in this study. Section 4 shows the experimental results in detail. Finally, section 5 provides conclusions and some future works.

2 BACKGROUND

The increasing focus on student-centered learning is driving the development of new technologies. It can be possibly achieved with the fusion of internet and communication technologies (ICTs). ICTs integrated into educational institutions have significantly altered the way in which instructors teach and students obtain knowledge. The change is able to adapt to the new development of education in order to facilitate the enhancement of teaching quality. Compared with face-to-face courses in the traditional education field, blended courses or online courses in educational cloud platforms can offer online learning resources for students and additional functions, such as forums, assignments, presentations and quizzes (Pina, 2012), which make students have more time to structure and organize their thoughts, and communicate simultaneously or even participate in multiple tasks at the same time (Cobo et al., 2011). These components support students' communication and collaboration, and enable them to share ideas, post problems, comment on posts by other students, and obtain feedback in online teaching-learning environments (Raghavan, Catherine, Ikbali, Kambhatla, & Majumdar, 2010). Thus, a large number of learning data can be collected by LMSs which can in turn be analyzed by institutions. The analytical results can significantly contribute to making achievement prediction (Romero & Ventura, 2010), which is one of the oldest and most useful applications of EDM.

Analyzing online interaction data is useful to identify how students perform their quizzes and exams. With these data, many useful EDM approaches have been widely applied in the achievement prediction models for assessing learning failure. An overview on predicting achievement exploiting different EDM techniques (e.g., SVM, NaïveBayes, and DecisionTree) was provided (Shahiri, Husain, & Rashid, 2015). Also, these approaches were employed to improve accuracy of predicting dropouts. More specifically, Lykourantzou, Giannoukos, Mpardis, Nikolopoulos and Loumos (2009) developed predictive models using neural network and multiple linear regression to achieve student performance prediction in e-learning courses. Smith, Lange and Huston (2012) employed NaiveBayes algorithm to predict learning failure. Shelton, Hung and Lowenthal (2017) claimed that the best predictive model could be generated using DecisionTree classification. Hu et al. (2014) exploited

C4.5, LGR and CART methods to identify at-risk students at three stages during a course, whose results demonstrated that the inclusion of EDM techniques contributed to the construction of an early warning system in an e-learning environment. These studies validated the effective predictive ability in accurately predicting learning failure. However, few studies have considered the importance of timing. Obviously, early guidance is a critical element in preventing learning failure (Jayaprakash et al., 2014). Predicting failure early enough is important to allow for appropriate guidance to reduce the risk of learning failure (Costa et al., 2017). Hence, this study focuses on building a model for the early identification of at-risk students.

3 DATA AND CLASSIFICATION TECHNIQUES

3.1 Data source

The data information with reference to students' online interaction was collected and accounted from the LMS "E-school Bags" in the smart educational cloud platform. 662 students in senior high school used the available portable android devices (PADs) to learn courses (e.g., mathematics) at any time and place. They were taught in the fall quarter from September, 2016 to January, 2017 about seventeen weeks of teaching and two weeks of final exams.

Achievement predictors come from the modules: attendance, resource, forum and assignment. In learning process, what students interacted with PADs, such as visited content pages, posted messages for question and answer, and made quizzes, was recorded as structured data (Sun et al., 2017). The raw data generated from the LMS was pre-processing to form predictor variables at each stage. Details are shown in table 1. Students are evaluated in the 5th, 9th (before midterm), 15th, and 19th week. The grade of pass is 60. Students' learning outcomes and grade distributions are presented in table 2.

Table 1: Data source.

Attribute type	Attribute name	Description
Predictors	Attendance	Number of learning online
		Duration of learning online
		Number of learning notes
	Resource	Number of viewing course resources
		Duration of viewing course resources
		Number of questioning in text
	Forum	Number of answering in text
		Number of questioning in hypermedia
		Number of answering in hypermedia
		Number of answering recommended
		Score of quizzes before class in average
		Score of quizzes during class in average
		Duration of quizzes during class

Target	Learning outcome	Score of exams after class in average
		Passed or failed the course

Table 2: Learning outcomes and grade distributions.

Number of students	Learning outcome		Score	
	Number of pass (score > 60)	Number of fail (score < 60)	Mean	Standard deviation
662	406	256	64.20	11.72

3.2 Classification techniques

To discriminate from students such as some students who are at risk of dropping from the blended learning and others perform better adequately, a classification analysis is performed to identify the level of student achievement. EDM approaches, k-nearest neighbor (KNN), NaiveBayes, support vector machine (SVM), and DecisionTree are applied to predict achievement for early identification of at-risk students (Shahiri et al., 2015). They are classification techniques based on supervised machine learning in the field of artificial intelligence. In recent years, deep neural networks such as convolutional neural networks (CNN) and DBN have achieved remarkable success in numerous classification tasks such as text classification and image identification. In the study, a CNN with four convolution layers and a simple DBN are trained. They are expected to present the potential for making early prediction of learning failure.

4 EXPERIMENTAL RESULTS

The accuracy (as Equation 1) and evaluation metrics are useful measures to assess the prediction performance. EDM techniques and several meaningful attributes (as Table 1) are used to assess early-stage predictive ability in weeks 5, 9, 15 and 19 in a 19-week semester. The experiments were made with data randomly divided into training and testing sets at a 500:162 ratio 100 times for the 2-class classification problem. Fig. 1 shows the classification results. The achievement prediction accuracy is promoted gradually along with the increase of week for each of the six chosen algorithms. Table 3 presents the confusion matrix of the DBN at Week 19. More specifically, DBN gives the accuracy of 0.66 at Week 5, improving to 0.67 at Week 9, 0.69 at Week 15 and 0.84 at Week 19. Compared with other approaches, DBN provides the best accuracy. Also, CNN shows a relatively better accuracy of 0.81 at Week 19 than other baseline algorithms including KNN, NaiveBayes, SVM and DecisionTree. Obviously, all baseline algorithms perform worse than deep learning algorithms. Their results, below 0.75, are poor in all four stages. This means the DBN model is pretty effective in predicting learning failure. By training the model with data that is randomly divided 100 times, the prediction performance of DBN is much more reliable.

$$\text{Accuracy} = \frac{\text{the number of students who are correctly identified}}{\text{the total number of students}} \quad (1)$$

Table 3: Confusion matrix of DBN result at week 19.

Prediction

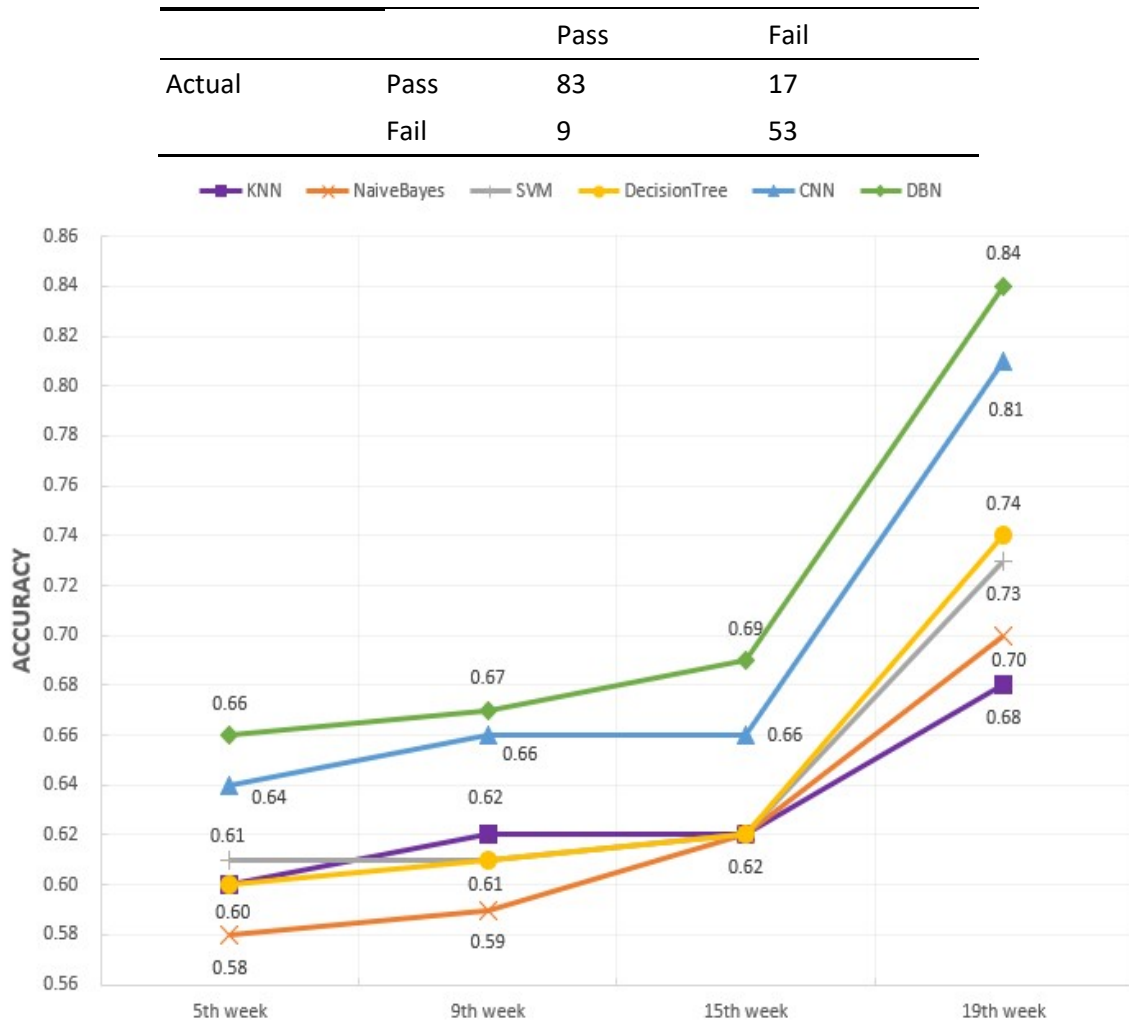


Figure 1: Achievement prediction.

5 CONCLUSIONS AND FUTURE WORKS

In this context, a predictive model capable of identifying students at risk of failure is built. The experimental results present a relatively high accuracy of early-stage prediction, but also indicate the limited early-stage predictive power due to data sparseness at the beginning of the course. Achievement prediction can give an early warning of learning failure risk, which can allow instructors and students to address the issue in time to rescue the students' final grade. A relatively high prediction accuracy contributes to the possibility of providing more accurate early-stage learning warning with real-time feedback. The findings for improvement in teaching and learning initiatives are important to maintain students' achievement and the effectiveness of learning process.

The study has limitations, for example, it is insufficient about the LMS data predictor variables. In a future study, more indicators, including students' emotion and learning behavior sequence, will be mined. More importantly, larger data sets will be collected for the application of other deep learning algorithms expected to significantly improve the accuracy of achievement prediction.

REFERENCES

- Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *JEDM | Journal of Educational Data Mining*, 1(1), 3–17. Retrieved from <http://jedm.educationaldatamining.org/index.php/JEDM/article/view/8>
- Cobo, G., Garcia, D., Santamaría, E., Moran, J. a, Melenchón, J., & Monzo, C. (2011). Modeling students' activity in online discussion forums: A strategy based on time series and agglomerative hierarchical clustering. *Proceedings of the 4th International Conference on Educational Data Mining*.
- Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17–29. <https://doi.org/10.1109/TLT.2016.2616312>
- Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247–256. <https://doi.org/10.1016/j.chb.2017.01.047>
- Freeman, A., Adams Becker, S., & Cummins, M. (2017). *NMC/CoSN Horizon Report: 2017 K-12 Edition*. Austin, Texas: The New Media Consortium. Retrieved from <https://www.learntechlib.org/p/182003/>
- Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2014.04.002>
- Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, 1(1), 6–47. <https://doi.org/10.18608/jla.2014.11.3>
- Long, P., Siemens, G., Conole, G., & Gasevic, D. (2011). Announcing open course: Learning and knowledge analytics. In *1st International Conference on Learning Analytics & Knowledge*. Banff, AB, Canada.
- Lykourantzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in E-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*, 60(2), 372–380. <https://doi.org/10.1002/asi.20970>
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers and Education*, 54(2), 588–599. <https://doi.org/10.1016/j.compedu.2009.09.008>
- Pina, A. A. (2012). *An overview of learning management systems*. Louisville, KY, USA: in Virtual Learning Environments: Concepts, Methodologies, Tools and Applications.
- Raghavan, P., Catherine, R., Ikbali, S., Kambhatla, N., & Majumdar, D. (2010). Extracting Problem and Resolution Information from Online Discussion Forums. In *Proceedings of the 16th International Conference on Management of Data, December 9–12, 2010*.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 40(6), 601–618. <https://doi.org/10.1109/TSMCC.2010.2053532>

- Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, 72(3), 414–422. <https://doi.org/10.1016/j.procs.2015.12.157>
- Shelton, B. E., Hung, J. L., & Lowenthal, P. R. (2017). Predicting student success by modeling student interaction in asynchronous online courses. *Distance Education*, 38(1), 59–69. <https://doi.org/10.1080/01587919.2017.1299562>
- Smith, V. C., Lange, A., & Huston, D. R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Network*. <https://doi.org/10.24059/olj.v16i3.275>
- Sun, B., Lai, S., Xu, C., Xiao, R., Wei, Y., & Xiao, Y. (2017). Differences of online learning behaviors and eye-movement between students having different personality traits. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education - MIE 2017* (Vol. 2017–Novem, pp. 71–75). <https://doi.org/10.1145/3139513.3139527>

Use of Feedback According to Students' Affective State during Problem Solving

Jingjing Zhang

Big Data Centre for Technology mediated Education, Beijing Normal University
Jingjing.zhang@bnu.edu.cn

Ming Gao

Research Centre of Distance Education, Beijing Normal University
mgao519@mail.bnu.edu.cn

Manolis Mavrikis

University College of London
m.mavrikis@ioe.ac.uk

Wayne Holmes

The Open University
wayne.holmes@open.ac.uk

Ning Ma

Advanced Innovation Center for Future Education, Beijing Normal University
horsening@163.com

ABSTRACT: Pattern-finding analytical techniques to improve our understanding of the use of feedback and scaffolding during problem-solving processes have attracted much attention. This study used a lag sequential analysis to unfold the learning patterns according to affective states during student problem-solving processes. The results have shown that the significant transitional sequences of learning activities before and after requesting feedback and seeking scaffolding differ between students associated with positive and negative affective states. This study highlights the importance of providing tailored support based on students' affective states, to further enhance their technology-mediated learning experience.

Keywords: Problem solving, learner support, affective states, learning analytics

1 INTRODUCTION

The importance of problem solving has been highlighted frequently in contemporary education (Greiff, et al., 2014). One of the most common ways to foster student problem-solving skills is to assign problem-based tasks to be completed in intelligent tutoring systems. In such learning environments, some form of support, such as feedback, scaffolding or elicited explanations, is usually provided to help students explore effectively (Liu, et al., 2004). In such explorations, students' affective states are, to a large extent, associated with their learning experiences and outcomes (Scotty et al., 2004). Positive affective states may contribute to learning (Csikszentmihalyi, 1990), while negative ones may undermine learning (although we acknowledge that some negative affective states can be necessary in learning, as students' progress towards understanding) (Woolf et al., 2009; Baker et al. 2010;

Grawemeyer et al., 2017). This is particularly the case in student problem-solving processes. Previous research has focused on how to detect students' affective states using varied methods, and has attempted to identify correlational or causal relationships between affective states and learning outcomes quantitatively (i.e. Calvo & D'Mello, 2010; VanLehn et al., 2017). Nevertheless, how student affective states affect their use of feedback and scaffolding during their problem-solving processes is still not fully understood (e.g. Grawemeyer et al., 2016). Thus, this study attempted to identify patterns of feedback requesting and scaffolding usage during problem-solving processes, and to explore whether there are any differences between students with different affective states.

2 METHODS

2.1 Participants and Setting

Our project involved 189 students (aged between 9 and 10 years) from six classes of three primary schools, which were all in or around Beijing, China. For approximately 45 minutes, the students were asked to complete 18 fractions-related tasks in a computer-based exploratory learning environment called Fractions Lab. Fractions Lab is designed to help students learn by interacting with different representations of fractions, while being aided by learner support such as different types of feedback and scaffolding, to solve given tasks. Built upon our previous work (e.g., Mavrikis, M., Holmes, W., Zhang, J., & Ma, N., 2018), Task 10 was selected for this study as the case to explore further how the use of feedback and scaffolding was associated with affective states. Task 10 was "Using two fractions with the same denominator, create an addition that equals $9/12$ ". This task was selected as students tended to present different affective states while working on it. In this study, 57 students were found to have positive affective states (i.e., they identified the task as "enjoyable" or "interesting") while 15 students were found to have negative affective states (i.e., they identified the task as "frustrating", "confusing" or "boring").

2.2 Data Analysis

While the students were interacting with Fractions Lab, their interactive logs (e.g. id, events) were collected and saved automatically. Events were categorized into nine different activities: *GenF* (generating fraction), *ChaF* (changing the denominator or numerator of a fraction), *TraF* (deleting fraction), *LabC* (dragging fractions to the balance to compare, add or subtract), *TasO* (opening the description of the present task), *TasR* (resetting the present task), *SeeS* (seeking scaffolding to solve the problem, such as, creating the equivalent fraction, changing the color of the numerator, etc.), *StaR* (showing the current states, such as, true, false or unfinished), and *FeeB* (requesting feedback or hints to resolve the task). The Mann-Whitney U test was used to test for the different uses of learner support (e.g. requesting feedback and seeking scaffolding), according to the two groups of students with different affective states. The lag sequential analysis method (Sackett, 1978) was adopted to compare patterns of feedback requesting (*FeeB*) and scaffolding seeking (*SeeS*), to identify the significant transitions with regard to these two types of learner support, by using the version 5.1 of Generalized Sequential Querier (GSEQ 5.1).

3 RESULTS AND DISCUSSION

3.1 Significant Transitional Sequences during Problem-Solving Process

Although there was no significant difference between the two groups of students associated with positive and negative affective states, in terms of how many times they requested feedback ($z=-.238$, $p=0.812 > 0.05$) and scaffolding ($z=-.151$, $p=0.880 > 0.05$), the transitional sequences of learning activities were significant. The lag sequential analysis was used to probe the significant transitional sequences during the students' problem-solving processes. To reach a statistically significant result of the sequence, a z-score higher than 1.96 (Bakeman & Gottman, 1997) was used to evaluate the significance of transition. As shown in Figure 1, the significant learning activities before and after feedback requesting (FeeB) and scaffolding seeking (SeeS) differed between the students associated with the different affective states.

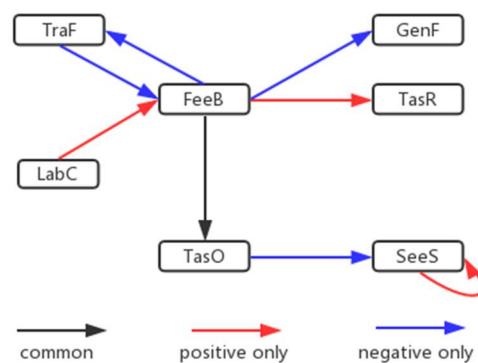


Figure 1: Significant sequences of feedback requesting and scaffolding seeking

The students who requested feedback after deleting a fraction were likely to indicate that they were frustrated, confused or bored. This perhaps implies that they failed to complete the task and had to request feedback in order to achieve the learning goal. In this matter, feedback was requested for the purpose of “telling me the answer”. After receiving feedback, they tended either to delete the fraction (TraF) or generate a new one (GenF) to make another attempt. Scaffolding (e.g. creating the equivalent fraction, changing the color of the numerator) was sought after they had opened the task, before they first explored how to solve the problem. This further confirms that the learner support embedded in Fractions Lab was used immediately to complete the task without any prior exploration. As Fractions Lab was defined as an exploratory learning environment, where feedback was designed to ask students to “think and explore”, or “try, fail, and learn” (Holmes et al., 2015), such open and reflective feedback may not provide “the support as the answer” that these students expected to receive. Thus, the students who expected learner support to provide them with answers were likely to feel frustrated, confused or bored.

Students who requested feedback after using the balance tool to confirm their results tended to enjoy or be interested in the learning task. These students with positive affective states seemed to be more proactive in terms of using scaffolding embedded in Fractions Lab, such as the balance tool, before requesting feedback. After receiving feedback, they tended to reset the task (TasR), which implies that they were not afraid to start over again. During such a learning process, they requested feedback to help them explore the learning task further. This was further confirmed by the significant transitional

sequences of scaffolding seeking with itself (as indicated in the self-loop from SeeS to SeeS in Figure 1). In this way, the students attempted to try out all possible scaffolding (e.g. creating the equivalent fraction, changing the color of the numerator) designed in Fractions Lab, which illustrates that these students were in an exploratory mode of learning.

4 CONCLUSION

In this study, an attempt was made to use pattern-finding analytical techniques (e.g. Mavrikis, 2010) to improve our understanding of the use of feedback and scaffolding during the problem-solving process. These important correlates of learning have been researched extensively (e.g. Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; Porayska-Pomsta, Mavrikis, & Pain, 2008) but outcomes remain conjectural. Learning analytics will provide a new perspective for examining transitional sequences of learning activities according to different affective states during the problem-solving process, and will thus inform future intervention in exploratory learning environments. This work is important in that it uses interaction patterns of requesting feedback and scaffolding to gain insights into student's reasoning processes. It further highlights the importance of students' affective states which significantly alter their behaviors, and in turn can be influenced by how learner support is provided in exploratory learning environments. Careful investigation of how students behave before and after the use of feedback and scaffolding in problem-solving intelligent environments will enable the provision of increasingly tailored support based on students' affective states, and further enhance their technology-mediated learning experiences.

REFERENCES

- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73 (3), 277-320.
- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*: Cambridge university press.
- Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241.
<https://doi.org/10.1016/j.ijhcs.2009.12.003>
- Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1), 18-37.
- Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media*, 29(3), 241-250.
- Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*: Harper and Row, New York.
- Grawemeyer, B., Mavrikis, M., Holmes, W., Gutierrez-Santos, S., Wiedmann, M., & Rummel, N. (2016). Affecting off-task behaviour: how affect-aware feedback can improve student learning (pp. 104-113). Presented at *the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, ACM. <https://doi.org/10.1145/2883851.2883936>

- Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction*, 1–40.
<https://doi.org/10.1007/s11257-017-9188-z>
- Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem-solving skills and education in the 21st century. *Educational Research Review*, (13), 74-83.
- Holmes, W., Mavrikis, M., Hansen, A., & Grawemeyer, B. (2015). Purpose and Level of Feedback in an Exploratory Learning Environment for Fractions. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (Vol. 9112, pp. 620–623). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-19773-9_76
- Liu, M., Bera, S., Corliss, S. B., Svinicki, M. D., & Beth, A. D. (2004). Understanding the connection between cognitive tool use and cognitive processes as used by sixth graders in a problem-based hypermedia learning environment. *Journal of Educational Computing Research*, 31(3), 309-334.
- Mavrikis, M. (2010). Modelling student interactions in intelligent learning environments: constructing bayesian networks from data. *International Journal on Artificial Intelligence Tools*, 19(6), 733–753.
- Mavrikis, M., Holmes, W., Zhang, J., & Ma, N. (2018). Fractions Lab Goes East: Learning and Interaction with an Exploratory Learning Environment in China. In: Penstein Rosé C. et al. (Eds.), *Artificial Intelligence in Education* (pp. 209-214). AIED 2018. Lecture Notes in Computer Science, vol 10948. Springer, Cham.
- Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction*, 18 (1), 125-173.
- Sackett, G. P. (1978). *Observing behavior: Theory and applications in mental retardation* (Vol. 1): University Park Press.
- Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2013, July). Affective state detection in educational systems through mining multimodal data sources. In *Educational Data Mining 2013*.
- VanLehn, K., Zhang, L., Burleson, W., Girard, S., & Hidago-Pontet, Y. (2017). Can a non-cognitive learning companion increase the effectiveness of a meta-cognitive learning strategy?. *IEEE Transactions on Learning Technologies*, 10(3), 277-289.
- Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4), 129-164.

How to Generate Actionable Predictions on Student Engagement: Hands-on Tutorial with Python Scikit-Learn

Erkan Er

Universidad de Valladolid

erkan@gsic.uva.es

ABSTRACT: The existing predictive research has been mostly based on post-hoc techniques that cannot be used in real-world practice as they are built only after the target action occurs in the context (e.g., dropouts). As a result, the impact on supporting pedagogy in real time has been limited. Building on past machine learning workshops and tutorials in LAK conferences, this tutorial session will introduce the machine learning approaches for creating actionable predictions (i.e., in-situ learning and transferring across courses) that can offer many utilities for designing real-world interventions. The participants will be guided through several hands-on examples to practice the use of these approaches in solving several real-world cases. Python Scikit-Learn will be used to implement the practice examples. At the end of the activity, the participants will reflect on their experience and share their opinions on the use of in-situ learning and transfer across courses techniques in their own research. This session will increase the awareness of LA researchers and practitioners about building actionable predictive models and will inspire future use of these approaches in real-world contexts.

Keywords: actionable predictive models, transferring across courses, in-situ learning, python, scikit-learn

1 TUTORIAL BACKGROUND

The area of predictive analytics has gained an increasing attention from the research community after the emergence of massive open online courses (MOOCs). Thus far, the prediction research has been based on the data from a single past course to build and test predictive models with post-hoc approaches (e.g., cross validation) (Gardner & Brooks, 2018). However, these approaches are not valid for real-world use since they require the true training labels which cannot be known until the target event takes place (e.g., dropouts) (Boyer & Veeramachaneni, 2015; Er, Bote-Lorenzo, Gómez-Sánchez, Dimitriadis, & Asensio-Pérez, 2017; Gardner & Brooks, 2018; Whitehill, Mohan, Seaton, Rosen, & Tingley, 2017). For example, Jiang, Williams, Schenke, Warschauer, & Dowd (2014) and Xing, Chen, Stein, & Marcinkowski (2016) attempted to predict if students would drop out using models that were trained with labels that can only be obtained once the course is over.

To overcome the limitations of post-hoc prediction models, several works explored the use of the *transferring across courses* approach, in which a prediction model is built using a completed MOOC and then used for designing interventions in a follow-up MOOC (Boyer & Veeramachaneni, 2015, 2016). MOOCs themselves indeed offer distinct opportunities that make transfer learning an advantageous approach (e.g., transferring across re-runs of the same course, or across courses from the same domain or with similar instructional design) (Boyer & Veeramachaneni, 2015). Nevertheless, there are not many studies that have investigated the potentials of transferring across

MOOCs in comparison to post-hoc prediction approaches. Authors in (Boyer & Veeramachaneni, 2016) and (Boyer, Gelman, Schreck, & Veeramachaneni, 2015) have tested the transferability of a dropout prediction model across different MOOCs. The results were quite promising, showing that different courses could be used to train a model to make predictions in another course. An increase in the accuracy of the predictions was noted when multiple courses were used to train the models, or when the training set was calibrated (e.g., maintain the learners in the training data that are more similar to the learners in the target course). Complementary to these findings, a recent study (Whitehill et al., 2017) has indicated that training a model on many other courses might lead to more accurate models compared to training on a course from the same discipline.

Different from transferring models across different MOOCs, Boyer & Veeramachaneni (2015) have proposed the in-situ learning approach that allows training a model based on proxy labels (e.g., students are considered dropout if they have no interactions for a specific week (Kurka, Godoy, & Von Zuben, 2016)). A few studies have investigated the use of in-situ learning in MOOCs. For example, Bote-Lorenzo & Gómez-Sánchez (2017, 2018) used in-situ learning to predict if there will be a decrease in student engagement at the end of a particular chapter (e.g., chapter 4) using the model trained on the preceding chapter data (e.g. chapter 3). Some other researchers (Boyer & Veeramachaneni, 2015; Kurka et al., 2016; Whitehill et al., 2017) have tested in-situ learning for building dropout prediction models that are transferable across different weeks within the same course and compared its performance with conventional transfer learning (using past courses).

Although transfer across courses and in-situ learning can provide actionable information for creating real-world interventions, their use is very limited in MOOC prediction research (Gardner & Brooks, 2018). Actionable information regarding students' future learning behavior can offer a wide range of pedagogical utilities. Such machine learning techniques to create timely actionable information, if widely adopted and practiced by researchers and practitioners, can promote the LA-empowered educational interventions in real-world practice. In an attempt to address this crucial gap, this tutorial session will introduce transfer across courses and in-situ learning techniques and demonstrate the participants real-world use of these techniques through several hands-on examples. Python Scikit-Learn (Pedregosa et al., 2012), one of the most widely used machine learning library in the field, will be used in the tutorial.

This tutorial is highly related with several past LAK workshops and tutorials, including, but are not limited to “Building predictive models of student success with the Weka toolkit” and “Python Bootcamp for Learning Analytics Practitioners”. These previous sessions have mainly focused on fundamental machine learning topics (e.g., unsupervised learning, text mining). Building on this evolving knowledge basis in the learning analytics community, the proposed session will motivate and inspire the LA researchers and practitioners to create actionable machine learning models that can be used for offering real-world interventions.

2 ORGANIZATIONAL DETAILS

The session will be organized into three parts. In the first part, the in-situ learning and transferring across courses will be introduced. In the second part, I will facilitate a hands-on activity about building machine learning models in Python. In the third part, building on the first two parts, the participants will use transfer across courses and in-situ learning approaches in practice for

generating actionable predictive models. At the end, I will facilitate a discussion among the participants regarding the potential uses of such approaches in their own research. For dissemination, I will summarize the participants' inputs and share the highlights through social media with relevant hashtags.

The proposed event is planned to be a half-day tutorial. It will be an open session where any interested delegate may register to attend. Although experience with Python is preferred, however, it is not mandatory.

The dissemination of the activity will be performed through several professional learning communities (e.g., SoLAR) and social media sites (e.g., twitter, LinkedIn, ResearchGate). A WordPress website will be used to share the learning materials in a tutorial format with participants registered for the tutorial. Additionally, a Python GitHub repository will be created to store all the coding and data to be used during the tutorial. This repository will be disseminated before the tutorial session. For recruitment, the participants will be asked to need to complete an online form¹, which will be released 2 weeks before the activity. The expected number of participants is 20. The participants will need to install Anaconda² on their laptops. Supplementary learning materials will be provided printed during the tutorial session.

3 TUTORIAL OBJECTIVES OR INTENDED OUTCOMES

The main objective of the proposed tutorial session is to teach the participants the concepts of transfer across courses and in-situ learning to enable the participants to put into practice their knowledge through several hands-on activities. The participants will reflect on their experience and share their ideas on the ways that transfer across courses and in-situ learning can relate with their own research (if possible) as well as the ways that they can be used for creating educational interventions. To disseminate these ideas about the pedagogical utilities of actionable predictions, social media will be used (e.g., hashtags in Twitter).

4 ACKNOWLEDGEMENTS

This tutorial session is organized as part of the project WeLearnAtScale³, funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 793317.

REFERENCES

Bote-Lorenzo, M. L., & Gómez-Sánchez, E. (2017). Predicting the decrease of engagement indicators in a MOOC. In *Proceedings of Seventh International Conference on Learning Analytics and Knowledge* (pp. 143–147). Vancouver, Canada. <https://doi.org/10.1145/3027385.3027387>

¹ <https://feelthedata.wordpress.com/2018/10/24/tutorial-lak19-scikit-learn-actionable-models/>

² <https://www.anaconda.com/download/>

³ <https://welearnatscale.gsic.uva.es/>

- Bote-Lorenzo, M. L., & Gómez-Sánchez, E. (2018). An approach to build in situ models for the prediction of the decrease of academic engagement indicators in Massive Open Online Courses. *Journal of Universal Computer Science*, 1. Accepted.
- Boyer, S., Gelman, B. U., Schreck, B., & Veeramachaneni, K. (2015). Data science foundry for MOOCs. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics* (pp. 1–10). Paris, France. <https://doi.org/10.1109/DSAA.2015.7344825>
- Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in Massive Open Online Courses. In *Proceedings of the 17th Conference on Artificial Intelligence in Education* (pp. 54–63). Madrid, Spain.
- Boyer, S., & Veeramachaneni, K. (2016). Robust predictive models on MOOCs: Transferring knowledge across courses. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 298–305). Raleigh, NC, USA.
- Er, E., Bote-Lorenzo, M. L., Gómez-Sánchez, E., Dimitriadis, Y., & Asensio-Pérez, J. I. (2017). Predicting student participation in peer reviews in MOOCs. In *Proceedings of the Second European MOOCs Stakeholder Summit 2017*. Madrid, Spain.
- Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 28(2), 127–203. <https://doi.org/10.1007/s11257-018-9203-z>
- Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & Dowd, D. O. (2014). Predicting MOOC performance with week 1 behavior. In *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 273–275). London, UK.
- Kurka, D. B., Godoy, A., & Von Zuben, F. J. (2016). Delving deeper into MOOC student dropout prediction. *CEUR Workshop Proceedings*, 1691, 21–27. <https://doi.org/10.1145/1235>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine learning in Python, 12, 2825–2830. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). MOOC dropout prediction: How to measure accuracy? In *Proceedings of the Fourth ACM Conference on Learning@Scale* (pp. 161–164). <https://doi.org/10.1145/3051457.3053974>
- Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, 58, 119–129. <https://doi.org/10.1016/j.chb.2015.12.007>

International Workshop on Technology-Enhanced and Evidence-Based Education and Learning

Rwitajit Majumdar, Ivica Botički and Hiroaki Ogata

Kyoto University

{majumdar.rwitajit.4a , boticki.ivica.5c , ogata.hiroaki.3e} @kyoto-u.ac.jp

ABSTRACT: The multi-disciplinary research approach of Learning Analytics (LA) has provided methods to understand learning logs collected during various teaching-learning activities and potentially enrich such experiences. This workshop aims to explore the frontiers of how technology can help to extract evidence of effective teaching-learning practices by applying the knowledge base of LA and developing novel techniques. It focuses on discussions on realizing a technology-enhanced evidence-based education and learning (TEEL) systems. We invite research articles conceptualizing foundations, methodologies and utility of TEEL systems. Further, we plan to have a focus group activity to validate an initial technical proposal of Learning Evidence Analytics Framework (LEAF) and draw a research road-map of log data-driven evidence-based education system.

Keywords: Technology-enhanced Evidence-based Education and Learning, TEEL, Learning Evidence Analytics Framework, LEAF

1 BACKGROUND

1.1 Motivation for the workshop

The purpose of Learning Analytics (LA) is “understanding and optimizing learning and the environments in which it occurs.” (Siemens, G., & Long, P. 2011). There has been workshops and tutorials in previous LAK conferences discussing about various methods, policies and data-crunching techniques to achieve the purpose. The concept of Evidence-Based Practices (EBP) has its root in medicine and coined by doctors at McMaster University in Hamilton, Ontario in early 1990s (Kvernbekk T., 2017). According to Kvernbekk (2017), EBP involves the use of the best available evidence to bring about desirable outcomes, or conversely, to prevent undesirable outcomes. Davies, P. (1999) reviews the concept of EBP in education. Literature takes various theoretical perspective such as Research-based education (Hargreaves, 1996), Literature-based education (Hammersley, 1997), Context-sensitive practice (Greenhalgh and Worrall, 1997). What is missing is any research agenda of how technology can support the process involving educational big data and the relevant discussions regarding issues in the current age of data-driven education.

In the Learning Analytics community, SOLAR, the term evidence has recently come up in the context of a workshop in 2018’s Learning Analytics Conference (LAK 18) regarding evidence-based institutional LA policy (Tsai Y.S., Gašević D., Scheffel, M., 2018; sheilaproject.eu) and in LAK 17 in work presented by Ferguson & Clow (2017) where they introduce Learning Analytics Community Exchange (LACE) project’s Evidence Hub. The Evidence Hub (<http://evidence.laceproject.eu/>) followed the evidence-based medicine paradigm to synthesize published LA literature and meta-

analyze four propositions about learning analytics: whether they support learning, support teaching, are deployed widely, and are used ethically. But neither of the works look at technological affordances required to extract evidence of learning from logged data and make it available for the practitioners to adopt in their own context. This workshop is on technology-enhanced evidence-based education and learning (TEEL) system. It aims to bring together researchers, practitioners and policy makers to discuss ways of conceptualizing evidence of learning success in different technology-enhanced learning contexts.

1.2 Contribution to research and alignment to LAK 2019

This workshop initiates a discussion on foundations, methodologies and utility of TEEL systems to extend the existing knowledge of learning analytics. It explores how to utilize existing LA infrastructures to capture teaching-learning practices, evaluate its effectiveness and recommend good practices back to the community of teachers. We ideate to refine our initial proposal on Learning Evidence Analytics Framework (LEAF), a framework that can be used for integrating an evidence-based education system in practice (see Fig1. below).

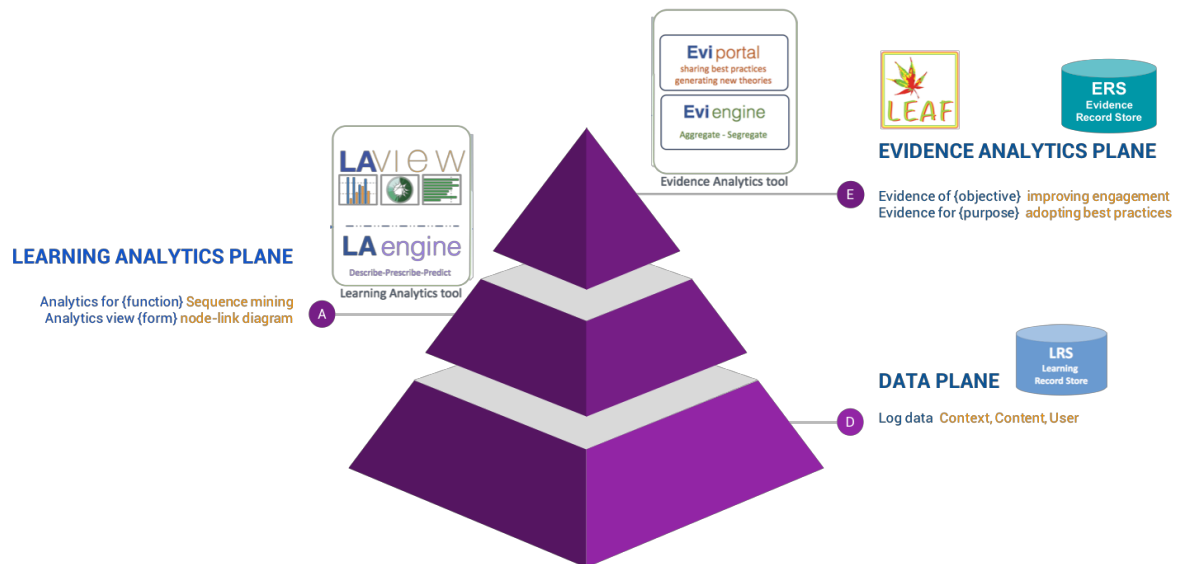


Figure 1: Planes of Analytics

The special theme of LAK2019 is on **Ways in which learning analytics can be used to promote inclusion and success**. Our workshop attempts to provide an operational perspective by discussing the technological infrastructure and methodologies to support extraction of evidence of learner success and developing a framework which connects researchers, educators, and policy-makers by sharing evidences amongst them.

2 WORKSHOP OBJECTIVES

2.1 Research Agenda and Outreach

With the open call for paper we aim to gather researchers and practitioners to present research findings in the workshop. The indicative research topics (not limited to) are listed below:

- Foundations for Technology-Enhanced & Evidence-Based Education & Learning
 - Technology Design Framework, Architecture and Platform
 - Evidence Data format in evidence record store
- Methodologies for Technology-Enhanced & Evidence-Based Education & Learning
 - Extraction of evidence from educational and learning log data
 - Meta-analysis of log data-driven evidences
 - Similarity measures of log data-driven evidence
 - Evaluation of evidences
- Utilizing Technology-Enhanced & Evidence-Based Education & Learning
 - Technological support for adoption of evidences in practice
 - Context-aware recommendation of evidence
 - Case studies of current instantiations
 - Privacy and Ethical issues

We created a website (<https://sites.google.com/view/teel-workshop>) where we shall list the activity outcomes from the workshop and maintain a hashtag #TEEL19 for outreach on the social media.

3 ACCEPTED PAPERS FOR DISCUSSION IN WORKSHOP

There were 9 accepted articles for discussion in this workshop. Authors of these articles were from 6 different countries.

Two of the articles present analysis of technology implementation focusing on teachers. In *Learning Analytics Dashboard Widgets to Author Teaching-Learning Cases for Evidence-based Education*, Majumdar et.al. (2019) components of LAVIEW, a learning dashboard to assist authoring of teaching-learning cases (TLC) by practitioners is described. The TLCs would enable to capture problems identified in a specific context, its indicators in terms of dashboard visualizations, solution and results. Authors propose it as the unit of analysis for evidence-based teaching and learning. In *Behind the Scenes: Designing a Learning Analytics Platform for Higher Education*, Chounta et.al. (2019) reports findings from stakeholder studies during development phase of a LA platform. Their LA platform is targeted for higher education academic institutions in Estonia and this study focus on the teachers' perspective.

Three articles propose models related to learner's artifact evaluation or log analysis to extract evidence. In *Quantitative Evaluation of Concept Maps: An Evidence-Based Approach*, Kadam et.al. (2019) propose automated evaluation algorithm of student submitted concept map

assignment. In *Modelling students' effort using behavioral data*, Moissa et.al. (2019) use interaction and eye gaze data to model student's effort. In *LASAT: Learning Activity Sequence Analysis Tool*, Mishra et.al. (2019) present a case-study of utility of various sequence analysis algorithms which assist in extracting and interpreting students' learning behaviors extracted as frequent patterns (sequence of activities) from their activity traces logged in computer-based learning environments. These algorithms, developed in Institute for Software Integrated Systems, Vanderbilt University, are packaged in a toolkit with the aim to make it accessible to wider community of researchers and practitioners.

Three articles focus on the context of MOOCs. In *Automated MOOC/SPOC Learning Design Verification based on Instructional Design Theories*, Lei et.al. (2019) propose a mechanism that can quickly visualize the courseware with faulty or at-risk designs that may cause obstacles for learners, which allows just-in-time revisions. In *Using Log Data to Evaluate MOOC Engagement and Inform Instructional Design*, Chai et.al. (2019) discusses a framework of MOOC engagement composed of learning-interface, learner-content and learner-community interactions. They illustrate how to utilize the framework with log data from 10 MOOC courses offered by Hong Kong University. In *CLEAR: Cohort-Level Evidence Analysis and Reflection Process as a methodology to assist MOOC Providers and Adopters for effective teaching-learning using MOOCs*, Warriem and Balaji (2019) discuss a case study of National Programme on Technology Enhanced Learning (NPTEL), a national MOOC initiative from India. They focus on the issue of persistent engagement of learners in MOOCs and propose a process flow that will assist the MOOC providers as well as institutions signed up with NPTEL to utilize the evidences available from previous offerings of courses and take meaningful actions on it.

Finally, in *Extracting Self-Direction Strategies and Representing Practices in GOAL System*, Li et.al (2019) provides an instance of building a framework for tracking self-directed actions of learners and illustrates how to utilize it for extracting evidence of best practices and self-reflection. The work is in the context of the GOAL system, where learners use their automatically collected self-data regarding learning and physical activities, to foster various self-direction skills.

ACKNOWLEDGEMENTS

The organizers would like to thank the reviewers for their comments on the submitted articles. This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

REFERENCES

- Chai Y., Lei C., Kwok Y. (2019) Using Log Data to Evaluate MOOC Engagement and Inform Instructional Design. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Chounta I., Pedaste M., Saks K. (2019) Behind the Scenes: Designing a Learning Analytics Platform for Higher Education. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Davies, P. (1999). What is evidence-based education? British journal of educational studies, 47(2), 108-121. DOI: 10.1093/acrefore/9780190264093.013.187

- Ferguson, R., & Clow, D. (2017). Where is the evidence? : A call to action for learning analytics. In Proc. of the 7th International Learning Analytics & Knowledge Conference (pp. 56-65). ACM.
- Greenhalgh, T. and Worrall, J.G. (1997) From EBM to CSM: The evolution of context-sensitive medicine, *Journal of Evaluation in Clinical Practice*, 3, (2), 105–8.
- Hammersley, M. (1997). Educational research and teaching: a response to David Hargreaves' TTA lecture. *British Educational Research Journal*, 23(2), 141-161.
- Hargreaves, D.H. (1996) Teaching as a Research-Based Profession: Possibilities and Prospects. Cambridge Teacher Training Agency Annual Lecture.
- Kadam K., Deep A., Prasad P., Mishra S. (2019) Quantitative Evaluation of Concept Maps An Evidence-Based Approach. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Kvernbekk, T. (2017) Evidence-Based Educational Practice, Oxford research encyclopedias,
- Lei C., Hou X., Wang J., Guo Y. (2019) Automated MOOC/SPOC Learning Design Verification based on Instructional Design Theories. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Li H., Majumdar R., Yang Y.Y., Akçapınar G., Flanagan B., Ogata H. (2019) Tracking Self-Direction Strategies and Representing Practices. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Majumdar R. (2018), Supporting Data-Driven Decision Making by Learners and Teachers, In Proc. of 26th ICCE, Manila, Philippines, Nov 2018.
- Majumdar R., Akçapınar A., Akçapınar G., Flanagan B. and Ogata H. (2019) Learning Analytics Dashboard Widgets to Author Teaching-Learning Cases for Evidence-based Education. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Mishra S., Munshi A., Rushdy M., Biswas G. (2019) LASAT: Learning Activity Sequence Analysis Tool. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Moissa B., Bonnin G., Castagnos S. and Boyer A. (2019) Modelling students' effort using behavioral data. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.
- Ogata H., Majumdar R., Akçapınar G., Hasnine N.M. and Flanagan B. (2018) Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning, In Proc. of 26th ICCE, Manila, Philippines, Nov 2018.
- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5), 30.
- Tsai Y.S., Gašević D., Scheffel, M. (2018). Workshop on Developing an evidence-based institutional learning analytics policy. In the 8th International Learning Analytics & Knowledge Conference. ACM.
- Warriem J.M., Balaji B. (2019) CLEAR: Cohort-Level Evidence Analysis and Reflection Process as a methodology to assist MOOC Providers and Adopters for effective teaching-learning using MOOCs. Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA, 2019.

Learning Analytics Dashboard Widgets to Author Teaching-Learning Cases for Evidence-based Education

Rwitajit Majumdar¹, Arzu Akçapınar¹, Gökhan Akçapınar^{1,2}, Brendan Flanagan¹, Hiroaki Ogata¹

Kyoto University¹, Hacettepe University²
{majumdar.rwitajit.4a, ogata.hiroaki.3e} @kyoto-u.ac.jp

ABSTRACT: In this paper, we introduce the components of LAView, a learning dashboard that assists teachers to author criteria for different teaching-learning cases. We define indicators as the basic unit to define the status of a situation and visualise that on the dashboard. This paper describes the technology design and workflow of the teacher as the user of the dashboard from setting indicator criteria to recording reflection of their results. We conclude with the utility of such a technology support in the context of evidence-based education.

Keywords: BookRoll, LAVIEW, Visual Analytics, Criteria setting, Evidence-based Education

1 LEARNING EVIDENCE ANALYTICS FRAMEWORK (LEAF)

Evidence-based education seeks to establish evidence in the context of teaching-learning practices (Davies, P., 1999; Ferguson, R., & Clow, D., 2017). While it is primarily done as a meta-analysis of the published literature, we attempt to extract evidence from practice. Our novelty lies in the approach to conceptualize evidence in practice by utilizing educational big data. We base our work on the Learning Evidence Analytics Framework (LEAF) (Ogata H., et.al. 2018). The components of LEAF are based on the LA platform proposed by Flanagan and Ogata (2017). It extends the infrastructure to include specific functionalities in the LA Dashboard and an Evidence Portal (see Figure 1).

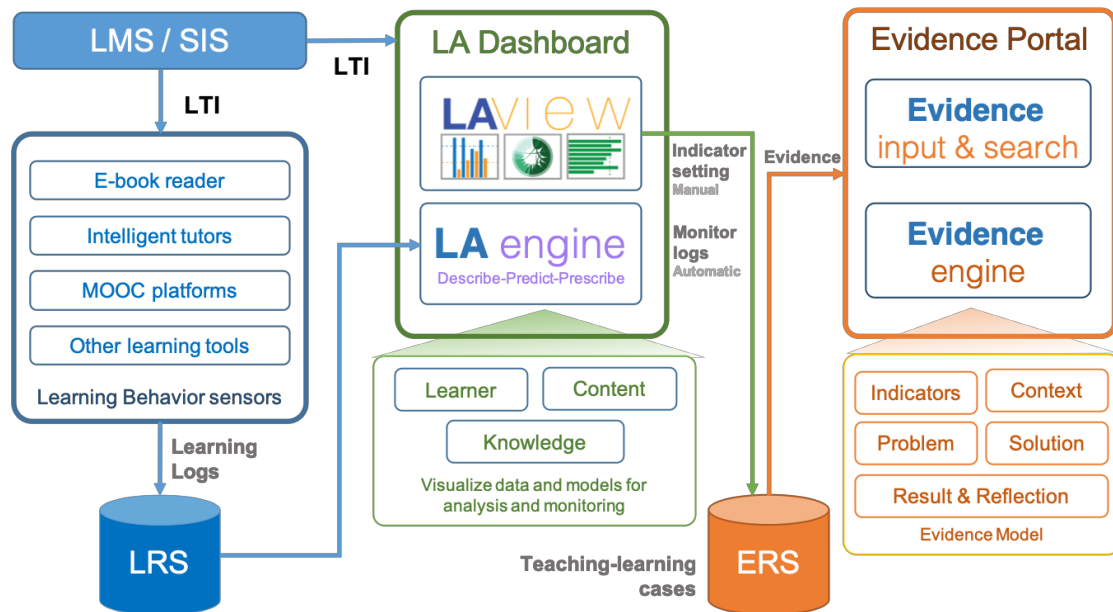


Figure 1: Components of the LEAF framework.

Using the LA platform helps to collect anonymous learning logs of students. For example, teachers can use a Learning Management System (LMS) to coordinate a course and upload reading content in an eBook reader linked to that LMS. BookRoll, the eBook system in our context assists instructors to support students' in-class and out-of-class learning activities. It has features to highlight important and difficult to understand text. Students can add memos or bookmark important pages. While students use BookRoll for browsing course material, their reading behaviors can be anonymously logged. Learning Logs of BookRoll reading is recorded in Learning Record Store (LRS) as an Experience API (xAPI) statements. We consider any similar tool which can log learner behavior as Learning Behavior Sensor (LBS). The LA dashboard has a backend LA engine and web-based front-end LAVIEW. The LA engine helps to analyze the log data and extract features and recording in database. This processed information and models regarding the learners, the content and their knowledge data is visualized in LAVIEW. The framework applies a two-way anonymization to the student data and supports all these processes in real-time. In the logs, students are represented by UUID to ensure their privacy. However, when user logs in to the system via LTI s/he can see the converted student ids based on their roles. Thus, the framework is also very flexible to connect to any other behavior sensors which has LTI. While the users interact in the dashboard to monitor and analyse the state of teaching and learning, the evidence portal gathers their interactions.

In this paper in the context of LAVIEW, we define the user workflow to gather evidence from practice and the corresponding features in the dashboard.

2 SUPPORTING ACTIONABLE ANALYTICS WITH LAVIEW

Our approach to designing technology-enhanced and evidence-based practice in education starts with systematically defining indicators of teaching-learning experiences in a specific scenario. These indicators are measurable attributes of the individual users or their interactions within the learning system. Our dashboard, LAVIEW, plays a central role to assist analysis of the visualized indicators to identify problems by teachers. Based on the problem that the teacher identifies, (s)he can think of possible solutions to mitigate it and then monitor its effectiveness. We are designing technology that can help to capture this process and reflect on the effectiveness of the practice as evidence. Conceptualizing such an evidence analytics system in education would push the boundaries of existing learning analytics infrastructures. We define the workflow for the teachers first.

2.1 Teachers Workflow Design

The teacher workflow is based on the DAPER model (see figure 2). The data collection is supported directly by the LA infrastructure. The indicators are either collected directly from data log or computed from the log. Typically, we envision that the learning analytics system developer would visualize various indicators based on the data that a particular system gathers and the features that are extracted from them. It is then visualized in dashboard to assist easier and useful interpretation by different stakeholders. For analysis phase, the teacher needs to specify the criteria to determining status of students based on those indicators. Based on the analysis, the teacher can implement certain intervention plan to mitigate the problem. Post-intervention the teacher can monitor the indicators while the system computes the change in the indicator values and stores it as results for the teachers to reflect on it. In the dashboard we want to assist the users monitor indicators, analyze and annotate status of problem based on those indicators and implement certain

solution interventions (for e.g. email based interventions as demonstrated in Majumdar 2019). This set of indicator-problem-solution-result-reflection is saved as a teaching-learning case in the evidence record store for further analysis.

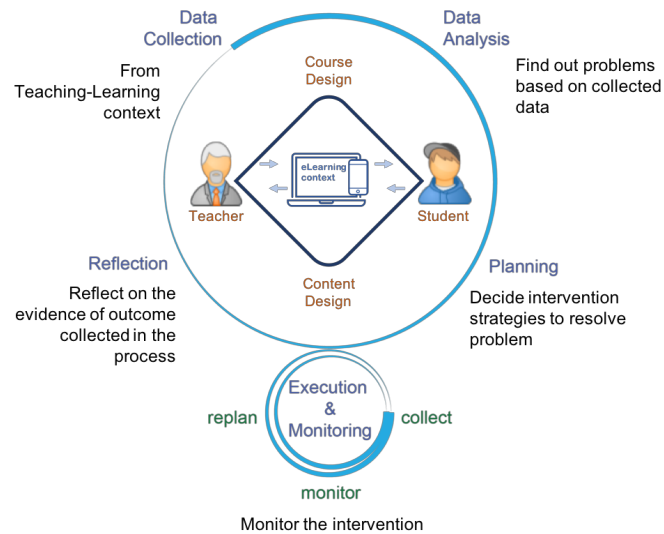


Figure 2: DAPER model-based workflow for intervention and evidence collection.

2.2 LAVIEW to support the teacher through DAPER workflow

We are designing LAVIEW dashboard as the unified tool that would assist instructors with the functionalities in the DAPER model. The collection phase is coordinated automatically by logging data from the ebook reader and Moodle. For analysis, teachers can first use the setting panel to set the criteria of each indicator based on which they get notification of the problem state. An example UI mock up is shown in Figure 3. Criteria can simple indicate desirable (green), ok (yellow), critical (red) zones based on the indicator value.

Displayed Indicators						Edit	Add	Info
Sl#	Display	File name	Indicator name	Description	Criteria			
1	<input checked="" type="checkbox"/>	marker.html	Marker count	Number of marker on the page	<div><div></div><div></div><div></div><div></div><div></div></div> 20			
2	<input checked="" type="checkbox"/>	memo.html	Memo count	Number of memos on the page	<div><div></div><div></div><div></div><div></div><div></div></div> 20			
3	<input checked="" type="checkbox"/>	reading.html	Reading completion	Portion of ebook read	<div><div></div><div></div><div></div><div></div><div></div></div> 100 %			
4	<input checked="" type="checkbox"/>	markerlist.html	Marker content	List of marker content in each page	<div>List of difficult words in content</div> <div>List of desired words in content</div>			
5	<input checked="" type="checkbox"/>	memolist.html	Memo content	List of memo content in each page	<div>List of difficult words in memo</div> <div>List of desired words in memo</div>			
6	<input checked="" type="checkbox"/>	pagejump.html	Reading sequence	Page jumps while reading				
7	<input checked="" type="checkbox"/>	enagement.html	Engagement score	Aggregated engagement score	<div><div></div><div></div><div></div><div></div><div></div></div> 100			

Figure 3: Criteria setting panel for indicators

In Figure 3, the indicators are in the context of BookRoll reading behaviors. The Markers and Memos are in terms of counts. Considering the act of annotating as active reading behavior, markers and memo count can indicate the level of active reading the learners are involved in. Similarly, reading completion and engagement are in terms of percentage and they can highlight the status of student's engagement. While setting the criteria if the teacher sets reading completion lower than 40% as low engagement, then dashboard can be used to notify both the teacher and the students respectively (see Figure 4a and 4b) for monitoring.

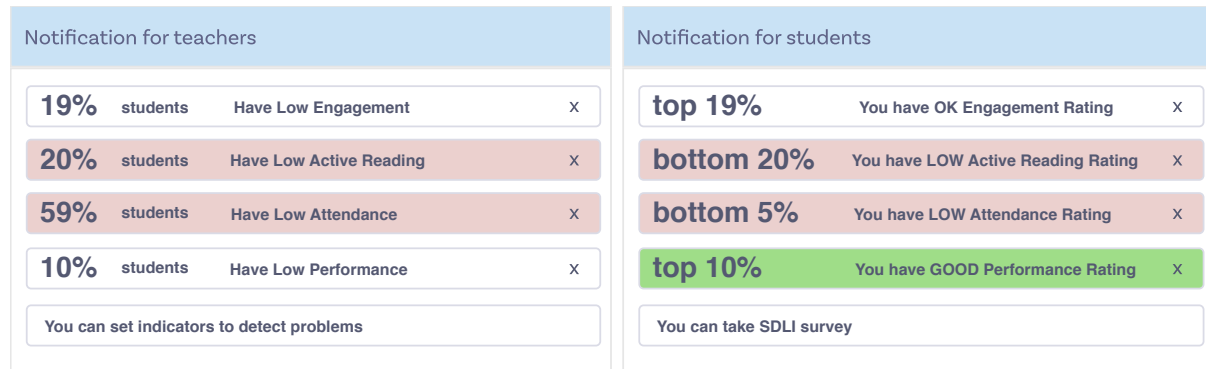


Figure 4: Notification panel for monitoring. Left – Teachers view, Right – Students view

2.3 Extracting evidence from teaching-learning practice



Figure 5: Sample information presented in LAView.

The dashboard contained various panels of visualized indicators for monitoring (see Figure 5). To assist users, we even add an overlay panel to every graph which gives explanation about each graph to the users. The ERS records all the information that is part of the earlier discussed workflow. It records the criteria set for each indicator, details of the context regarding which course and content,

the solution plan of intervention in case there is a problem identified and the result of the solution. These actions of analysis, planning and monitoring by the teachers are saved in as xAPI statements in the Evidence Record Store (ERS). Context anonymized dataset in the LRS can be used to retrieve the whole case details during evidence search. Each of these records are saved in the ERS as a teaching-learning case which can then be analysed for extracting evidence.

Figure 6 illustrates an example of an overall workflow. The teacher sets the criteria value for indicators which is saved in the system. Based on that criteria the system puts notification on the instructor's dashboard. The instructor can select to email the cohort of students in a particular criteria zone (red, yellow or green) by selecting a predefined editable message. Once the message is sent the indicator criteria, problem identified based on cohort definition, and intervention (email message) is saved in the teaching-learning case. After a period of designated time period the result of the intervention is also added to the case. Such a record captures the cycle of DAPER model and we plan to use the case for extracting evidence.

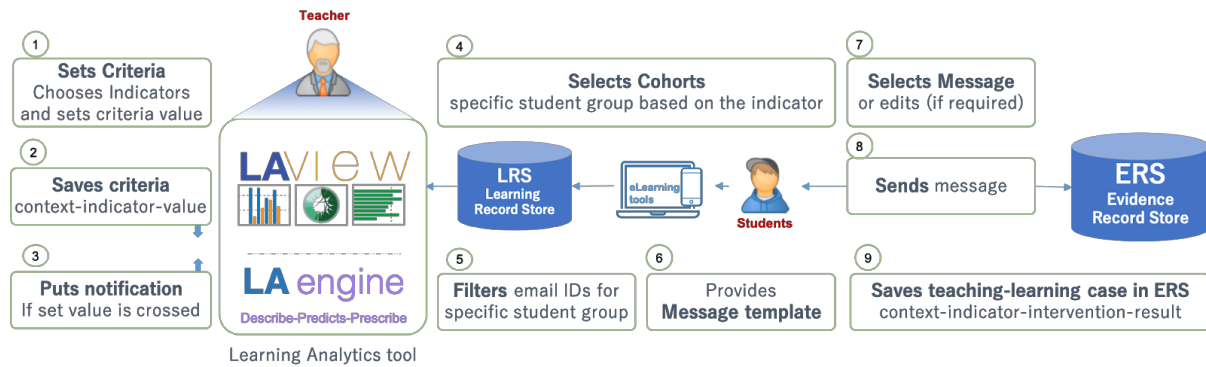


Figure 6: Example of an overall workflow with LAVIEW

An example of record of the TLC is presented in table 1 based on the previous example workflow.

Table 1: Description of the parameters captured in a Teaching-Learning Case (TLC)

Sl.	Parameter	Description	Example
A	Context	Details of course-student-content.	
1	<i>Institutional profile</i>	Details of the institute.	
1.1	Country	Name of the country.	Taiwan
1.2	Institution type	University/School/Corporate training.	University
2	<i>Course profile</i>	Details of the specific course.	
2.1	Field or Subject	Name of the course or subject.	Introduction to programming
2.2	Mode of instruction	How is the course offered, for example face-to-face, eLearning, MOOCs.	Face-to-face

2.3	Language	Language of instruction.	English
2.4	Time	The date and duration of the course.	Fall 2018, semester long
2.5	Units	Number of units.	14
2.6	Pre-requisite	Pre-requisites required for the course.	none
2.9	Class size	Number of registered students.	122
3	<i>Learner profile</i>	Details of the cohort of learners.	
3.1	Demographics	Distribution of learners.	Undergraduate first year
4	<i>Content profile</i>	Details of content.	
4.1	Learning content	Course content and its link.	<BookRoll link of content>
B	Indicators	Measurable parameters defining the problem and highlight results.	
5	Indicator definition	The definition of the data or its computed feature and description	Percentage completion
C	Problem	Problem identified through analysis	
6	Problem definition	Describes the problem and how to identify them from the indicators	Low engagement if percentage completion is less than 40%
D	Solution	Solution to mitigate the problem	
7	Plan definition	Description of the plan and associated content for it.	Email sent to low engagement students: <body message>
8	Review period	Period to review the indicators after the plan is implemented.	1 week
E	Results	Results of the implemented plan	
7	Dataset	The indicator values in the context across time.	<link to dataset 1 week before and after the intervention>
8	Reflection	Reflection of the teacher or student	<i>The tone of the message in the email seems critical for the motivating the low-engagement students.</i>
F	Metadata	The data related to the case	
9	Timestamp	The time the record was updated	ISODate("2018-12-03T20:48:08.099Z")
10	Rating	The rating of the case for the evidence.	4

3 PLANS FOR TESTING AND VALIDATION

The current system is under active development and we propose to open it up to teachers, such that they can use the various components in actual practice. We would follow a co-design paradigm by observing log data and getting feedback from the teachers who use the system.

3.1 Sample

To conduct such a research, we selected teachers who were already trained for offering course by using some LMS. We invited the teachers who successfully completed a MOOC based faculty development course on Educational Technology. Total 533 participants completed this course. Participants were from across 16 different states in India. These instructors were across 15 disciplines including Engineering, Humanities, Language, Science, Law, Pharmacy and Commerce to name some. Majority of them are from Engineering (377) and in that too in computer science (175). Teachers have diverse teaching experience 1 – 10 plus years.

3.2 Method

We offered the infrastructure associated with LEAF to the interested teachers and such that they can conduct their next semester-long course on the platform. We choose Moodle as the LMS. Teachers shall use BookRoll as the ebook-reader and the LAVIEW dashboard with that Moodle. We shall set-up a course on the same moodle and register the teachers there. This course would be used for coordination and training of the various components in the system.

While the teachers conduct their course, we shall log their Dashboard components utilization. We plan to gather an initial dataset of teaching-learning case from this pilot. It shall help us to validate the process and the actual structure of the collected data too.

To initiate this in an immersive and pertinent way (Warriem, 2014), we had a face-to-face workshop with teachers during mid-December 2018 following which we launch the coordination course on the Moodle.

4 CONCLUSION

In this research article we take a position to extend the notion of evidence in the evidence-based education from meta-analysis of published works to educational BIG data gathered from actual teaching-learning scenarios. This complements the existing research-based evidence by finding evidence in practice. Based on LEAF, a framework design which defines and supports gathering all the associated parameter from such an instance of practice, we illustrate a dashboard design to supports it. We give an example of teaching-learning case (TLC) that notes the context, problem, solution and indicators related to a teaching-learning scenario. It gives a micro-view of the evidence. A collection of such TLCs can be aggregated or analyzed based on its parameters to get a macro view of the evidence. We presented the details of the technical components and illustrate how it supports the DAPER workflow model to generate the evidence parameters and store it as xAPI statements. Also, keeping the components in LEAF as standard learning analytics infrastructure and standard data structures making it easier to adopt by interested institutions which has existing resources.

Our approach to commence an evidence-based practice in education supported by technology starts with systematically gathering indicators of learning in a specific scenario and then analyzing visualized indicators in the analytics dashboard to identify problems. Teacher can design intervention to mitigate it and then monitor its effectiveness. We believe technology can help to capture this process and reflect on the effectiveness of the practice as evidence. Conceptualizing such an evidence analytics system in education would push the boundaries of existing learning analytics infrastructures towards a technology-enhanced and evidence-based education and learning.

REFERENCES

- Davies, P. (1999). What is evidence - based education? *British journal of educational studies*, 47(2), 108-121.
- Ferguson, R., & Clow, D. (2017). Where is the evidence? : A call to action for learning analytics. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference* (pp. 56-65). ACM.
- Flanagan, B., & Ogata, H. (2017) Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In *Proceedings of the 25th ICCE 2017*, New Zealand, Nov 2018.
- Majumdar R., Yang Y.Y., Li H., Akçapınar G., Flanagan B. and Ogata H. (2018) Supporting Learner's Development of Self-Direction Skills using Health and Learning Data, In *Proceedings of 26th ICCE*, Manila, Philippines, Nov 2018
- Ogata H., Majumdar R., Akçapınar G., Hasnaine N.H, and Flanagan B. (2018) Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning, In *Proceedings of 26th ICCE*, Manila, Philippines, Nov 2018
- Warriem J.M., Murthy S., Iyer S., A2I: A Model for Teacher Training in Constructive Alignment for Use of ICT in Engineering Education, In *Proceedings of 22nd International Conference on Computers in Education*, Nara, Japan, November- December 2014

Behind the Scenes: Designing a Learning Analytics Platform for Higher Education

Irene-Angelica Chounta, Margus Pedaste, Katrin Saks

University of Tartu

{chounta, margus.pedaste, katrin.saks}@ut.ee

ABSTRACT: In this paper we share our experience from designing a Learning Analytics platform to support the needs of stakeholders from a higher-education academic institution in Estonia. We present the design framework and the architecture of our platform and we discuss how we aim to address challenges imposed by context. For the design of the platform, we carried out interviews with students, teachers and stakeholders from the institution's administration in order to gain insight with respect to the needs of users. Here, we report our findings from these studies, but we specifically focus on the teachers' perspective. Finally, we conclude to a discussion about lessons learned from our interviews with teachers and the proposed design framework of the LA platform in its first steps.

Keywords: learning analytics, design framework, teachers, requirements

1 INTRODUCTION

The use of computational methods to analyze the learning process and to improve the learning outcomes is commonly described by the term "*Learning Analytics*" (Siemens, 2013). Learning Analytics (LA) in Higher Education mainly aim to support students and instructors in monitoring, mirroring and guiding (Jermann, Soller, & Muehlenbrock, 2001) by providing adaptive and personalized feedback. Usually, feedback is offered through student or teacher dashboards using visualizations and graph representations (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). Such dashboards present informative statistics and visualizations of "meaningful" student activity. That is, student actions that may indicate either learning or some kind of disruption of the learning process (Dyckhoff, Zielke, Bültmann, Chatti, & Schroeder, 2012). It is argued that the use of metrics of student activity may provide false assessments of learning, mainly because such metrics come from data-driven approaches and are not theoretically grounded using pedagogical reasoning (Duval, 2011; Gašević, Dawson, & Siemens, 2015). In addition, activity metrics, charts and statistics can be interpreted in more than one ways leading to misunderstandings and misinterpretations (Spada, Meier, Rummel, & Hauser, 2005).

Our main objective is to design and implement learning analytics and feedback mechanisms to support the practice of stakeholders from the academic community of the University of Tartu. The University of Tartu is a leading centre of research and training in Estonia and it consists of 4 faculties: Faculty of Arts and Humanities, Faculty of Social Sciences, Faculty of Medicine and Faculty Science and Technology. It offers a wide range of bachelor, master and PhD study programs for about 13000

students¹. Stakeholders in this context are the students and the teachers (or instructors) of the university. The learning analytics we aim to design, are based on computational models that aim to assess the student's academic performance, to identify risks and to prevent possible failures (such as drop-outs) and to provide personalized and adaptive feedback to students. By computational models, we mean predictive approaches to assess a dependent variable (for example, academic performance) with respect to independent variables (for example, points earned in the current course from assignments, students' contribution in group projects or group discussions, resources access patterns, etc.). As data inputs, we use three data sources: a) demographics and data about the student's history, as recorded by the Study Information System (SIS) of the university; b) data from courses that the student has participated, as recorded by the university's Learning Management System (LMS); c) data from direct student input, such as questionnaires and learning artefacts (for example, homework). The goal is to use the assessments of the computational analytics to provide appropriate interventions (for example, feedback and recommendations) for students in order to improve learning outcomes and for teachers in order to support their practice.

We strive to follow an evidence-based (Ogata, Majumdar, Akçapınar, Hasnine, & Flanagan, 2018) design approach and design a computational approach that can be backed up by rigorous pedagogy. Most importantly, we want to provide tools to teachers and students that "make sense". That is, tools that can support their needs and that can be easily and effortlessly integrated in their every-day practice. Educational technologies and, in particular, learning analytics are topics that attract research interest. However, successful integration of new technologies and computational tools into the classrooms has been so far slow and hard to achieve (Ferguson et al., 2016). Teachers, in particular, feel disconnected from research outcomes and don't see how new technologies support their needs². In this paper, we present our experiences from designing a new learning analytics platform with the goal to bridge the gap between research and practice. In particular, during the design phase of the platform, we followed a socio-technical approach. We asked stakeholders (teachers, students, administration and policy makers) to contribute to the design by participating in interviews and focus groups. Here, we focus on the teachers' perspective, as it was captured in a focus group and we discuss how their input contributed to the design framework of the learning analytics platform.

2 METHOD OF STUDY

To support the design of the learning analytics platform, we conducted interviews and focus groups with stakeholders in two rounds (**Figure 1**). In the first round, the aim was to discuss with stakeholders potential LA mechanisms (in total, we asked the stakeholders to review 21 LA mechanisms) – both for students and teachers – to support different objectives of the contemporary learning approach (Pedaste & Leijen, 2018) and how we can adapt these mechanisms to facilitate our university's needs. For the first round, we carried out two focus group interviews. The first interview was conducted with teachers, program directors, LMS administrators and a specialist in educational technology (N=10), all having long-term experience with LMSs. The participants of the second interview were undergraduate students (N=6) who all had one or two years of experience with LMSs.

¹ <https://www.ut.ee/en/university/general>

² <https://www.edsurge.com/news/2018-09-26-what-can-machine-learning-really-predict-in-education>

In the second round, we focused on teachers' practices and needs. Therefore, we carried out a series of activities over an academic semester where we asked from four teachers – who were sharing a blended-learning course – to use a set of educational technologies to organize this course. At the end of the semester, we carried out a focus group discussion with 3 out of the 4 teachers who worked with us during the semester (from now on we refer to them as I1, I2 and I3). During the discussion we went over the teachers' work practices, we discussed about their needs and how technology addressed these needs, as well as their expectations from learning analytics. The discussion was facilitated by an experienced research in Human-Computer Interaction topics, Educational Technologies and Learning Analytics. The discussion lasted for about an hour and it was recorded - after having acquired the instructors' consent. After the end of the discussion, the recordings have been transcribed and analyzed.

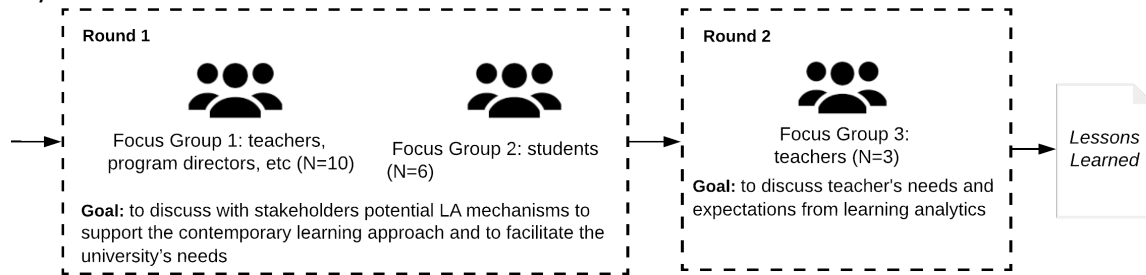


Figure 1. The interviews and focus groups process described in this work

3 LESSONS LEARNED

Here, we present the outcomes from the two discussion-with-stakeholders rounds in the form of “lessons learned”. In this sense, we use stakeholders’ input as practical guidelines that can support us in the design and implementation of our LA framework. From the first round, the analysis of the focus groups discussions showed that self-regulation and subject knowledge acquisition makes sense for stakeholders to be supported in combination by LA. Stakeholders also stated that more attention should be paid to supporting collaboration and subjective well-being. At the same time, it was mentioned that the value of LA requires more in-depth analysis. Our focus group interviews showed students to be slightly more positive towards using different applications of LA, whereas only a few scenarios were considered useful by most of the teachers and high variation was found in teachers’ questionnaires replies (Saks, Pedaste, & Rannastu, 2018). In order to further explore the high variation in teachers’ perspective regarding LA, we conducted the second round of focus group discussions (described in section 2) only with teachers. This discussion was structured in three parts that explored teachers’ user experience with educational technologies, their perception about usability of such technologies and future directions that could be supported by LA.

The analysis of the discussion showed that teachers are in favor of LA tools that support tracking the progress of students with respect to competencies or skills and they envisioned a technology that would allow them to track progress regarding different activities in one bigger picture (I3: “*We wanted our students to upload their tasks, their pictures and we wanted to see how they change these... we wanted to see their progress*”). They pointed out that the nature of the course did not allow them to act on a predefined plan, but they had to adapt their teaching strategies to the students’ needs (I3: “*it was the professional development course and you lay on students’ needs*”). This made the need for LA tools to support their practice more prominent. Due to the blended-learning nature of the course, the teachers used in combination various technologies (for example Google Apps, LMS and other

educational software). This made it difficult for them to track the students' progress and interactions with learning objects and therefore the teachers pointed out that there was a need for a tool that would provide them with an overview of students' activity (I3: *"we would like to see how the students make these changes... we wanted to see the progress but couldn't find the right tool for that"*). With respect to the way we present information about students' activity, the teachers first of all mentioned issues of privacy. In particular, the teachers informed us that students are usually uncomfortable when sharing information or materials with their peers (I2: *"in my group the problem was that this was visible to everybody and the students said I don't want to put anything there"*). At the same time, teachers have concerns about the visibility level of their own materials and information. When they don't have a clear idea about the visibility status of their activities, it makes them feel uncertain and leads then to take additional action (for example, to send emails) in order to confirm that the students can see certain information.

For the last part of the focus group, we asked the teachers to discuss what kind of expectations they have from technology. Teachers stated that they strongly feel the need for tools that will support them to manage their time efficiently and at the same time allow them to have a clear picture of how (and how often) the students engage with learning material and activities. This helps them to assess the students' progress and to plan their teaching strategies (I2: *"for me it's important to know that the student has not disappeared, but he visits from now and then. Another thing I follow is that they regularly practice their exercises. If they don't, I usually send out emails and remind them"*). They pointed out the need for tools that present basic traffic information. We followed up and asked them what kind of input they would like to receive from the system (visualizations, alarms, text messages). The instructors responded that graphs and percentages are difficult to read and require time to understand and interpret. One of the instructors referred to a past brainstorm session they had and brought up an idea from this session: *"the idea was that when a student has not logged in for a number of days, then the program automatically sends the student a little friendly note e.g. 'is everything ok?' 'please come and visit us'. At the same time the teacher will also receive a note that these students received that messages. If a student repeats this behavior, then the teacher gets a report based on the number of messages a student has received"* (I2). The same instructor stated that it is important for them not to have to follow each and every student on a regular basis but only to receive information on critical issues. The other instructors agreed that they are in favor of some kind of automated assessment that they could use to further investigate but they also pointed out that they would like to control the amount of information they receive (I3: *"I'm not sure I want too much information automatically. Maybe I prefer to do that manually"*).

4 PRACTICAL IMPLICATIONS AND FUTURE WORK

We used the input from the focus groups in order to inform the design framework of the learning analytics platform at the University of Tartu (**Figure 2**). In particular, we focused on serving teachers' needs and therefore we put emphasis in their requirements. An outcome from our discussions with teachers was that even though teachers want to have a clear picture about students' activity, they often don't have time to review visualizations about students' progress or to go through statistics. On the contrary, they would prefer to receive automated or semi-automated messages or assessments that would use in order to further investigate specific cases. To that end, our design framework integrates LA tools that provide teachers with automatic assessments of student's performance or explicit alerts of potential problems. Such tools aim to assist teachers in adapting to student's needs

easier, faster and to support them in deciding whether an intervention (and potentially what kind of intervention) is necessary (Chounta & Avouris, 2016; Holstein, McLaren, & Aleven, 2017).

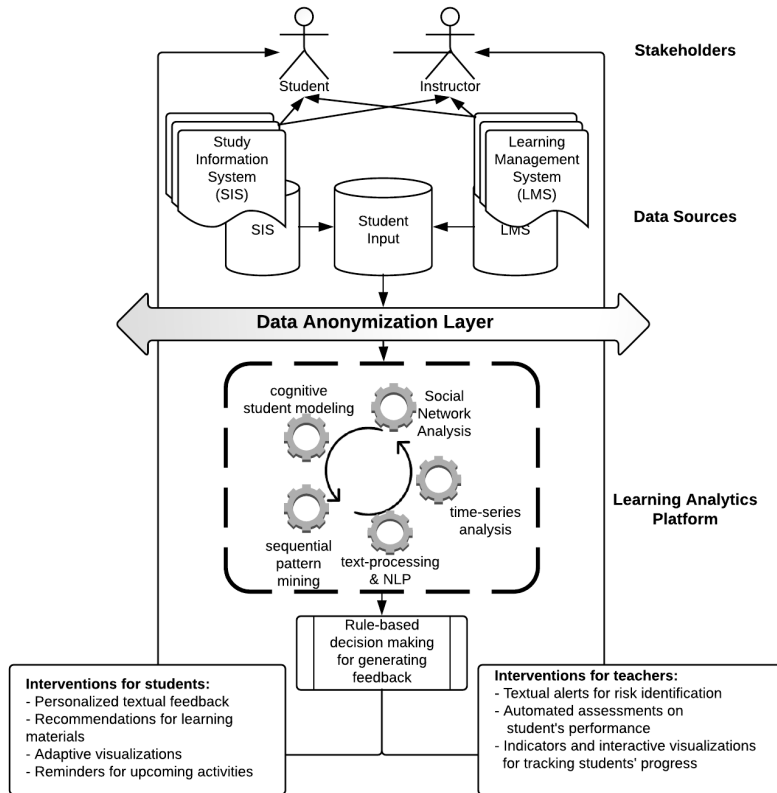


Figure 2. The proposed design framework of the Learning Analytics Platform at the University of Tartu after the interviews with stakeholders

Teachers also stated that it is important for them to track the progress of students with respect to specific skills and competencies. This is a well-established practice: Intelligent Tutoring Systems use the concept of Mastery Learning in order to provide learning materials or feedback to students who practice specific skills. To do that, they maintain individual student models (one model for each student) that provide an assessment of the student's knowledge state (Corbett, Koedinger, & Anderson, 1997). Similarly, we aim to apply cognitive modeling approaches to capture cognitive development (Chounta, Albacete, Jordan, Katz, & McLaren, 2017) and dynamic competence assessment of individual students using learning analytics to assess students' performance. Achieving this step will bring us closer to providing personalized and adaptive feedback to students as well as informative monitoring mechanisms to support teachers in planning and guiding.

ACKNOWLEDGEMENTS

This work was partly supported by the Estonian Research Council grant PSG286.

REFERENCES

- Chounta, I.-A., Albacete, P., Jordan, P., Katz, S., & McLaren, B. M. (2017). The "Grey Area": A computational approach to model the Zone of Proximal Development. In *European Conference on Technology Enhanced Learning* (pp. 3–16). Springer.

- Chounta, I.-A., & Avouris, N. (2016). Towards the real-time evaluation of collaborative activities: Integration of an automatic rater of collaboration quality in the classroom from the teacher's perspective. *Education and Information Technologies*, 21(4), 815–835.
- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In *Handbook of Human-Computer Interaction (Second Edition)* (pp. 849–874). Elsevier.
- Duval, E. (2011). Attention Please!: Learning Analytics for Visualization and Recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 9–17). New York, NY, USA: ACM. <https://doi.org/10.1145/2090116.2090118>
- Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, 15(3), 58–76.
- Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., ... Vuorikari, R. (2016). Research evidence on the use of learning analytics: Implications for education policy.
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71.
- Holstein, K., McLaren, B. M., & Aleven, V. (2017). Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 257–266). ACM.
- Jermann, P., Soller, A., & Muehlenbrock, M. (2001). From mirroring to guiding: A review of the state of art technology for supporting collaborative learning. In & K. H. P. Dillenbourg A. Eurelings (Ed.), *Proceedings of the European Conference on Computer-Supported Collaborative Learning EuroCSCL-2001. Maastricht, The Netherlands* (pp. 324–331). Maastricht, Pays-Bas. Retrieved from <http://telearn.archives-ouvertes.fr/hal-00197377>
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, B. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning. In *26th International Conference on Computers in Education*.
- Pedaste, M., & Leijen, Ä. (2018). How Can Advanced Technologies Support the Contemporary Learning Approach? In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 21–23). IEEE.
- Saks, K., Pedaste, M., & Rannastu, M. (2018). University Teachers' and Students' Expectations on Learning Analytics. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 183–187). IEEE.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400.
- Spada, H., Meier, A., Rummel, N., & Hauser, S. (2005). A New Method to Assess the Quality of Collaborative Process in CSCL. In *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The Next 10 Years!* (pp. 622–631). Taipei, Taiwan: International Society of the Learning Sciences. Retrieved from <http://dl.acm.org/citation.cfm?id=1149293.1149375>
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10), 1500–1509.

LASAT: Learning Activity Sequence Analysis Tool

Shitanshu Mishra^a, Anabil Munshi^b, Marian Rushdy^c, Gautam Biswas^d

Institute for Software Integrated Systems, Vanderbilt University

Emails: {^ashitanshu.mishra, ^banabil.munshi, ^cmarian.n.rushdy,

^dgautam.biswas}@vanderbilt.edu

ABSTRACT: Learning Activity Sequence Analysis Tool (LASAT) is a collection of sequence analysis algorithms developed at Institute for Software Integrated Systems, Vanderbilt University, with the purpose of extracting and interpreting students' learning behaviors extracted as frequent patterns (sequence of activities) from their activity traces logged in computer-based learning environments. LASAT includes several algorithms for analyzing temporal sequence data – such as, sequential pattern mining (SPM), differential sequence mining (DSM) and Hidden Markov Model-based learner modeling. LASAT also includes tools for pre-processing and organizing log data for analysis. In this paper, we present the LASAT toolkit with an aim of making these algorithms accessible to the wider community of researchers and practitioners. We review cases from the learning analytics literature, which have employed LASAT algorithms to demonstrate the use of the tool in supporting evidence-based pedagogical decision making, specifically in the context of learner modeling in computer-based learning environments (CBLE). This paper demonstrates the applicability of LASAT for a range of applications that span from studying learners' cognitive and strategic processes to affect transitions that together form the basis for understanding self-regulated learning processes.

Keywords: Sequential pattern mining, sequence clustering, evidence-based pedagogy, technology enhanced learning environments, intelligent tutoring systems

1 INTRODUCTION

“One of the factors leading to the recent emergence of learning analytics (LA) and educational data mining (EDM) is the increasing quantity of analyzable educational data” (Baker & Inventado, 2014). LA and EDM methodologies have been transforming educational research by providing means to extract useful information from large sets of learning-teaching datasets. The applicability of these methodologies spans a range of educational technology and learning sciences research and practice. A review of LA and EDM research by Papamitsiou and Economides (2014) had suggested four distinct major axes of LA/EDM empirical research: 1) Pedagogy-oriented issues that include but are not limited to student modeling, prediction of performance, assessment and feedback, reflection and awareness; 2) Contextualization of learning that includes the positioning of learning within specific conditions and attributes; 3) Networked learning that focuses on the social aspect of learning, for example in the case of MOOCs, where learning is a product of interactions among large-scale diverse cohort; and 4) Educational resources handling that focuses on intelligent recommender systems which organize and suggest educational resources from pools. This paper is delimited to the learning analytics techniques which address pedagogy-oriented issues spanning from studying learners' cognitive and metacognitive processes or affect transitions. More specifically, these LA techniques study learners' self-regulated learning (SRL) behaviors by extracting and analyzing patterns from learner activities, in relation to the prediction of performance, metacognition, and self-awareness,

generating evidence-based (Ogata et al., 2018) feedback services and recommendation of resources.

We present an LA-toolkit, Learning Activity Sequence Analysis Tool (LASAT), which is a collection of sequence analysis algorithms that help in finding and interpreting students' learning behaviors extracted as frequent patterns from their activity traces logged in computer-based learning environments. These algorithms were developed at Institute for Software Integrated Systems, Vanderbilt University. The algorithms in the current version of LASAT include Sequential pattern mining (SPM), Differential Sequence Mining (DSM) and Hidden Markov Model based sequence analysis (HMM). The LASAT toolkit aims at making these algorithms accessible to the wider community of researchers and practitioners. This paper presents a review of cases from LA and EDM literature, which have employed LASAT algorithms and demonstrated how LASAT supports evidence-based pedagogical decision making.

In order to carry out the literature review we followed four distinct steps: a) Searching for literature that has employed any of the LASAT algorithms, b) Reviewing this literature to identify their primary studies, c) Analyzing the studies to identify the context, LASAT algorithm used, indicators, and results, and d) Synthesizing these results to identify the plausible contexts where LASAT can be useful and the educational constructs that LASAT can provide insight into.

2 ALGORITHMS IMPLEMENTED IN LASAT

Computer-based learning environments (CBLEs) have the capability to log large volumes of data that capture learners' interactions with the environment at a very fine-grained level. Since learners' activities within the learning environment are a product of their internal cognitive and metacognitive processes, capturing these activities in the log traces can provide an opportunity for researchers to identify interesting learning behaviors from this activity data that give an insight into learners' self-regulated learning (SRL). Panadero et al. (2016) have shown how the definition of SRL has shifted over the years from a trait (that is static and can be analyzed simply by self-report questionnaires) to a process (that is dynamic and unfolds as a sequence of cognitive, metacognitive and affective processes). It is therefore apparent that CBLE logs, which capture sequences of learners' activities, are particularly suited to analyzing SRL in its most current and widely accepted process-based definition. The LASAT toolkit contains several algorithms (discussed below) to analyze learners' SRL processes from logged activity sequences.

2.1 Sequential pattern mining

Sequential pattern mining (SPM) involves the extraction of patterns from sequence data (Agrawal & Srikant, 1995). LASAT provides an implementation of SPM that extracts patterns from learners' action sequences in a CBLE. If such patterns are identified above a minimum frequency threshold for a majority of observed learners, they could be interpreted as "learning strategies" that learners applied within the environment. SPM may also be used to extract patterns involving other aspects of SRL such as affective states. Patterns obtained from mining sequences of affect observed during learning may be inferred as learners' affective transitions. In the presence of more contextual information such as learner performance, these patterns obtained from SPM can be analyzed and interpreted in greater depth.

LASAT SPM Module

- **Input:** Learners' logged activity/affect sequences
- **Output:** Frequent "patterns" or action/affect-subsequences observed in input data
- **Parameters:** i-frequency (frequency of occurrence of a pattern for a learner), s-frequency threshold (minimum percentage of learners who exhibit a given pattern), max-gap between consecutive items in the input sequences considered for finding patterns

2.2 Differential Sequence Mining

The DSM module (algorithm described in Kinnebrew et al., 2013) is an extension of the SPM module, where frequently observed patterns can be compared between two pre-defined groups of learners (experimental vs control, high performers vs low performers, collaborative vs individual learners, etc.). SPM is performed separately for each of the two pre-defined groups, and patterns which show statistically significant differences between the two groups are shown as output. This helps understand behavioral differences between two groups of learners (e.g., strategies that are used frequently by high performers versus those used frequently by low performers).

- **Input:** Category that differentiates learners in each group (e.g.: score>10=High performer, score<1=10=Low performer), Logged activity sequences for each group
- **Output:** Frequent "patterns" observed only in group1, only in group2, and in both groups
- **Parameters:** Statistical significance test used (t-test/Mann-Whitney), max p-value considered, i-frequency, s-frequency threshold, max gap between consecutive input actions

2.3 Bayesian HMM – based Clustering

The HMM-clustering module performs probabilistic clustering of learners based on Hidden Markov Models (HMMs) generated from logged action sequences. While SPM and DSM provide specific behavioral patterns observed in learners, HMMs allow for an overall temporal learner model where learners shift between a number of (hidden) states (*most probable actions in each state are identified by the emission probabilities associated with the actions in that state*). HMM-clustering goes one step beyond generating HMMs - by also grouping learners into clusters based on the HMMs generated from their action sequences. The HMM-clustering algorithm implemented in LASAT uses the Bayesian Information Criterion (BIC) to determine the best number of hidden states, and Partition Mutual Information (PMI) to determine the best number of clusters.

- **Input:** Logged action sequences of each learner
- **Output:**
 - PMI value for different number of clusters (to find the *bestNumberOfClusters*)
 - BIC values for different number of states (to find the *bestNumberOfStates*)

- HMM generated with *bestNumberOfStates* for each of the *bestNumberOfClusters* clusters (initialState probabilities, state transition probabilities, emission probabilities for actions in each state) Learners in each of the *bestNumberOfClusters* clusters
- **Parameters:** maxNumberOfStates, maxNumberOfClusters, numberOfRepeats

3 LASAT USER INTERFACE

The landing page of the LASAT user interface contains the list of various algorithms available in the LASAT toolkit. Users can select any of the algorithms to start the data analysis process. Figure 1 shows the analysis screen for one of the algorithms (SPM). On the left pane of the analysis-screen, users can configure the algorithm parameters and set the input datasets. The right side of the analysis-screen compiles the results corresponding to different input parameters.

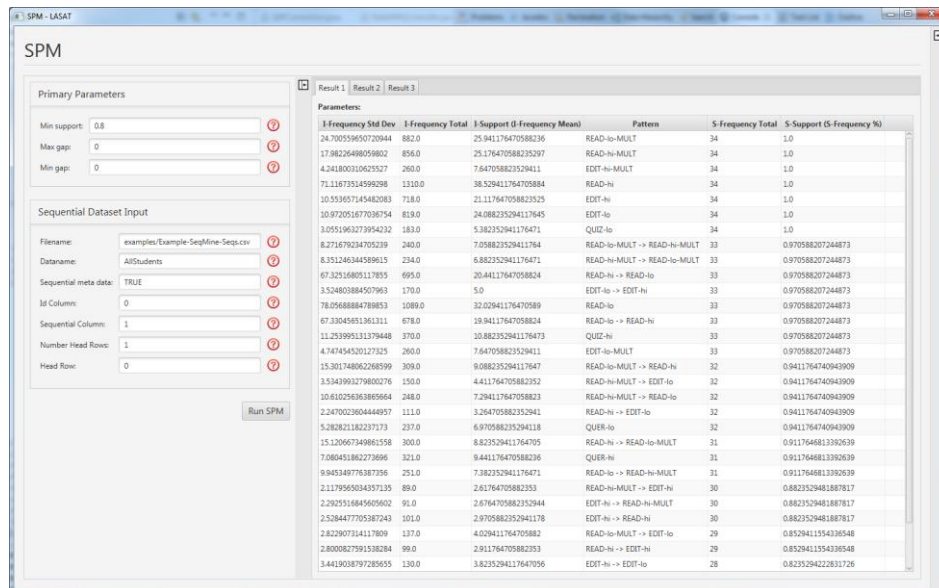


Figure 1. Analysis-Screen of LASAT

4 DIFFERENT CASES THAT HAVE EMPLOYED LASAT ALGORITHMS

LASAT algorithms are employed in the analysis of sequence logs generated from various computer-based learning environments. Table 1 shows an exemplary list of such cases. Munshi et al. (2017), Kinnebrew & Biswas (2012) and Jeong, et al. (2010) analyzed data coming from various deployments of Betty's Brain, an open-ended learning environment where learners model causal relations by teaching a virtual agent named Betty. Different learner actions in Betty's Brain include reading of resource pages, editing causal links between concepts in the concept map, quizzing Betty to test her knowledge, etc. Jiang, et al. (2015) analyzed data coming from Virtual Performance Assessments (VPAs). "VPAs are online 3D immersive virtual environments that assess middle school students' science inquiry skills, in line with state and national standards for science content and inquiry processes" (Baker & Clarke-Midura, 2013). Various learner actions within VPAs include navigating around the virtual environment, making observations, gathering data, conducting laboratory experiments, etc. Bouchet et al. (2012) have worked on data generated while learners interacted with MetaTutor, a multi-agent and adaptive hypermedia-based ITS that fosters self-regulated

Table 1: Example cases that have employed LASAT algorithms

Reference	Algorithm	Data/Context	Construct addressed	Result
Munshi et al. (2017)	DSM	Learner actions data in Betty's Brain from 87 six graders. Climate change topic.	Identifying sequences of learner actions which are statistically significantly different between low performers and high performers	Three sequences of actions found to be significantly different. Two of them frequent in high performing groups and one in the low performing group.
Jiang et al. (2015)	SPM	Action logs in Virtual Performance Assessments (VPAs) system from 2,431 students in grades 7-8 in science classes.	Identifying student patterns of behavior over time, different between novice and experienced students	Novice students engaged in more exploratory behaviors as compared to more experienced students.
Kinnebrew and Biswas (2012)	DSM	Learning interaction trace data in Betty's Brain from middle school class	Identify and compare segments of students' productive and unproductive learning behaviors	Identified differentially significant action patterns for productive and unproductive learning behaviors, in the cases of both 'High' and 'Low' performers.
Bouchet et al. (2012)	DSM	51 college students activity data. Complex science topic in MetaTutor	Identifying differentially frequent activity patterns between the student groups and interpret these patterns in terms of relevant learning behaviors.	High-performing students tend to be better at quickly identifying the relevance of a content to their sub goal, are more methodical and strategic.
Jeong et al. (2010)	HMM	Activity sequences from Betty's Brain coming from 6,298 learners (adult professionals)	Analysis of productive learning behaviors in a structured inquiry cycle	High-performing students have more linear and consistent learning behaviors.

learning by presenting challenging human biology science content to learners. The logged learner actions are related to reading, monitoring and use of strategy.

Munshi et al. (2017) presented an application of the DSM algorithm to identify patterns of frequent learner actions (interpreted as "learning strategies") which are significantly different ($p < 0.05$) between high and low performing learners while modeling causal relations related to climate change in the Betty's Brain environment. Kinnebrew and Biswas (2012) employed the DSM algorithm to identify and compare segments of students' productive and unproductive learning behaviors, again in the context of climate change topic and Betty's Brain environment. Bouchet, et al., (2012) employed DSM to identify differentially frequent activity patterns between the student groups in the context of Biology topic in MetaTutor. Jiang et al. (2015) presented an application of SPM to identify

student patterns of behavior over time, different between novice (students who are new to VPAs) and experienced students (students who have worked with VPAs before). Jeong et al. (2010) have shown an application of HMM algorithm to analyze productive learning behaviors in a structured inquiry cycle in the context of Betty's Brain used by adult learners. In addition to the list shown in the table 1, there are number of other research (for example, Biswas, et al. (2014), Kinnebrew, et al. (2013), Biswas, et al. (2010), Jeong, et al. (2008), Li & Biswas (2002), etc.)) that demonstrate the application of LASAT algorithms for range of research purposes.

5 DISCUSSION AND CONCLUSION

From the use cases described in the previous section, we see that LASAT algorithms are useful in many ways. In the cases of Munshi, et al. (2017), Bouchet, et al. (2012) and Jeong, et al. (2010), we see that LASAT algorithms can help in empirically defining the behavioral learning characteristics of high performers and low performers. Similarly, in the case of Kinnebrew & Biswas (2012), DSM helped in characterizing patterns of actions as productive and unproductive learning behaviors. Identification of these characteristics has manifold benefits for ITS designers and researchers. Firstly, the identification of productive and unproductive action sequences can provide insight into how learners interact with the features of the learning environment and how does any specific pattern of interaction support or impede learning. Secondly, the characterized action patterns serve as models that can inform time and content of the scaffolds (e.g. formative feedback) to be provided to the learners. It should be noted that Kinnebrew & Biswas (2012) extracted two levels of action sequence insights: a) first is the characteristic behavior(s) of high vs low performers, and b) the second is an account of the productive and unproductive action sequences within the high and low performing cohorts. This two levels of evidence are crucial because one may want to use the productive action sequences within the low performing cohort to inform the design of the scaffolds for the low performers, instead of looking at the prominent action sequences of the high performers. Jiang, et al. (2015), on another hand, have shown that LASAT can also help researchers in understanding the differences in the characteristics of novices and experienced users of the learning environment, i.e. LASAT can extract evidence about how crucial is the learner's familiarity with the learning environment for the achievement of her learning objectives.

LASAT algorithms enable the researchers to extract finer evidence by analyzing learner's micro-actions in any technology-enhanced learning environment. These evidences can be further used to a) Develop a rich understanding of local learning mechanisms; b) Design or curate appropriate pedagogical responses for the learners; c) Determining the time when any pedagogical response should be given to the learner; d) Develop rich understanding about similarities and differences between different cohorts of learners. LASAT toolkit can be used by any researcher or practitioner who wishes to collect sequential learning data. In future, we wish to add more algorithms to the toolkit. We would also like to build a generalized software framework where anyone from the community can append new algorithms to the suite. We also look forward to applying LASAT algorithms to a more varied scenario. The LASAT tool can be accessed by submitting a download request at: <https://wp0.vanderbilt.edu/oele/software/>.

REFERENCES

- Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Data Engineering*, 1995. Proceedings of the Eleventh International Conference on (pp. 3-14). IEEE.

- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer, New York, NY.
- Baker, R. S., Clarke-Midura, J. (2013). Predicting successful inquiry learning in a Virtual Performance Assessment for science. In *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization*, 03-214.
- Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & ROSCOE, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, 5(02), 123-152.
- Biswas, G., Kinnebrew, J. S., & Segedy, J. R. (2014, June). Using a cognitive/metacognitive task model to analyze students learning behaviors. In *International Conference on Augmented Cognition* (pp. 190-201). Springer, Cham.
- Jeong, H., Biswas, G., Johnson, J., & Howard, L. (2010, June). Analysis of productive learning behaviors in a structured inquiry cycle using hidden markov models. In *Educational Data Mining 2010*.
- Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008, June). Using hidden Markov models to characterize student behaviors in learning-by-teaching environments. In *International conference on intelligent tutoring systems* (pp. 614-625). Springer, Berlin, Heidelberg.
- Jiang, Y., Paquette, L., Baker, R. S., & Clarke-Midura, J. (2015). Comparing Novice and Experienced Students within Virtual Performance Assessments. *International Educational Data Mining Society*.
- Kinnebrew, J. S., & Biswas, G. (2012). Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. *International Educational Data Mining Society*.
- Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM| Journal of Educational Data Mining*, 5(1), 190-219.
- Li, C., & Biswas, G. (2002). Applying the hidden Markov model methodology for unsupervised learning of temporal data. *International Journal of Knowledge Based Intelligent Engineering Systems*, 6(3), 152-160.
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, Brendan. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning. In *26th International Conference on Computers in Education (ICCE 2018)*, at Manila.

Quantitative Evaluation of Concept Maps: An Evidence-Based Approach

Kapil Kadam ^a, Anurag Deep ^b, Prajish Prasad ^c, Shitanshu Mishra ^c

Indian Institute of Technology Bombay

{^a kapilkadam, ^b anuragdeep4949, ^c prajish.prasad, ^d shitanshu} @iitb.ac.in

ABSTRACT: Evidence-based approach in educational contexts advocates sustained interaction between practitioners and researchers. In this paper, we present a case of how a practitioner-researcher partnership can help in implementing a robust evaluation method from the analysis of evidence. In the present study, concept maps have been used as a tool for assessment of students' learning. Typically, teachers are required to evaluate concept maps manually. Efforts have been made to automatically evaluate concept maps which are still an active research area. We have come up with a semi-automatic evaluation of concept maps, using log data generated from a concept mapping tool. The log data contains various learner actions such as add/modify concepts and links. Manual evaluation is done on a subset of concept maps. We then apply learning analytic algorithms to the log data to generate a goodness model of the concept map. The generated model can be further used to evaluate the remaining concept maps. The model also shows that several learner actions, such as adding links and connections, seem to be predictors of good concept maps. We discuss how these predictors can serve as an evidence-based practice for teachers and students to incorporate these actions in their teaching and learning. In this paper, we discuss the research study and analysis method in the context of evidence-based practice.

Keywords: Evidence-Based Approach, Learning Analytics, Concept Map

1 INTRODUCTION

In the 21st century, teaching-learning continuum focuses on an evidence-based approach. Prerequisites of this approach are the alignment between curriculum, pedagogy and assessment (Luke, 1999). There has been growing emphasis on collection, analysis and interpretation of information about learners to inform teaching and learning. The process of evidence collection, analysis and interpretation must be an explicit and accountable one to achieve quality educational outcomes by students. The value of the evidence is in its understanding followed by applying appropriate strategies to improve student learning. Practitioners who effectively use assessment data have been pioneers in bringing change in the local classroom (Protheroe, 2001).

Hsieh & O'Neil (2002) reported that the concept map strategy is simple to use and effective on problem-solving along with meaningful learning. Besides these, it also affects learners' achievements and interests (Aghakhani, et al. 2015). Research studies show that the concept mapping can significantly improve learning when compared with lecturing. Concept maps have been widely used as a formative assessment and conceptual knowledge representation tool. In this study, we explored the application of evidence-based practice for evaluation of concept maps. A semi-automatic evaluation was performed in which sixty-three concept maps were manually categorised into three categories. The analysis of the log data was done using rules based and decision tree

algorithms for further evaluation. The concept maps generated during the intervention were used as evidence of students' learning. The analysis of concept mapping steps served as evidence of productive and unproductive learner actions. We then mapped the steps performed during the evaluation to the LEAF framework (Ogata, et al. 2018).

2 BACKGROUND RESEARCH

2.1 Evidence-Based Practice

In a teaching-learning scenario, evidence includes teacher observations, tests, peer assessments and, formative and summative assessments. It can be used for assessment at different levels which include individuals, groups, courses etc. Learning performance of students can be improved by using evidence in the following ways (Bruniges, 2005; Kvernbekk, 2017):

- As a diagnostic method to improve teaching
- As a motivational method to focus on learners' strengths and weaknesses
- As a means of communication of learners' achievements or course effectiveness

2.2 Concept Map and Its Assessment

Novak & Cañas (2008) described concept maps as graphical tools used to represent and organise knowledge. It consists of concepts, connected by linking phrases. Two concepts connected by a linking phrase represent a unit of meaning (propositions).

Traditionally concept maps have been evaluated based on criteria or via a human-based rubric. The criteria map represents an expert's map which is then used to compare with the concept map of learners. It ensures control of quality and quantity of propositions. Based on the difference identified, instructors give appropriate feedback to the learners (Trumpower & Sarwar, 2010). Concept map assessment varies regarding the emphasis on its features. For some, the emphasis is more on hierarchy and cross-links while others focus on the number of concepts. Typically, they are evaluated manually, which is time-consuming and tiring. Active research has been done to address this issue, and many automatic assessment methods have been proposed based on a computerized assessment (Hirashima, et al. 2011; Pailai, et al. 2017). Most software provides functionalities of construction of concept maps along with automatic assessment. The automatic assessment compares the generated maps with the criteria map for effective assessment. Traditionally these automatic assessments have been criticized for no flexibility during assessment which can be implemented while doing the manual assessment. The computerized assessment requires the strict rules for calculating the concept map score. Attempts have been made to increase flexibility by including features like assessment using graph theory, synonym words etc. There hasn't been much emphasis on the application of learning analytics on concept map data. Wu et al. (2016) examined learners' behavioral patterns of a concept map tool in a collaborative concept map-based online discussion environment. The analysis aimed to understand how the collaborative concept map activity enhanced discussion and social knowledge construction. Wang et al. (2017) developed a concept mapping tool that offers navigational support in the form of hyperlinks, where nodes in the concept map are linked to segments of text. Concept map features (such as total nodes, total links, link/node ratio) and learner actions (such as total actions, navigation actions) were used to model

the generative strategies of learners. However, their focus was not the assessment of the concept maps but to examine how the navigational support and visual aids in concept mapping support generative learning.

3 EVALUATION

3.1 Procedure and Instrument

In this paper, we report an evaluation study in which video lecture was used to teach two topics on Tree data structure and Linked list data structure in a computer science (CS) undergraduate course. After watching the first video on trees, participants draw the concept map on the same topic by using a concept mapping software known as CmapTools (Cañas et al., 2004). The same activity was repeated for the second topic which was on the linked list. We conducted the study with first-year engineering undergraduate CS students of introductory programming course. The research question for this study was “What are the significant predictors in determining the quality of the concept map?”

3.2 Data Analysis

Following steps were performed to analyze the concept map log generated during the study. The steps are data compiling, data pre-processing and applying the algorithm (Figure 1). These steps will be discussed in the subsequent sections.

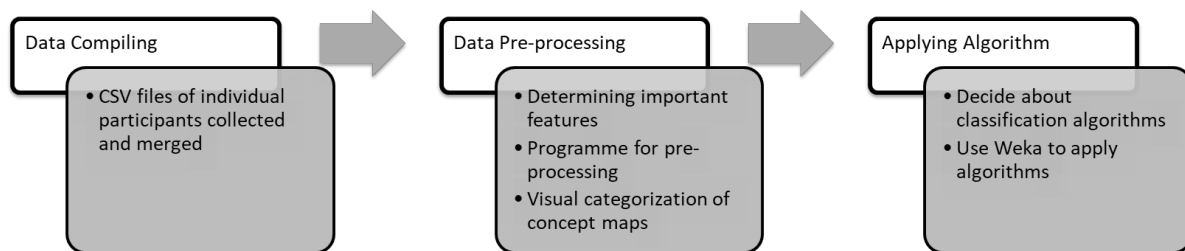


Figure 1: Data analysis process and corresponding steps

3.2.1 Data compiling

In total 63 concept maps were generated. Each of these maps had a raw file (*.cmap) and a CSV file. The CSV file included fields like DateTime [dow mon dd hh:mm:ss:mils zzz yyyy], user id, step number, event id, action type, entity type, entity id and entity description. An example of the values in for `<data, user id, step number, event_id, action_type, entity_type, entity_id, entity_desc>` is `<"Thu Oct 29 15:35:32:81 IST 201X", abc@gmail.com, 1, 1PKDX7MK4-1M6YRQK-159, Add, Concept, ge:1PKDX7MK5-1R553B4-15B, 'Linked list' x:43 y:40 w:44 h:26>`. In this step, all the CSV files were merged into a single file which was used for further analysis in the next step.

3.2.2 Data pre-processing

The following steps were performed in the data preprocessing phase.

- We combined action type and entity type to get concept mapping actions: *Add Concept, Add Connection, Add Linking Phrase, Delete Concept, Delete Connection, Delete Linking Phrase, Modify Text Concept, Modify Text Linking Phrase*.

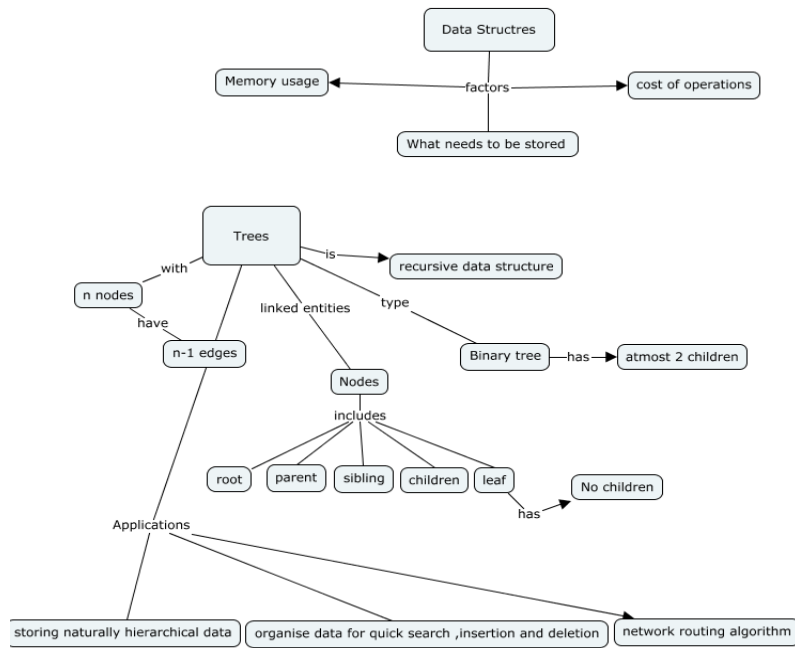


Figure 3: Representative of an average concept map

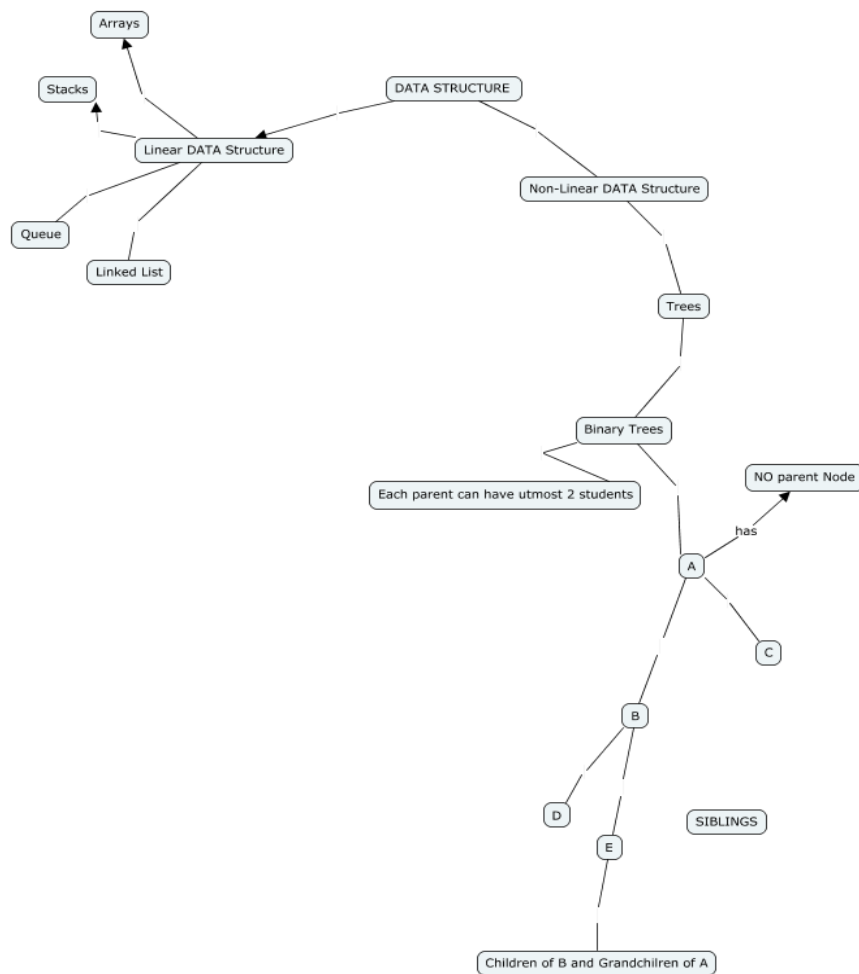


Figure 4: Representative of a bad concept map

3.2.3 Data Analysis

We used Weka software (V3.6.13) for further analysis and applied various Decision Tree and Rule-Based Classifier algorithms.

3.3 Results

We used these algorithms for analysis of concept map for trees (T) and for linked list (L). Besides this we used two combinations of features:

- **Feature Set 1:** Containing 12 features - *Add Concept, Add Connection, Add Linking Phrase, Delete Concept, Delete Connection, Delete Linking Phrase, Modify Text Concept, Modify Text Linking Phrase, number_concepts_added_deleted, number_connections_added_deleted, number_linking_phrases_added_deleted, CMAP_quality*, and
- **Feature Set 2:** Containing 5 features - *number_concepts_added_deleted, number_connections_added_deleted, number_linking_phrases_added_deleted, CMAP_quality*, and *number_of_steps*

Since we have combined the concept maps T and L, we realised that the average number of steps in both T and L are different. Hence we cannot include the number of steps as a feature. We also included Add and Delete features along with the delete ratio features. In all the cases the cross-validation folds were kept 5. The results of our analysis are summarised in Table 1. For example, in the row, Analysis of concept map T by using 12 features through rule-based algorithms (JRIP) yielded an accuracy of 60.46% along with one rule.

Table 1: Summary of analysis results

Concept map	Feature Set	Algorithm	Type	Accuracy (%)	Result
T	1	Rule based	JRIP	60.46	1 rule
T	1	Rule based	PART	62.79	5 rules
L	1	Rule based	JRIP	-	0 rules
L	1	Rule based	PART	55	2 rules
T + L	2	Rule based	JRIP	68.25	2 rules
T + L	2	Rule based	PART	68.25	4 rules
T + L	2	Decision tree	J48	69.84	Size 9 Leaves 5
T + L	1	Decision tree	J48	69.3	Size 19 Leaves 10

Representative examples of rules and decision trees which emerged (Figure 5) are shown below.

Example of rules:

- number_of_steps > 325 AND number_connections_added_deleted ≤ 0.638554: Average (4.0)
- number_of_steps ≤ 325 AND number_of_steps > 139: Average (32.0/9.0)
- number_of_steps ≤ 237: Bad (8.0/2.0)
- number_linking_phrases_added_deleted ≤ 0.44: Bad (15.0/1.0)

Example of a decision tree:

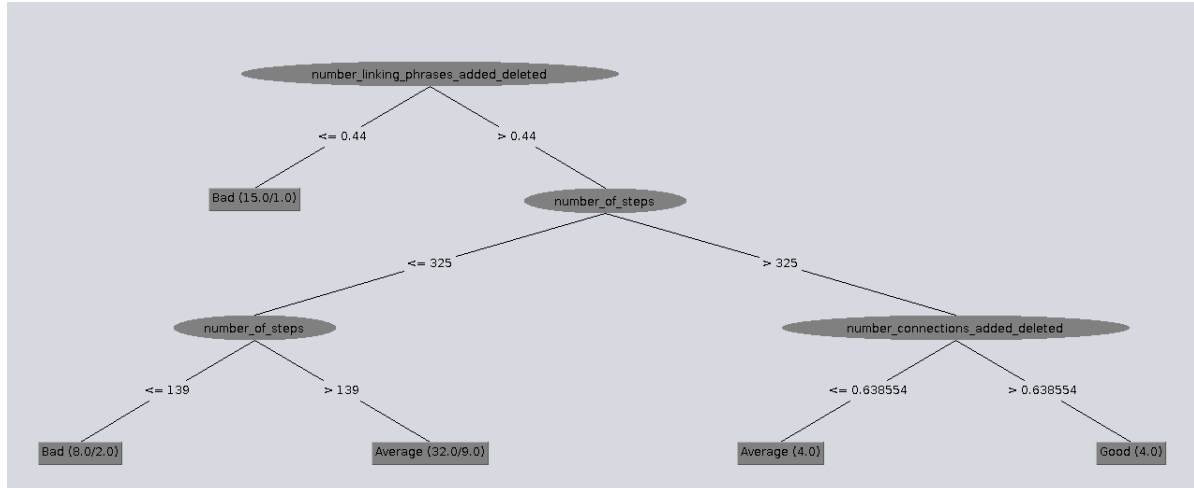


Figure 5: Decision tree generated after analysis of concept maps (T+L)

4 DISCUSSION AND CONCLUSION

We conducted this research with an aim to understand how the semi-automatic evaluation of concept maps and learner action logs can be analysed from the perspective of evidence-based practice. We implemented a study with participants who generated concept maps after experiencing video-based learning intervention. The analysis of concept map was done both manually and through the implementation of the classification algorithms. Two categories of classification algorithms used were - Decision Tree and Rule-Based Classifiers. Deletions of Linking Phrases seem to be a significant predictor of concept map quality. If the percentage of the ratio of linking phrases remaining to the ratio of total linking phrases is less than 44% (deletions were more than 56% of the total linking phrases), then the concept map was considered a bad concept map. Quality of the concept map also depends on the number of steps. Analysis showed that if the number of steps falls below a threshold (in this case 139), the concept map can be classified as a bad concept map. However, since the two concept maps on an average require different steps, we cannot generalize this to both the concept maps. Operations on Linking Phrases and Connections seem to be a significant predictor of concept map quality rather than addition/deletion of concepts.

The evidence-based practice advocates for the sustained interactions between the stakeholders. In a teaching-learning scenario, both the stakeholders - teachers and researchers should be brought together with an aim to improve the learning experience of the students. In the case discussed in this paper, practitioner interacted with researchers and informed them about their difficulties in manually evaluating concept maps. Researchers, thereafter applied learning analytics to create models of 'good', 'average' and 'bad' concept maps. These models helped in providing

evidence about students' learning, as the practitioner used them to identify students who poorly performed in the concept map based assessment. More specific results from researchers' analysis, for example operations on Linking Phrases and Connections were significantly better predictor of concept map quality as compared to addition and deletion operations, can act as evidence that practitioner can use in making informed decisions when they observe their students struggling with concept map creation activities.

This case can be seen as an instance of the implementation of Learning Evidence Analytics Framework (LEAF) which is based on evidence-based practice (Figure 6). In this case, generation of the concept maps and their log data corresponds to the data plane. Learning analytics plane includes evaluation of concept maps by the use of algorithms. This plane also includes the emerged rules and decision trees which can be used to evaluate the remaining set of concept maps. Finally, the evidence analytics plane corresponds to the informed decisions which can be taken by the practitioners based on results of learning analytics plane.

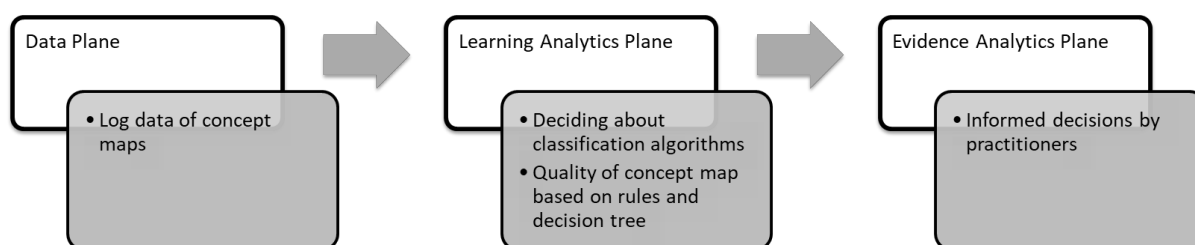


Figure 6: Alignment between LEAF framework and our steps

Our work has a few limitations. Actions for a 'good' concept map could not be analysed effectively. Most of the rules were for 'average' and 'bad' concept maps. This was due to the relatively smaller percentage of 'good' concept maps in the data set. Another limitation is that we applied this evaluation method only to two topics. As part of our future work, we would like to address these limitations along with the issue of the small sample size of participants and sustained interaction with practitioners.

REFERENCES

- Aghakhani, N., Sharifnia, H., Eghtedar, S., & Torabizadeh, C. (2015). The Effect of Concept Mapping on the Learning Levels of Students in Taking the Course of Nursing Care of Patients with Glandular Diseases Subject in Urmia University of Medical Sciences, Iran. *Jundishapur Journal of Chronic Disease Care*, 4(2).
- Bruniges, M. (2005). An evidence-based approach to teaching and learning.
- Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., ... & Carvajal, R. (2004). CmapTools: A knowledge modeling and sharing environment.
- Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2011, June). Kit-build concept map for automatic diagnosis. In *International conference on artificial intelligence in education* (pp. 466-468). Springer, Berlin, Heidelberg.
- Hsieh, I. L. G., & O'Neil Jr, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior*, 18(6), 699-715.

- Kvernbekk, T. (2017) Evidence-Based Educational Practice, in Oxford research encyclopedias, DOI: 10.1093/acrefore/9780190264093.013.187
- Luke, A. (1999, October). Education 2010 and new times: Why equity and social justice still matter, but differently. In Education Queensland online conference. Retrieved on (Vol. 1, No. 07, p. 2010).
- Novak, J. D., & Cañas, A. J. (2008). The theory underlying concept maps and how to construct and use them.
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, B. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning. In 26th International Conference on Computers in Education (ICCE 2018), at Manila.
- Pailai, J., Wunnasri, W., Yoshida, K., Hayashi, Y., & Hirashima, T. (2017). The practical use of Kit-Build concept map on formative assessment. *Research and Practice in Technology Enhanced Learning*, 12(1), 20.
- Protheroe, N. (2001). Improving Teaching and Learning with Data-based Decisions: Asking the Right Questions and Acting on the Answers. *ERS Spectrum*, 19(3), 4-9.
- Trumpower, D. L., & Sarwar, G. S. (2010). Formative structural assessment: Using concept maps as assessment for learning. In Fourth International Conference on Concept Mapping, Vina del Mar, Chile. Retrieved January (Vol. 22, p. 2013).
- Wang, S., Walker, E., & Wylie, R. (2017, June). What Matters in Concept Mapping? Maps Learners Create or How They Create Them. In International Conference on Artificial Intelligence in Education (pp. 406-417). Springer, Cham.
- Wu, S. Y., Chen, S. Y., & Hou, H. T. (2016). Exploring the interactive patterns of concept map-based online discussion: a sequential analysis of users' operations, cognitive processing, and knowledge construction. *Interactive Learning Environments*, 24(8), pp. 1778-1794.

Modelling students' effort using behavioral data

Barbara Moissa, Geoffray Bonnin, Sylvain Castagnos and Anne Boyer

Université de Lorraine – LORIA, Nancy, France

{barbara.moissa, geoffray.bonnin, sylvain.castagnos, anne.boyer}@loria.fr

ABSTRACT: Students' effort is often considered a key factor for students' success. It has several related definitions, none of which is widely adopted. In this paper, we define students' effort as the experienced cognitive load, which is the total amount of cognitive resources used during the execution of a given task. We propose an effort model to quantify students' effort based on this construct. Our approach uses behavioral measures (i.e., interaction and eye gaze data). Our preliminary results show that the eye gaze measures have an intermediary relationship with effort, while the interaction measures have a weak relationship with effort and seem slightly complementary to eye gaze measures.

Keywords: students' effort, cognitive load, descriptive analytics, eye gaze data, interaction data

1 INTRODUCTION

Decades of studies have shown that student's success is strongly dependent on their effort (Hill, 1990; Swinton, 2010; Scariot et al., 2016). Being able to accurately measure students' effort can therefore lead to a better understanding of its relationship with learning outcomes, and to the design of new tools to help teachers identify students who are struggling or not truly engaged in their learning. However, measuring the effort is a particularly difficult task. One difficulty is that despite all the interest raised by this concept, student effort has no widely adopted definition (Meltzer et al., 2001). For instance, some define it as just "the amount of studying" (Schuman, 2001), while others define it in a more specific manner, e.g., "the amount of time and energy that students expend in meeting the formal academic requirements established by their teacher and/or school" (Carbonaro, 2005).

Not surprisingly, researchers have used a variety of approaches to measure students' effort. These approaches include the time spent on learning tasks (Schuman et al., 1985; Hill, 1990) and grades assigned by teachers (Nagy, 2016; Swinton, 2010). Despite being easy to acquire, the time spent on learning tasks can be considered as unreliable, as a student may spend a long time on some activity precisely because he is not making much effort. Grades given by teachers (and even by students themselves) are time consuming, and cannot be automated.

Other approaches are related to the exploitation of behavioral data. These include the work of (Scariot et al., 2016), who proposed the modeling of students' effort using Moodle's log data as a behavioral measure. Their assumption was that greater students' participation on Moodle means greater students' effort. This assumption makes sense, since having a good attendance, delivering learning tasks on time (or not), and other actions (Carbonaro, 2005), are good attitudes expected from students. Another related study is the one from Huptych et al. (2017) which uses the total number of clicks given by a student in each activity to measure the effort. One downside of this approach is that

they are only able to measure students' effort on activities that require students' to interact by clicking.

In this paper, our aim is to continue this last line of research and to propose a model that exploits behavioral data to quantify students' effort. We rely on the Cognitive Load Theory and on its methods for measuring the cognitive load. More precisely, we study the ability of different behavioral measurements to measure the effort, and propose different combination approaches to enhance the accuracy of the measurements, while being applicable to different types of learning activities.

The reminder of this paper is as follows: Section 2 describe the Cognitive Load Theory and some approaches for measuring cognitive load. Section 3 describes the dataset we collected and Section 4 shows our results using this dataset. Finally, Section 5 closes the paper.

2 THE COGNITIVE LOAD THEORY

The Cognitive Load Theory (CLT) was proposed by Sweller (1988). According to this theory, learning is the development and automation of schemas in the working memory and the storage of those schemas in the long-term memory for easy access and use (Paas et al., 2003; Leppink, 2017). The theory states that the learning design must take into account the limitations of the working memory in order to avoid underload and overload, allowing the storage of the learning schemas (Leppink, 2017). In case of underload, i.e., if the student does not exert the proper amount of effort, no new information will be stored in the long-term memory. On the other hand, in case of overload, i.e., if the student exerts a great amount of effort, there will be no cognitive resources left to store new information in the long-term memory. Overload typically happens when a task is too hard for a given student. For instance, the task may require too much knowledge in order to be completed. Processes that do not allow the creation of learning schemas may also consume the available cognitive resources, e.g., when a student divides his attention between different information sources in different spaces or times.

Many researchers consider the cognitive load as a form of student effort (Paas & Van Merriënboer, 1993; Paas et al., 2003; Leppink, 2017). Moreover, the CLT gained lots of attention since it was first proposed, becoming a consolidated research field. One goal of this field is to accurately measure the cognitive load. Although it cannot be measured directly, it can be inferred through other measures that are believed to have a high correlation with it (Xie & Salvendy, 2000). Those measures can be classified in four categories: subjective, performance, physiological and behavioral measures (Chen et al., 2016).

Subjective measures are a popular way of measuring the cognitive load (Paas et al., 2003; Shi et al., 2007). This approach is based on the assumption that people are capable to introspect their cognitive processes and report the mental effort exerted (Leppink, 2017). It consists in asking the participants to self-assess their cognitive load in the middle of a task (Shi et al., 2007) or immediately after the task is over (Chen et al., 2016), being unsuitable for applications that require real-time data. This approach has been shown to be sensitive to small differences, valid, reliable and unobtrusive (Paas et al., 1994, Paas et al., 2003), and is often used as a baseline to measure the correlation of other measures with the cognitive load (Chen et al., 2016).

Performance measures are based on the assumption that the experienced cognitive load will reflect on task outcomes, and correspond to grades, the number of correct exercises, etc. Although in the educational context it is often assumed that the more effort students exert on their learning task, the higher their outcomes will be, it is possible for two students to achieve the same outcome for a given task and exert different levels of effort on it (Paas et al., 2003). This can be explained by the potential differences in previous knowledge of the students (Hau & Salili, 1996). Thus, Paas & Van Merriënboer (1993) argue that performance measures should be combined with other types of cognitive load measures to assess the efficiency of the instructional design.

Physiological measures are based on the assumption that the increase of the experienced cognitive load leads to physiological changes (Paas et al., 2003). These changes affect various body properties, such as temperature, heartbeat, pupil dilation, brain waves, etc. (Kramer, 1990). One advantage of the corresponding measures is that they can be captured at a high rate and with a high degree of sensitivity (Paas & Van Merriënboer, 1994), and can therefore capture variations of cognitive load over time. However, they typically require the use of specific technologies that can bias the learning experience.

Behavioral measures capture objectively and implicitly the subjects' actions (Chen et al., 2016). Examples of behavioral measures include eye activity such as blink frequency, fixation frequency and fixation duration (Beatty & Lucero-Wagoner, 2000), speech features (Khawaja et al., 2007; Yin et al., 2007), linguistic features (Khawaja et al., 2014), mouse usage (Arshad et al., 2013), digital pen input (Ruiz et al., 2007; Yu et al., 2011), gait patterns (Verrel et al., 2009), and head movements and mouth openness.

Usually, researchers only use these measurements to detect increases and decreases of the cognitive load, or to classify the cognitive load into categories such as low, medium and high. Another key point when measuring the cognitive load is that using several measures instead of just one in isolation usually increases the accuracy of the measurements by reducing noise and eventually overcoming the lack of data (Mulder, 1992; Chen et al., 2016).

3 DATASET

Our goal in this paper is to propose an effort model based on behavioral data. In order to acquire such data, we wanted to rely on a course material that would contain various contents (text, images, graphics, etc.) and a clearly defined educational objective. We chose the context of language learning, because this field allows to acquire knowledge and to restore it in a relatively short time, since the encoding of all aspects of the language and its restitution are mainly based on verbal short-term memory and verbal working memory (Baddeley, 2003).

In order to avoid a bias related to a mastery of the language prior to our study, we chose the Esperanto language, which is in little use and not studied at school. After a comparative study of the

different Esperanto online learning sites, we selected the iKurso website¹ for its simple and complete course, accompanied by various exercises.

We recruited participants through a university mailing list. 14 French volunteers passed the test, 8 students, 1 engineer, 2 researchers, 2 PhD students, and 1 post-doctoral student. Our panel is composed of 6 women and 8 men. 7 are between 18 and 25 years old, 3 between 26 and 30 years old, 2 between 31 and 35 years old, and 2 between 36 and 40 years old. None of them knew Esperanto prior to our study (as answered in a questionnaire). In order to maximize the engagement of the participants in the tasks, we decided to set up a lottery whose outcome was dependent on the scores obtained on the final evaluation.

When each subject arrived, we briefly presented the material (an eye-tracker and a computer) and calibrated the eye tracker. The subject was then faced with an instruction page after which the site opened, and the learning phase began. Each subject had the opportunity to browse the different pages of the course with no time limitations. Once the participant was finished with the learning phase, he was directed to an evaluation questionnaire with 21 questions (11 sentences to translate and 10 multiple-choice questions), and could not return to the course. The experiment ended when the subject submitted his answers. Table 1 provides a summary of the scores obtained by the participants.

Table 1: Summary of users' scores on the evaluation questionnaire

	<i>Grammar score (/11)</i>	<i>Vocabulary score (/11)</i>	<i>Translation score (/11)</i>	<i>MCQ (/10)</i>	<i>Global score</i>
Mean	4.28	2.92	1.92	8.78	10.23
Median	3.50	2.50	1.00	9.00	10.00
Standard deviation	3.66	2.33	2.58	1.42	3.53

During the learning phase, users' gaze data were retrieved using a Tobii X1 Light eye-tracker and the software Tobii Studio. We manually defined 336 areas of interest (AOIs) for all course elements required in the evaluation questionnaire. In total, for each subject, we extracted 18 characteristics (Marchal et al., 2016; Marchal et al., 2018): number of fixation points, cumulative duration, mean and standard deviation of fixations, cumulative length, mean and standard deviation of saccades, sum, mean, and standard deviation of the absolute and relative angles of the visual path; length of the first fixation relatively to the edge of the screen, number of dynamic AOIs obtained with the DBSCAN clustering algorithm, and entropy measures.

4 MODELLING STUDENTS' EFFORT

As mentioned in the previous section, none of the subjects had prior knowledge about Esperanto. We therefore assume their effort is the main factor influencing their outcome, and consider their score at the final questionnaire a reliable measure of effort in this case. In order to study how students' effort can be effectively measured and modeled using behavioral data, we now analyze the correlations of

¹ <https://ikurso.esperanto-france.org>

the different measurements available in the dataset with these scores, and rely on the Spearman's rank correlation coefficient.

4.1 Measuring the effort using raw indicators

We first focus on the correlations of the raw indicators extracted from the dataset. Specifically, we provide the values of the following indicators: time spent on task (TaskTime), total number of page views (#Hits), total number of clicks (#Clicks), total number of keystrokes (#Keystrokes), total time of fixations (FixationsTime), and total number of fixations (#Fixations). The correlations between these indicators and the global scores obtained by the participants are shown in Table 2. As can be seen, only small correlations were obtained. The strongest correlation we found is the one with the total number of page views (#Hits) with a coefficient of 0.28. Surprisingly, the weakest correlations are the eye gaze indicators (FixationsTime and #Fixations) with correlations coefficients near to zero.

Table 2: Raw indicators correlation with scores

Type	Indicator	Correlation Level
Interaction	TaskTime	0.21 Low
	#Hits	0.28 Low *
	#Clicks	0.14 Low
	#Keystrokes	0.10 Low
Eye gaze	FixationsTime	-0.08 Low *
	#Fixations	0.00 Low *

4.2 Combining several interaction indicators

As mentioned previously, one means of increasing the accuracy of the measurements is to combine different measures. We thus created new indicators using the three following types of combinations:

1. *Average time between actions*: These combinations correspond to the total duration of the session divided by the number of occurrences of a type of action. These actions are pages views (#Hits), clicks (#Clicks) and keystrokes (#Keystrokes). We also combined all of them by adding them together (#Actions):

$$AvgTimeBetweenHits = TaskTime \div \#Hits \quad (1)$$

$$AvgTimeBetweenClicks = TaskTime \div \#Clicks \quad (2)$$

$$AvgTimeBetweenKeystrokes = TaskTime \div \#Keystrokes \quad (3)$$

$$AvgTimeBetweenActions = TaskTime \div \#Actions \quad (4)$$

2. *Weighted actions*: This indicator combines three types of actions (page views, clicks and keystrokes) in a different way. The number of clicks (#Clicks) and keystrokes (#Keystrokes) are used as a means to weight the importance of the page views (#Hits)²:

$$\text{WeightedActions} = \text{\#Hits} \times (\text{\#Clicks} + \text{\#Keystrokes} + 1) \quad (5)$$

3. *Eye gaze*: These indicators correspond to the average time (duration) of the fixations (Equation 6), and the average elapsed time between fixations (Equation 7):

$$\text{AvgTimeFixations} = \text{FixationsTime} \div \text{\#Fixations} \quad (6)$$

$$\text{AvgTimeBetweenFixations} = \text{TaskTime} \div \text{\#Fixations} \quad (7)$$

The correlations between the indicators described above and the global scores of the participants are presented in Table 3. Although still generally small, the resulting correlations are higher than the correlations coefficients shown in Table 2. Perhaps the most surprising outcome is the medium correlation (0.54) of the average duration of fixations (AvgTimeFixations). Moreover, the correlation for the weighted actions (WeightedActions) and for the average time between clicks (AvgTimeBetweenClicks) is 0.35, which is an improvement compared to the previous highest interaction correlation coefficient of 0.28 for the page views (#Hits).

Table 3: Combined indicators correlation with scores

Type	Indicator	Correlation	Level	
Interaction	AvgTimeBetweenHits	0.05	Low	
	AvgTimeBetweenClicks	0.35	Low	*
	AvgTimeBetweenKeystrokes	0	Low	
	AvgTimeBetweenActions	-0.18	Low	
	WeightedActions	0.35	Low	*
Eye gaze	AvgTimeBetweenFixations	0.36	Low	
	AvgTimeFixations	0.54	Medium	*

4.3 Towards an effort model

We finally proceeded to create the effort model by choosing one interaction and one eye gaze indicator to combine, expecting them to be complementary and therefore able to increase the correlation with the scores when combined. Beside the average time between fixations, two previous interaction indicators had the same highest correlation values (highlighted in Table 3). Of course, the

² This equation contains an addition of one to avoid having a value of zero when the task does not require any clicks or keystrokes to be completed (e.g., reading a text).

chosen eye gaze indicator is the average duration of fixations (AvgTimeFixations) as it is the highest correlation (also highlighted in Table 3).

Both types of indicators have different ranges and distributions. Thus, before combining these indicators, we normalized them by using an exponential function with an upper limit of one (Equation 8). We then combined the interaction indicators (AvgTimeBetweenClicks and WeightedActions) and the eye gaze indicator (AvgTimeFixations) into the model as shown in Equation 9³.

$$\text{norm}(x) = 1 - e^{-\alpha x} \quad (8)$$

$$\text{Effort} = w_1 \times (\text{norm}(\text{WeightedActions})) + w_2 \times (\text{norm}(\text{AvgTimeFixations})) \quad (9)$$

The correlations results for these combinations are shown in Table 4. As can be seen, the highest correlation is 0.59 for the combination of WeightedActions ($w_1 = 0.7$) with AvgTimeFixations ($w_2 = 0.3$). Overall, we were thus able to increase the correlation by more than 100% compared to the initial correlations.

The other combination (AvgTimeBetweenClicks and AvgTimeFixations) did not improve the correlation (i.e., the highest correlation still equal to the correlation of the eye gaze indicator alone), which suggests that the average time between clicks, which is not as flexible to the type of task as the WeightedActions indicator, is not complementary to the eye gaze indicator. This is not surprising, as people usually watch the screen when using the mouse.

Table 4. Comparison of the combined indicators

AvgTimeBetweenClicks and AvgTimeFixations	0.54
WeightedActions and AvgTimeFixations	0.59

5 CONCLUSION

A number of definitions have been proposed for the concept of students' effort. Usually, these definitions are related to various students' behaviors such as attending classes, delivering assignments on time, participating in class, etc. The Cognitive Load Theory allows the understanding of effort as the cognitive load experienced by students. The theory also explains that learning occurs when we create and automate schemas in the working memory and then store them in the long-term memory, stating that the limitations of the working memory should be respected in order to allow learning.

This theory gained lots of attention and several researchers are now looking for new and more effective ways of measuring the imposed cognitive load (i.e., effort) in order to detect and avoid underload and overload. Different means of acquiring measurements of the cognitive load exist, and

³ We empirically determined the value of the parameters that led to the higher correlation coefficients. We did however not systematically optimize these values.

they can be classified into subjective, performance, physiological and behavioral measures. However, those measurements only allow to classify or identify an increase or decrease in the cognitive load.

In this paper, we proposed a new model that allows to quantify students' effort, as opposed to identified or classified as shown in the cognitive load research. Despite the limitations of our dataset (i.e., absence of effort ratings and number of subjects) and the need of further investigation, our study adopts some new approaches that can offer a better understanding of students' effort. First, it relies on the Cognitive Load Theory that explains how learning occurs, the limitations that should be considered while designing learning activities (i.e., the working memory has a limit that should be respected), and also the role of effort in achieving learning outcomes. Second, it uses interaction and eye gaze data which can reduce noise and account for missing data (Mulder, 1992; Chen et al., 2016). Third, the model can still be used if only part of the data is available, even though it would result in a smaller correlation (e.g., if only the interaction data are available, the correlation coefficient would be 0.35 instead of 0.59).

Our results show that our approach is able to provide measurements that have a medium correlation with effort. Especially, our proposed combination increased the correlation score by more than 100% compared to raw indicators. This approach can be exploited to develop effort-based educational tools and help both teacher and students. For instance, with proper dashboards, teachers could be able to identify students who are not well engaged into learning (i.e., not exerting enough effort on their learning tasks) and students who are struggling (i.e., exerting too much effort on their learning tasks, possibly meaning that they lack previous knowledge to handle the proposed tasks). Similarly, students could identify if they are just not engaged enough or if they are struggling with the proposed learning tasks, leading them to seek help from their teachers and classmates. The proposed model can also be used in fully automated tools. For instance, recommendation systems could identify how much effort a student should exert in a given moment and recommend the ideal learning tasks, with the goal of promoting his engagement.

In the future, we intend to create a new dataset and run the model to see how it performs and how it can be further enhanced. Later, we want to add new behavioral (e.g., movements) and physiological measurements (e.g., pupil dilation, and skin temperature). Our ultimate goal is to be able to exploit our effort model to provide engaging recommendations to the students.

REFERENCES

- Arshad, S., Wang, Y., & Chen, F. (2013). Analysing mouse activity for cognitive load detection. In *Proc. of the 25th Australian Computer-Human Interaction Conf.: Augmentation, Application, Innovation, Collaboration* (pp. 115–118). ACM.
- Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience* 4, 829–839.
- Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of Psychophysiology* (pp. 142–162). Hillsdale, NJ: G. Berntson & L. G. Tassinar.
- Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, 78(1), 27–49.

- Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-31700-7>
- Hau, K.-T., & Salili, F. (1996). Prediction of academic performance among Chinese students: Effort can compensate for lack of ability. *Organizational Behavior and Human Decision Processes*, 65(2), 83–94.
- Hill, L. (1990). Effort and reward in college: A replication of some puzzling findings. *Journal of Social Behavior & Personality*, 5(4), 151–161.
- Huptych, M., Bohuslavek, M., Hlosta, M., & Zdrahal, Z. (2017). Measures for recommendations based on past students' activity (pp. 404–408). ACM Press. <https://doi.org/10.1145/3027385.3027426>
- Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *Int. Journal of Human-Computer Interaction*, 30(5), 343–368.
- Khawaja, M. A., Ruiz, N., & Chen, F. (2007). Potential speech features for cognitive load measurement. In *Proc. of the 19th Australasian Conf. on Computer-Human Interaction: Entertaining User Interfaces* (pp. 57–60). ACM.
- Kramer, A. F. (1990). *Physiological Metrics of Mental Workload: A Review of Recent Progress*.
- Leppink, J. (2017). Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences*, 12(5), 385–391.
- Marchal, F., Castagnos, S., & Boyer, A. (2016). A First Step toward Recommendations Based on the Memory of Users. In *Proc. of the 28th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI)*. IEEE.
- Marchal, F., Castagnos, S., & Boyer, A. (2018). First Attempt to Predict User Memory from Gaze Data. *Int. Journal on Artificial Intelligence Tools*.
- Meltzer, L., Katzir-Cohen, T., Miller, L., & Roditi, B. (2001). The Impact of Effort and Strategy Use on Academic Performance: Student and Teacher Perceptions. *Learning Disability Quarterly*, 24(2), 85–98.
- Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, 34(2–3), 205–236.
- Nagy, R. (2016). Tracking and Visualizing Student Effort: Evolution of a Practical Analytics Tool for Staff and Student Engagement. *Journal of Learning Analytics*, 3(2), 165–193.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38(1), 63–71.
- Paas, F., Van Merriënboer, J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1), 419–430.
- Paas, F., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35(4), 737–743.
- Paas, F., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351–371.
- Ruiz, N., Taib, R., Shi, Y. D., Choi, E., & Chen, F. (2007). Using pen input features as indices of cognitive load. In *Proc. of the 9th Int. Conf. on Multimodal Interfaces* (pp. 315–318). ACM.
- Scariot, A. P., Andrade, F. G., Silva, J. M. C. da, & Imran, H. (2016). Students Effort vs. Outcome: Analysis Through Moodle Logs. In *IEEE 16th Int. Conf. on Advanced Learning Technologies* (pp. 371–372). Austin, TX, USA: IEEE.

- Schuman, H. (2001). Comment: Students' Effort and Reward in College Settings. *Sociology of Education*, 74(1), 73–74.
- Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces*, 63(4), 945–966.
- Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). ACM.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Swinton, O. H. (2010). The effect of effort grading on learning. *Economics of Education Review*, 29(6), 1176–1182. <https://doi.org/10.1016/j.econedurev.2010.06.014>
- Verrel, J., Lövdén, M., Schellenbach, M., Schaefer, S., & Lindenberger, U. (2009). Interacting effects of cognitive load and adult age on the regularity of whole-body motion during treadmill walking. *Psychology and Aging*, 24(1), 75–81.
- Xie, B., & Salvendy, G. (2000). Prediction of Mental Workload in Single and Multiple Tasks Environments. *Int. Journal of Cognitive Ergonomics*, 4(3), 213–242.
- Yin, B., Ruiz, N., Chen, F., & Khawaja, M. A. (2007). Automatic cognitive load detection from speech features. In *Proc. of the 19th Australasian Conf. on Computer-Human Interaction: Entertaining User Interfaces* (pp. 249–255). ACM.
- Yu, K., Epps, J., & Chen, F. (2011). Cognitive load evaluation of handwriting using stroke-level features. In *Proc. of the 16th Int. Conf. on Intelligent User Interfaces* (pp. 423–426). ACM.

Using Log Data to Evaluate MOOC Engagement and Inform Instructional Design

Yuqian Chai¹, Chi-Un Lei², Yu-Kwong Kwok¹

¹Department of Electrical and Electronic Engineering, The University of Hong Kong

²Technology-Enriched Learning Initiative, The University of Hong Kong
{u3004409, culei, Ricky.Kwok}@hku.hk

ABSTRACT: Traditional educational studies verify the performance of courses through questionnaires, interviews and observations, which can be an arduous task for researchers. It is easier to verify the effectiveness of online courses as all the interactions between students and the courseware are recorded. However, the utilization of these activity data is lack of theoretical framework. In this paper, we propose to utilize learning interaction theory and web analytics knowledge to evaluate MOOC engagement and inform instructional design. This framework is composed of learner-interface, learner-content and learner-community interaction. 15 indicators derived from web analytics are proposed to help teachers better understand the engagement level of their courses in three interaction dimensions.

To illustrate how the above analysis can facilitate teaching in practice, we used log data of 10 MOOCs owned by The University of Hong Kong on edX. Results and corresponding insights are offered. 10 experts are invited to evaluate the proposed framework. Most of them have showed positive attitudes. In the future, we will cooperate with MOOC designers and verify whether this framework can help them teaching and improve MOOC engagement.

Keywords: MOOCs, Log Data, Learning Analytics, Learning Design

1 INTRODUCTION

The flexibility of the learning conditions and the ability to record and process a large amount of data are two major advantages of online learning over traditional learning. Traditional educational studies analyze the learner behavior through expert observations, questionnaires, interviews and other methods (Lodico & Voegtle, 2010). These methods require substantial efforts and generate overestimated results since less motivated students are less likely to participate in the surveys or interviews. Compared with these methods, it is easier for online behavior analysis to generate unbiased results as all students' online behaviors will be analyzed. Since all the data is automatically stored in LMS, data collection is relatively effortless.

Online behavior analysis can facilitate MOOC development in two ways: (1) building a real-time learning dashboard to help teachers monitor the learning status of the whole class (Schwendimann et al., 2017); (2) generating the course's report every year or semester which can help teachers review the performance of their courses. Molenaar et al., (2018) proved that dashboards can influence teaching progressively. Ogata et al., (2018) have proposed the Learning Evidence Analytics Framework (LEAF) for evidence-based education. However, there were little details on how to extract evidences from log data. In addition, the indicators adopted by existing dashboards remained also lack of theoretical framework. Therefore, it is necessary to adopt a theoretical framework with higher level indicators to verify the engagement of online courses.

There are many engagement studies examining the interaction in distance education. Interaction has been considered as one of the most important components for effective traditional learning and

online learning. Moore (1989) proposed that learning process should be classified into three types of interactions (learner-content, learner-learner and learner-instructor interaction). A new type of interaction, learner-interface interaction, has been identified in distant education (Hillman et al., 1994). This theoretical framework can be utilized to evaluate the engagement performance of online courses and offer actionable insights to teachers.

We utilize the knowledge of web analytics and learning interaction theory and to propose a MOOC engagement evaluation framework. In detail, the major contributions of this paper can be summarized as follows:

- We re-define four web analytics terminologies in the learning context which help develop indicators for online courses evaluation framework and teacher-facing dashboards;
- We utilize the learner interaction theory to propose a MOOC engagement evaluation framework to verify the engagement of MOOCs in learner-interface interaction, learner-content interaction and learner-community interaction;
- We use 10 MOOCs' log data to demonstrate how to conduct the behavior analysis with the MOOC engagement evaluation framework and identify the strengths and problems in these courses and their course components.

The rest of the paper is organized as follows. In Section 2, we give a short introduction of the dataset involved in this study. In Section 3, we re-define the web analytics terminologies in the online learning context to facilitate evaluating online courses' performance. We utilize the learning interaction theory to propose the course evaluation framework in Section 4. The implementation of the above framework is described in Section 5 and conclusions are given in Section 6.

2 DATASET DESCRIPTION

Ten MOOCs owned by The University of Hong Kong on edX are analyzed in this study. Overall, our dataset contains around 20 million logs generated by 30450 students. The rich diversity of these ten MOOCs can be reflected in academic category, instructional design, duration, availability of historical data, etc. Details of the ten MOOC information can be checked in Table 1.

Table 1: Details of the 10 MOOCs.

cid	course id	#students	Days
0	HKU01x/14	5521	73
1	HKU01x/15	2936	80
2	HKU02.1x	2564	37
3	HKU02.2x	1260	41
4	HKU03x/15	3803	70
5	HKU03x/16	2213	63
6	HKU04x/15	3571	49
7	HKU04x/16	2000	38
8	HKU05.1x	4927	42
9	HKU06.1x	1497	42

3 WEB ANALYTICS IN ONLINE LEARNING

Traditionally, researchers verified the effectiveness of courses via students' academic performance, questionnaires and interviews. However, it is not applicable for MOOCs as not many MOOCs contain compulsory and rigid assessments. Therefore, we need to verify the engagement of MOOCs without assessment data. To solve this problem, web analytics knowledge is involved as online courses share the same major data source (log data) with web analytics.

Web analytics is defined as the assessment of a variety of data to help create generalized understanding of the visitor experiences (Peterson, 2004). With web analytics, people can verify the engagement of their online marketing campaigns for continual improvement. If we consider each course as a website and each learner as the website visitor, we can easily measure the engagement with web analytics indicators. In the following part, we will first give the online learning context definitions of four important web analytics concepts:

- **Session/Visit:** When a learner triggers the next event within 30 minutes, these events will be considered as within one session/visit. The intention is to avoid the cases where students may have left the system but have not logged out. This can be shown as long staying time without triggering any events. Average number of sessions can reflect the course's ability in keeping learnings frequently checking the courseware. In addition, average duration of sessions can be considered as the rough learning time of students. Courses with longer average session can be considered as more attractive in gaining learners' attention.
- **Page View:** In web analytics, a page view is considered as one load of a web page. The interpretations of a page are different in different platforms because of the different course structures. Since we extracted the data from edX, one page is identified as one subsection in this study and correspondingly, one page view is one view of the loaded subsection.
- **Interested/Time-engaging Learner:** A learner who spends more than n minutes will be considered as an interested/time-engaging learner. The Interested learner proportion can be an important indicator to measure the possibility of the courses in attracting learners' attention to stay longer. Unlike web analytics, we will use the median time student spent across courses instead of the average time student spent as the value of n .
- **Heavy/Action-engaging Learner:** The learner with more than n events will be regarded as heavy or action-engaging. The heavy/action-engaging learner proportion is the percentage of learners who trigger more than n events. It can be used to measure the ability of courses in driving learners to actions. The median number of events students triggered will be the threshold to distinguish whether learners are action-engaging.

4 MOOC EVALUATION FRAMEWORK

Interaction plays a very important role in traditional classrooms and distant education. Many researchers have dedicated on interaction studies in learning context. One of the most classical studies is proposed by Moore (1989). He classified learning interactions into three types: learner-content, learner-instructor and learner-learner interaction. Furthermore, learner-interface interaction has been identified as the fourth type of interaction in distant education (Hillman et al., 1994). These categorizations can help us verify the engagement of online courses.

First, we need to map the learners' online activities with these four types of interactions. Students' interactions with the courseware belong to the learner-interface interaction while students' activities with htmls, videos and problems belong to the learner-content interaction. As all the learner-

instructor and learner-learner interactions happen in the forum, we merge them into learner-community interaction. Therefore, our evaluation framework measures the learning engagement of MOOC in learner-interface, learner-content and learner-community interaction. Each type of interaction will be measured based on several indicators. The indicator summary of each interaction type is given in Table 2.

Table 2: Indicators of learner-interface, learner-content and learner-community interaction.

Learner-Interface Interaction	Learner-Content Interaction		Learner-Community Interaction
<ul style="list-style-type: none"> • weekly time spent • triggered events • active learner proportion • time-engaging learner proportion • action-engaging learner proportion 	html	<ul style="list-style-type: none"> • active reader proportion • average page views 	<ul style="list-style-type: none"> • forum active learner proportion • number of learner posts • number of instructor posts • average replies of threads
	video	<ul style="list-style-type: none"> • active watcher proportion • average page views 	
	problem	<ul style="list-style-type: none"> • active problem-solver proportion • average triggered events 	

4.1 LEARNER-INTERFACE INTERACTION

In learner-interface interaction, general learning activities between students and the courseware are considered. This type of interaction was recognized by Hillman (1994). Students' learning experience with interface can be identified via the following five aspects:

- **weekly time spent:** It refers to the total staying time of all sessions over the past complete week. It is assumed that students will spend more time in the course if they find the course attractive. Therefore, this indicator can reflect whether a course is time level engaging;
- **total triggered events:** This indicator is the number of events learners have triggered across weeks. We will compare the average total triggered events among courses. Courses with higher total triggered events can be considered more action level engaging;
- **active learner proportion:** Active learners are the learners who access the courseware at least once. This concept has been widely used in the domain of learning analytics, such as edX Insights¹. The active learner proportion is the ratio of active learners to all registered learners. We adopted the active learner proportion to better compare among courses;
- **time-engaging learner proportion:** This is the percentage of learners who spend more than n minutes in the course. A course with higher time-engaging learner proportion can be more attractive. We use the median duration of all sessions (25 minutes) in the dataset as n .
- **action-engaging learner proportion:** A learner who triggers more than n events is an action-engaging learner. Higher action-engaging learner proportion indicates the course is well designed in driving students into actions. Here, n is median triggered events (20) in dataset.

4.2 LEARNER-CONTENT INTERACTION

¹ <https://insights.edx.org/courses/>

Activities where learners encounter with learning materials are considered in the learner-content interaction, such as watching the video, answering the MCQs, etc. This interaction is considered as one of the most key factors in learning process (Moore, 1989).

Basically, there are three common types of course materials in online learning environments which are html (static course content), video and problem. Different courses have different proportions of these materials. We will assign indicators based on the type of course materials. Details of these three types and corresponding indicators are given below.

4.2.1 Interaction with htmls

For online courses, pages with static course content (slides, reading materials, etc.) will be considered here. Interactions with such materials include viewing and closing the pages. Therefore, how many learners access the html and how many times they viewed are considered to measure the engagement level between learners and static course content.

- **active reader proportion:** Learners who have accessed the page will be considered as active of the corresponding html. The active reader proportion can give teachers an overview on their static course content' s ability in attracting learners' attention. If reading and guiding materials are attractive to learners and have high referring ability, learners will be willing to access more htmls which. This can be reflected as high average active reader proportion;
- **average page views:** Different from active reader proportion, this indicator measures the htmls' ability in keeping learners coming back. Reading materials with high average page views can be informative or interesting and motivate learners to repetitively read it.

4.2.2 Interaction with videos

Video lectures have been considered as one of the most important online learning materials (Wang & Kelly, 2017). Unlike htmls, students have more interactions with video (play, pause and jump). Due to the limitation of edX, the exact watching duration cannot be derived from log data. Thus, we measure learner-video interactions via active watcher proportion and average page views.

- **active watcher proportion:** It refers to the students who have played the video. This indicator can help us measure the referring ability of videos. If the videos have high referring ability to other videos, learners tend to access other videos after watching one and this will be reflected as high overall active watcher proportion;
- **average page views:** This indicator refers to the number of times active learners have watched the video. Reloading and refreshing video behaviors are filtered out. Videos which has high average page views can be considered as well delivered.

4.2.3 Interaction with problems

Problem is a more active learning material format compared with html and video. The participation rate is used to evaluate the interaction between students and problems. As some of the problems are in the participation level without grading, the rate of correctness is not considered here. The problem interaction is measured via the following metrics.

- **active problem-solver proportion:** Learners who manipulate with the problem will be considered as active problem-solver. Active problem-solver proportion is the participation rate of problems. If the course has high participation rate, it may indicate that this course can motivate learners to actions;

- **average triggered events:** Well-designed problems can increase students' motivation to interact with problems and reflect as high average triggered events. However, if the number of triggered events is too high, it may indicate that learners are gaming the system and such behaviors should be filtered out.

4.3 LEARNER-COMMUNITY INTERACTION

The learner-community interaction is the combination of the learner-learner interaction and learner-instructor interaction. Feeling of isolation is one of the biggest concerns with online education (Hodges & Kim, 2010). Therefore, it is very important for us to measure the engagement of learner-community interaction. In forums, learners discuss course content and build connection with fellow participants while instructors monitor the dynamics of the course and offer some help to students who encountered problem in real time (Almatrafi & Johri, 2018). We measure the learner-community interaction based on the following metrics:

- **active learner proportion:** It refers to the number of learners who are active in the forum to all learners. Students who create the thread, reply others, vote, etc. are considered as active in the forum;
- **number of learner posts:** Posts refer to thread, reply and comment. Total number of posts can reflect the active level of the forum;
- **number of instructor posts:** Staff's posts can boost learners' participation to a certain degree (Almatrafi & Johri, 2018). We use the number of instructor posts across courses as one of the indicators to measure the interactions between learners and instructors;
- **average replies of threads:** Number of replies from classmates or teachers and the average reply time of each thread can measure the level of interaction between students and the community (Idowu & McCalla, 2018). For the sake of convenience, we used the number of replies of threads to measure the engagement level of forum.

5 PRACTICE OF MOOC EVALUATION FRAMEWK

The engagement of 10 MOOCs are evaluated with the framework in Section 4. Insights are offered to demonstrate how this framework can facilitate teaching and instructional design. As the original course id is long and difficult to be shown in the figure, we use a number (cid) to represent a course.

5.1 LEARNER-INTERFACE INTERACTION COMPARISON

We utilized five indicators introduced in Section 4.1 to measure the engagement of ten MOOCs in learner-interface interaction. The results are displayed in Figure 1 and insights are as follows:

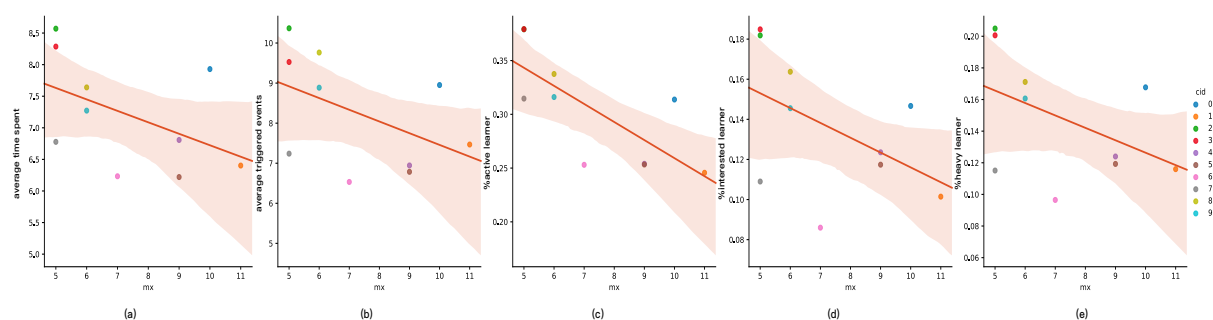


Figure 1: Comparison of the learner-Interface interaction

- Online engagement has roughly negative relationship with course duration as shown in Figure 1. It may be better to split the long duration MOOC (e.g., 10-12 weeks) into several short MOOCs (e.g., 4-5 weeks). HKU2.1x (cid: 2) and HKU2.2X (cid: 3) are two parts of the one MOOC. After splitting into two MOOCs, students' engagement with interface has been continually high during the whole process. Thus, teachers may avoid making their MOOCs too long in duration (e.g., 10-12 weeks) and try to split into several short duration MOOCs;
- It is necessary to make certain changes or improvements based on students' performance in the previous cohorts. HKU01x/14 (cid: 0) and HKU01x/15 (cid: 1) belong to the same MOOC. With no changes in the second run (HKU01x/15), all parameters except active learner proportion have decreased indicated by the independent t test results (weekly time spent: $t=8.124$, $p<0.001$; average triggered events: $t=7.861$, $p<0.001$; time-engaging learner proportion: $t=2.12$, $p<0.05$; action-engaging learner proportion: $t=2.197$, $p<0.05$);

5.2 LEARNER-CONTENT INTERACTION COMPARISON

Learners' interactions with learning content are evaluated based on the material type. As html, video and problem are three types of learning materials, we will discuss them separately. With the indicators mentioned in Section 4.2, we can evaluate the engagement of learner materials on single file level, section level and course level. For convenience's sake, learning materials are evaluated on the course level in this paper.

5.2.1 Learner-Html Interaction Comparison

We use active reader proportion and average page views to measure learner-html interaction. Results are displayed in Figure 2 and insights are as follows:

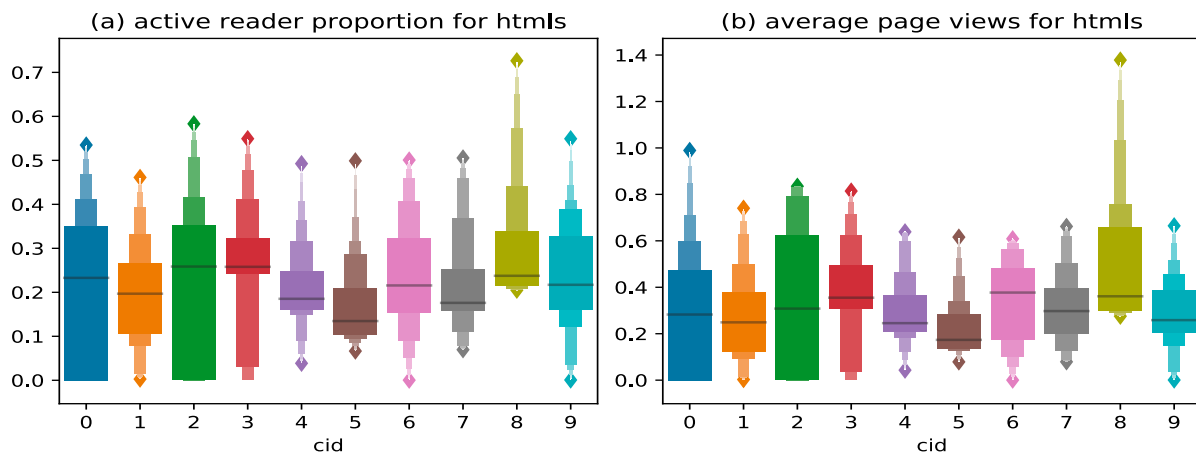


Figure 2: Comparison of the learner-html interaction

- Static course content of HKU06.1x (cid: 9) should be re-designed. From Figure 2(b), it has the lowest average page views among 10 MOOCs and most of its page views are less than 1.3. After checking this course, we found that the potential problem may be (1) the guiding materials are a short description of corresponding section content without any encouraging words and figures and (2) the recommendation list is too long (over 20 materials) while without pointing out the relationship between course content and recommendation materials. Instructors can make corresponding changes to static course content.

5.2.2 Learner-Video Interaction Comparison

Learner-video interactions are evaluated based on two indicators: active watcher proportion and average page views. Figure 3 displays the comparison results and insights are as follows:

- Different sessions of the same course tend to have the similar active watcher proportion and average page views (cid: 0-1, 2-3, 4-5, 6-7). The possible reason would be that video lectures will not be changed in the re-runs;
- HKU03x/15 (cid: 4) and HKU03x/16 (cid: 5) should improve the referring ability among video lectures. Though the active watcher proportion of these two courses are low (around 0.1), the average page views are high in Figure 3. It indicates that learners who have watched their videos like their videos and repeatedly watch them. However, most of learners did not access their videos. Instructors may improve the referring ability among video lectures so that learners can access more videos after watching one.

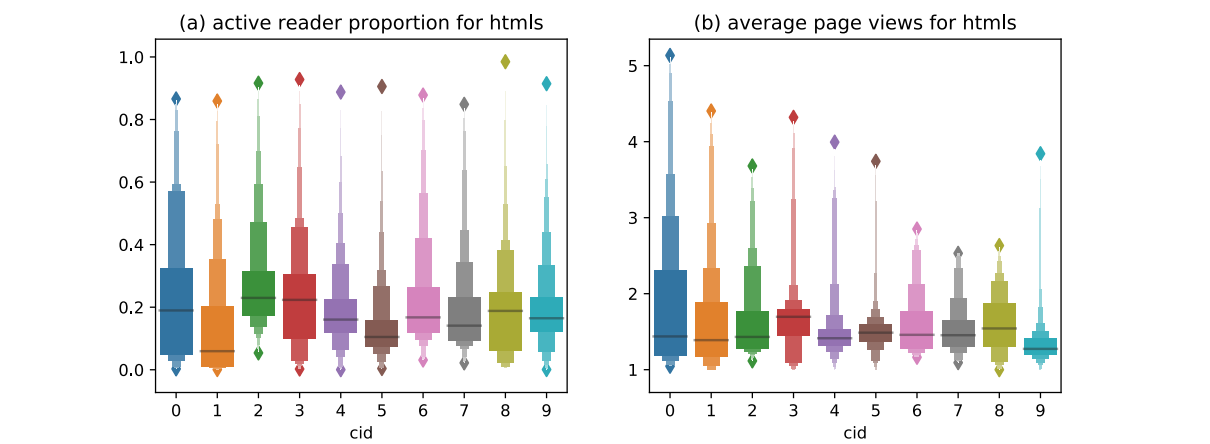


Figure 3: Comparison of the learner-video interaction

5.2.3 Learner-Problem Interaction Comparison

Indicators introduced in Section 4.2.2 are involved to evaluate the engagement level of learner-problem interaction. Results are depicted in Figure 4 and insights are as follows:

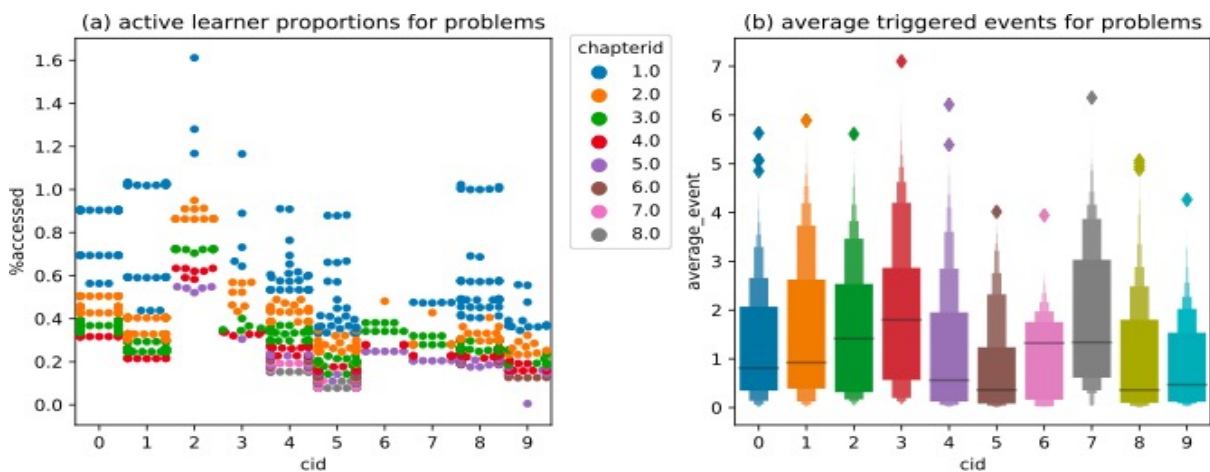


Figure 4: Comparison of the learner-problem interaction

- The active problem-solver proportion is highly correlated with the delivering time. The active problem-solver proportion gradually decreases when the chapter id² becomes larger. Teachers can utilize this to better design their problems;
- Teachers need to focus on the quality of problems. In Figure 4, one dot represents a problem. We can see that there is no direct relationship between the number of problems and active problem-solver proportion or average triggered events from Figure 4;
- Instructors of HKU04x/15 (cid: 6) and HKU04x/16 (cid: 7) can consider revising problem formats and content. In Figure 4, their active problem-solver proportion remains low. After checking the course contents, the possible reason is that all problems of this course are text input which need many steps to finish. Therefore, learners are less willing to attend even in the first week. It will be better if teachers can reduce the steps of text input questions.

5.3 LEARNER-COMMUNITY INTERACTION COMPARISON

Lastly, we compared 10 MOOCs' learner-community engagement level using the five indicators introduced in Section 4.3. Results are displayed in Figure 5 and insights are as follows:

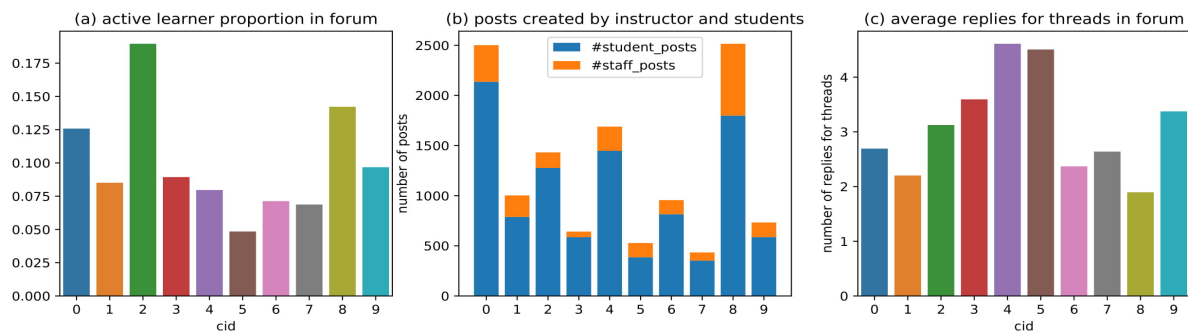


Figure 5: Comparison of the learner-community interaction

- From Figure 5, the active learner proportion and total number of posts in forum are relatively higher when there are more staff's posts;
- Teachers of HKU03/16 (cid: 5) should offer one forum entry below the course materials. In Figure 5, active learner proportion in forum and total number of posts are low compared with other MOOCs which have the similar number of staffs' posts. After observing its course forum, we found that instructors generated many posts in the forum. However, few learners have noticed those posts because it is not part of the course activities. From Figure 5(c), average replies of threads in this course are high. This indicates that learners who noticed the discussion in the forum are willing to share their thoughts in the forum.

5.4 MECHANISM EVALUATION OF THIRD-PARTY EXPERTS

To verify the effectiveness of our framework, ten third-party instructional designers are invited to comment on the framework and insights we displayed before. Nine of them showed positive attitudes on the framework and were willing to use the framework for their next MOOC while the remaining expert preferred utilizing log data by herself to observe students' behaviour. 4 experts held high expectations for clickstream data. They believed it would reduce teachers' burdens and generate more objective results. Some of them also mentioned that clickstream data can tell the teachers what

² The order of delivering time. Welcome and Farewell sections are not considered here

happens in the course but cannot identify the specific reasons behind. To find the reason, they would like to discuss with data scientists and verify the potential reasons by traditional approach (interviews, questionnaires, etc.). Therefore, we plan to cooperate with more MOOC instructors and check whether this framework will help them teaching.

6 CONCLUSION

In this study, we utilize the learning interaction theory first proposed by Moore to evaluate the engagement level of MOOCs in three dimensions (learner-interface, learner-content and learner-community interaction). 15 indicators are proposed based on web analytics and assigned to each dimension of the framework.

We demonstrate how to utilize the engagement evaluation framework with 10 MOOCs' log data offered by HKU. Some insights are derived, such as courses with short duration (4-5 weeks) tend to have higher engagement in terms of learner-interface interaction. It may be better to split the long duration MOOC (10-12 weeks) into several short MOOCs. 10 experts are invited to give comments on the proposed framework and 9 of them showed positive attitudes. Some experts have concerns about finding the specific reasons behind the bad performance. Therefore, we are considering cooperating with MOOC instructors to use our framework and further check whether there will be improvement in engagement after instructors adopting the framework.

REFERENCES

- Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice* (Vol. 28). John Wiley & Sons.
- Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., ... & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1), 30-41.
- Molenaar, I., & Knoop-van Campen, C. (2018). How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*.
- Moore, M. G. (1989). Three types of interaction.
- Hillman, D. C., Willis, D. J., & Gunawardena, C. N. (1994). Learner-interface interaction in distance education: An extension of contemporary models and strategies for practitioners. *American Journal of Distance Education*, 8(2), 30-42.
- Peterson, E. T. (2004). *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram.
- Wang, S., & Kelly, W. (2017). Video-Based Big Data Analytics in Cyberlearning. *Journal of Learning Analytics*, 4(2), 36-46.
- Hodges, C. B., & Kim, C. (2010). Email, self-regulation, self-efficacy, and achievement in a college online mathematics course. *Journal of Educational Computing Research*, 43(2), 207-223.
- Almatrafi, O., & Johri, A. (2018). Systematic Review of Discussion Forums in Massive Open Online Courses (MOOCs). *IEEE Transactions on Learning Technologies*.
- Idowu, O. M. I., & McCalla, G. (2018). Better Late Than Never but Never Late Is Better: Towards Reducing the Answer Response Time to Questions in an Online Learning Community. In *International Conference on Artificial Intelligence in Education* (pp. 184-197). Springer, Cham.
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, Brendan. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning. In *26th International Conference on Computers in Education (ICCE 2018)*, at Manila.

Automated MOOC/SPOC Learning Design Verification based on Instructional Design Theories

Chi-Un Lei, Xiangyu Hou, Jiaqi Wang, Yanxiang Guo

Technology-Enriched Learning Initiative, University of Hong Kong

Email: culei@hku.hk, hxiangyu@hku.hk, wjq1015@hku.hk, yxguo@hku.hk

ABSTRACT: Teachers often work with course development teams to implement MOOCs and SPOCs. However, existing MOOC instructional quality analysis often requires manual effort and is not supported by instructional design theories. In this paper, we propose an automated MOOC/SPOC learning design verification mechanism based on instructional design theories. The mechanism can quickly visualize the courseware with faulty or at-risk designs that may cause obstacles for learners, which allows just-in-time revisions. The mechanism can facilitate the process of verifying the course design and assessing the quality of the course, and eventually minimize learning hurdles encountered by learners.

Keywords: instructional design, verification, visualization, MOOC, SPOC

1 INTRODUCTION

Plenty of massive open online courses (MOOCs) and small private online courses (SPOCs) have been recently developed for enriching learning experiences (Lei et al., 2015; Guo et al., 2014). In order to rigidly incorporate multifaceted contents in MOOCs and SPOCs, teachers often intensively work with other colleagues to develop their courses. For example, a course development team had spent 4045 working hours for producing three MOOCs, and 22% and 24% of the production effort (in terms of working hours) was contributed by the course teacher and project manager, respectively (Hollands and Tirthali, 2014). E-learning technologists and multimedia professionals were also involved.

In practice, course quality assurance (QA) takes a significant amount of time and effort in the course development process. However, discussions on MOOC instructional quality and QA are mainly about manually analyzing MOOCs (Gamage et al. 2015; Lowenthal and Hodges 2015; Stracke 2017; Margaryan 2015). Among these frameworks, Margaryan's framework is supported by instructional design theories. However, no detailed evaluation scheme was proposed by the team. Therefore, due to the tight production schedule in reality, these frameworks are not practically helpful for adoption to analyze and rectify the design of the course. Currently, instructional designers have not yet fully explored using tools to minimize the effort and time needed for the quality assurance process.

In this paper, we aim to propose an automated MOOC/SPOC learning design verification mechanism. Based on instructional design theories gathered by the literature, the mechanism can identify and visualize faulty or at-risk courseware designs in the actually implemented courseware, from course structure level to learning object level. Such weak designs often cause obstacles for learners in participating learning activities in the course. As a result, learners either ask for peers' support, skip the problematic learning section, or even drop out of the course. The proposed mechanism can facilitate the process of self-verifying the course design and self-assessing the quality of the courses

for instructional designers. In other words, the mechanism can minimize learning hurdles encountered by learners and prevent undesirable outcomes (e.g. leading to ineffective learning) (Davies, 1999).

This paper describes how the proposed verification mechanism can be used for MOOC and SPOC designs. Faulty and at-risk designs based on instructional design theories are illustrated in Section 2. The technical implementation of the proposed mechanism is described in Section 3. The mechanism has been evaluated through identifying faulty and at-risk designs of MOOCs and SPOCs. The analyzed result is presented in Section 4. Based on the analyzed result, course instructional designers were agreed for further course design revisions. Further development and adoption of the proposed mechanism are presented in Section 5.

2 TEACHING DESIGN BASED ON ANALYTICS AND LEARNING THEORIES

2.1 Analytics-informed Teaching Design Pattern

Standardized learning design patterns (Laurillard 2013) have been proposed for modeling the learning journey, such that learning design patterns can be shared in the teaching community. An automated learning design synthesizing mechanism has been recently proposed for clustering teaching and learning design patterns in MOOCs (Davis et al., 2018). However, the team has not explored how these clustered learning design patterns could be described by existing instructional design theories or could be used for course design or verification. Meanwhile, a “teacher inquiry into student learning” model has been proposed for combining learning analytics and learning design (Alhadad and Thompson, 2017). Based on the model, a learning design studio (Law et al., 2017) has been recently developed for guiding the development of courseware designs. However, this studio requires manual course structure modeling, which is effort-demanding.

Due to the convenience of collecting learning data from learning management systems, traditional evidence-based education framework proposed by Davies (Davies, 1999) had been recently revamped by researchers (Ferguson, 2017). For example, DAPER framework (Ogata et al., 2018) had been proposed to systematically collect data, identify teaching and learning problems through statistical computations and visualizations, measure adopted interventions for producing evidence-based teaching-learning cases (TLCs), and finally reflect on aggregated TLCs for deriving good teaching and learning practices. Some of the research challenges proposed by DAPER include i) how to evaluate evidences, and ii) how to support teachers and learners to apply evidence-informed teaching and learning practices.

2.2 Faulty and At-risk Designs based on Instructional Theories and Practices

A collection of high-level design principles (corollaries) for effective and efficient instructions can be found in Merrill’s “First Principles of Instruction” study (Merrill, 2002). These principles are grouped into five big categories: problem-centered, activation, demonstration, application, and integration. Studies showed that there may be a decrement in learning when the learning design process violates or fails to implement one or more of these principles. After analyzing the courseware through these principles, in order to remove learning hurdles (i.e., prevent undesirable outcomes), revisions may include i) re-organizing learning objects for a coherent learning sequence, ii) changing problematic

settings of the learning object, and iii) adding new contents and objects according to the learning design principles.

Selected principles for implementation are shown in Table 1. The selection is based on whether the principle can i) rapidly identify the quality (or “health”) of the course for just-in-time learning design interventions, and ii) directly and automatically analyze the courseware source file package (See Section 3). For example, principles related to the “Demonstration” category requires understanding of context inside the object, which is usually course dependent and cannot be generalized. Therefore, we have not modeled principles related to the “Demonstration” category. For illustrations, we have modeled 4 out of 15 principles in the Merrill’s study (Merrill, 2002).

Table 1: High-level Instructional design principles essential for promoting students’ learning

Principles	Corresponding category	Description
Show tasks	Problem-centered	Learners are shown the problem that they will be able to solve after completing a module. (Van Merriënboer et al. 1997)
Structure	Activation	Learning is promoted when learners are provided or encouraged to recall a structure that can be used to organize the new knowledge. (Dijkstra et al. 2012)
Coaching	Application	Learners are guided in their problem solving by appropriate feedback and coaching. (Dijkstra et al. 2012) [Simplified version]
Reflection	Reflection	Learners can reflect on, discuss & defend their new knowledge or skill. (Laurillard 2002)
Adequate contents	N/A	Contents in the learning object are adequate.
Relevant parameters	N/A	Parameters in the learning object are relevant and within a reasonable range.

For ease of identification, we claim that a courseware has a “faulty” (Critically pedagogically problematic: Students cannot proceed to learn) or “at-risk” (Potentially pedagogically problematic: Students can proceed to learn, but students learn in-effectively) design if one of the following situations is true: i) The amount of learning objects in the learning journey is not in an appropriate proportion; ii) The content is pedagogically insufficient for learning; and iii) The learning object cannot convey the message clearly due to problematic technical settings in the learning object.

3 AUTOMATED LEARNING DESIGN VERIFICATION AND VISUALIZATION

The proposed mechanism can help i) identify problematic settings in learning objects through course-level and object-level verification, and ii) visualize the course design with identified problematic learning objects through course-level visualizations. The mechanism can be adopted in any learning management systems that can export courseware design packages (e.g. (Open) edX, Moodle and Canvas). For other LMSs, course developers can also manually analyze the course and import data based on principles shown in Table 1. In this paper, we used Open edX courseware packages for illustrations. In Open edX, the course design is represented by a zipped package of

courseware XML files which specify course objects, including Chapters, Sections, Subsections, Units, and Components/Learning objects, course structure, course assets and course settings.

3.1 Course-level and Object-level Verification

Verification is to ensure the implemented learning design does not violate pre-defined learning design rules. Learning design rules specify restrictions to ensure all learning components function correctly. Table 2 shows detailed design rules for edX/Open edX courses with their corresponding pedagogical problem severity as well as violations of learning design principles listed in Table 1. In the verification process, learning design and course object design parameters are first extracted from XML files. Parameters are then automatically checked for faulty or at-risk designs. If there is no violation, then the learning design passes the inspection. If any faulty or at-risk designs are identified, they will be reported through visualizations, for revision of the courseware design.

Table 2: Detailed design rules for edX/Open edX courses, based on principles shown in Table 1.

Course Structure					
Item description	Severity	Violation	Item description	Severity	Violation
There is no course image or course overview on the landing page.	At-risk	Learner guidance	Number of assessment objects is different from the assessment setting.	Faulty	Relevant parameters
Section, Subsection or Unit is empty.	Faulty	Adequate contents	There is no forum in the course.	At-risk (SPOC) Faulty (MOOC)	Reflection
There is no problem in a Section.	At-risk	Show tasks	There is no introduction for first-time learners.	At-risk	Learner guidance
Learning Object					
Item description	Severity	Violation	Item description	Severity	Violation
The transcript is not available for videos.	At-risk	Learner guidance	The provided link is broken.	Faulty	Adequate contents
The video, images or iframe objects cannot be loaded.	Faulty	Adequate contents	Forum category and subcategory have not been named.	At-risk	Structure
Video, HTML, question, third-party object or forum has not been named.	At-risk	Structure	There is no Learning Tools Interoperability (LTI) ID, LTI URL for third-party objects connected by LTI.	Faulty	Relevant parameters
There is no correct answer in the question.	Faulty	Relevant parameters	There is no hint/feedback for questions.	At-risk	Feedback
There are no pedagogical settings for the assessment question, including the number of maximum attempts, the time required between attempts, the selection of the option for answer retrieval and question resetting.				At-risk	Feedback

3.2 Course-level Visualization

To facilitate instructional designers and teachers analyzing the course design, the tool will visualize the overview of the courseware design, including the number of sections, subsections, units, learning objects and their corresponding locations. The overview can help instructional designers estimate the workload of each section and re-structure contents in sections if contents among sections are not balanced or aligned. Furthermore, identified faulty and at-risk designs, based on instructional design rules shown in Table 2, will also be shown in the visualization. Based on the verification results, problematic components identified in the verification process will be labeled in a subsection level. Based on the visualization, instructional designers can identify faulty/at-risk objects and decide on possible revisions of the courseware.

4 EVALUATION OF THE PROPOSED MECHANISM

4.1 Adoption for Analyzing the Design of a Launched On-Campus SPOC

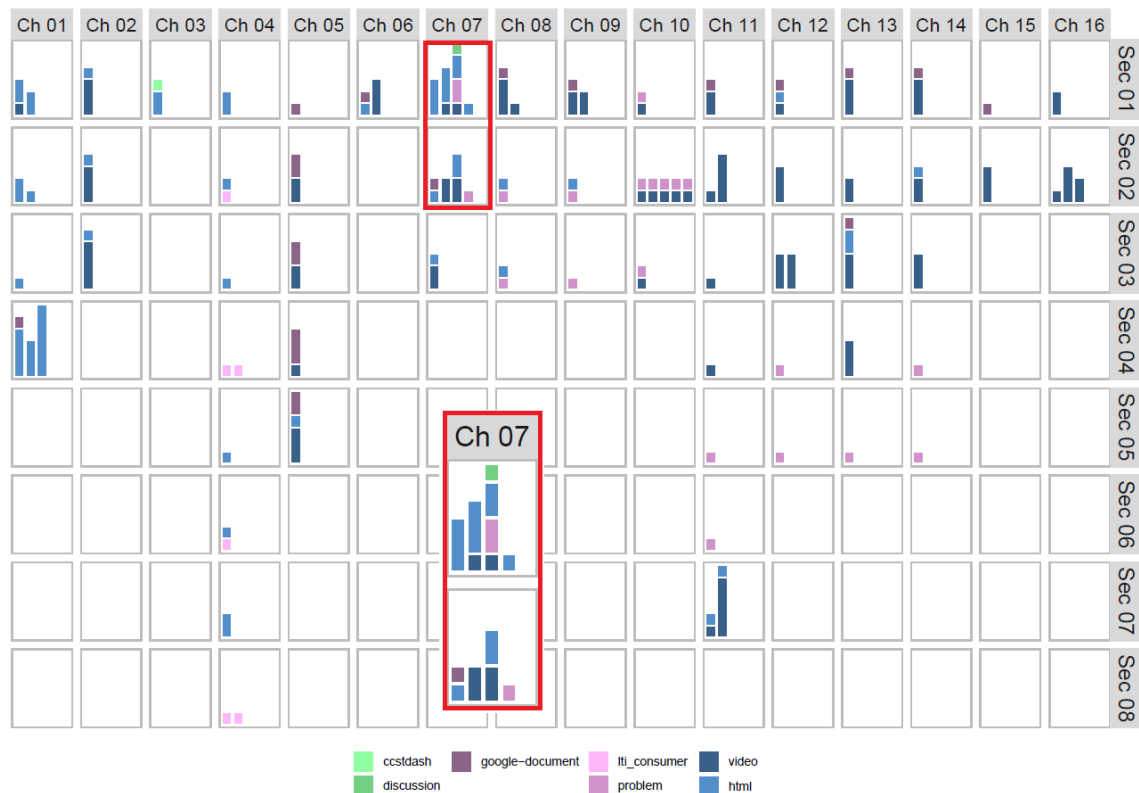


Figure 1: The verified SPOC design: course structure. The middle bottom figure is a zoomed figure showing the structure of Chapter 7 Sections 1 and 2.

We have adopted the proposed mechanism in a 13-week on-campus general education SPOC. The course was delivered using the flipped classroom approach: students remotely learned concepts of algorithmic design via online videos and quizzes in advance followed by face-to-face learning activities. As shown in the course structure visualization (Fig. 1), each block represents a Section in the course. Its horizontal location is the index of the Chapter, and the vertical position is the index of the Section. The stacked bars in the block show the structure of Section inside. To be specific, the

number of bars means the number of subsections in each section, where each bar describes units (contents) inside the Subsection. Colors of the bar indicate different types of contents (e.g. Green: Logistics-related; Purple: Assessment-related; Blue: Content-related). The figure shows that the contents are unevenly distributed among chapters, however, it is typical since contents are arranged according to topics rather than teaching weeks. Videos have been heavily used for online activities (Chapters 2-16) except the logistic announcements section (Chapter 1). Google Docs have also been used for online collaborative writing, which is shown to be effective as a pre-class activity.

The instruction for visualization of warnings (Fig. 2) is similar with the course structure, substituting the component type with the issue type (e.g. Red: Faulty; Gray: At-risk), with Section 0 indicating chapter-level issues. The figure shows that questions in chapters can have a better design, such as including more hints or providing concept clarification sessions during the classwork session.

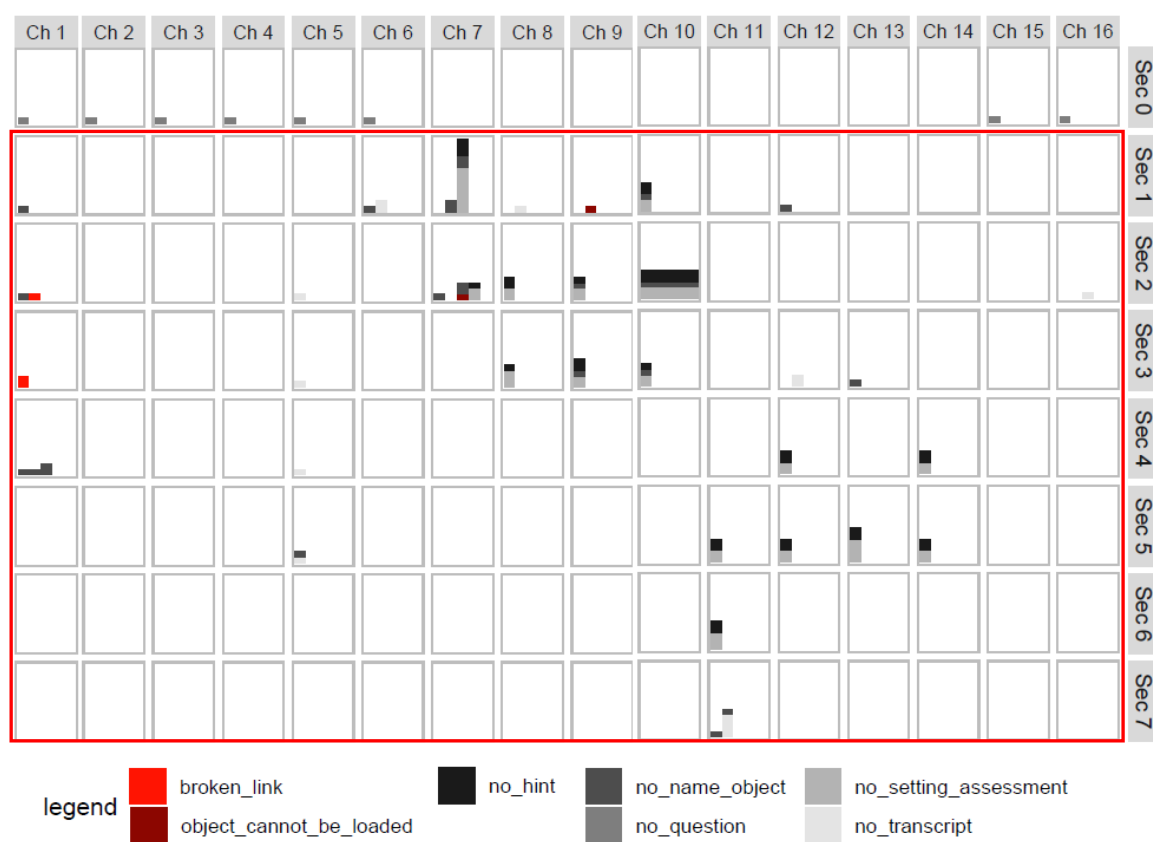


Figure 2: The verified SPOC design: faulty/at-risk components.

The analyzed results had been sent to the course teacher for reflections. She reflected that the analyzed results listed the learning elements with real-time insights on the pedagogical strength, for prompt revisions. She was agreed for revising the course design next cohort, based on generated insights. She reported that it was common that some learning materials were not updated in time, especially when there were multiple offerings of one course in the same academic year. The mechanism addresses these issues with a design of not allowing faulty content to be published and flagging the content that is at risk. She recommended the mechanism can offer in-depth pedagogical guidance through aggregating (sub-)section-level content, since the pedagogy is manifested in not only a single learning activity, but also a series of learning objects arranged in particular sequences.

4.2 Adoption in Analyzing the Design of a Work-In-Progress MOOC

We adopted the proposed mechanism to analyze a MOOC that is in the development stage. The course is about calculus and differential equations and is taught by two teachers. The structure visualization (Fig. 3) shows that there are many interactive learning components which provide a variety of learning experiences for learners that could not be experienced in face-to-face sessions. However, the course still has faulty and at-risk designs (Fig. 4), and thus is not ready for launch. For example, the assessment tasks are not yet well designed, without feedback or learner's guidance provided for questions. Furthermore, the learning design for the first part of the content (Chapters 2-8) is different from the second part of the content (Chapters 9-14) (e.g. how assessments are arranged in the Section level). This is caused by the difference of teaching rationales of the two teachers. After exploring the design, the learning designer decided to redesign the course by i) including more assessment tasks with informative feedback and hints, ii) simplifying contents shown in Chapters 3 and 4, iv) introducing more contents for Chapters 5-8, and iii) revise the contents of chapters for minimizing faulty and at-risk design issues.

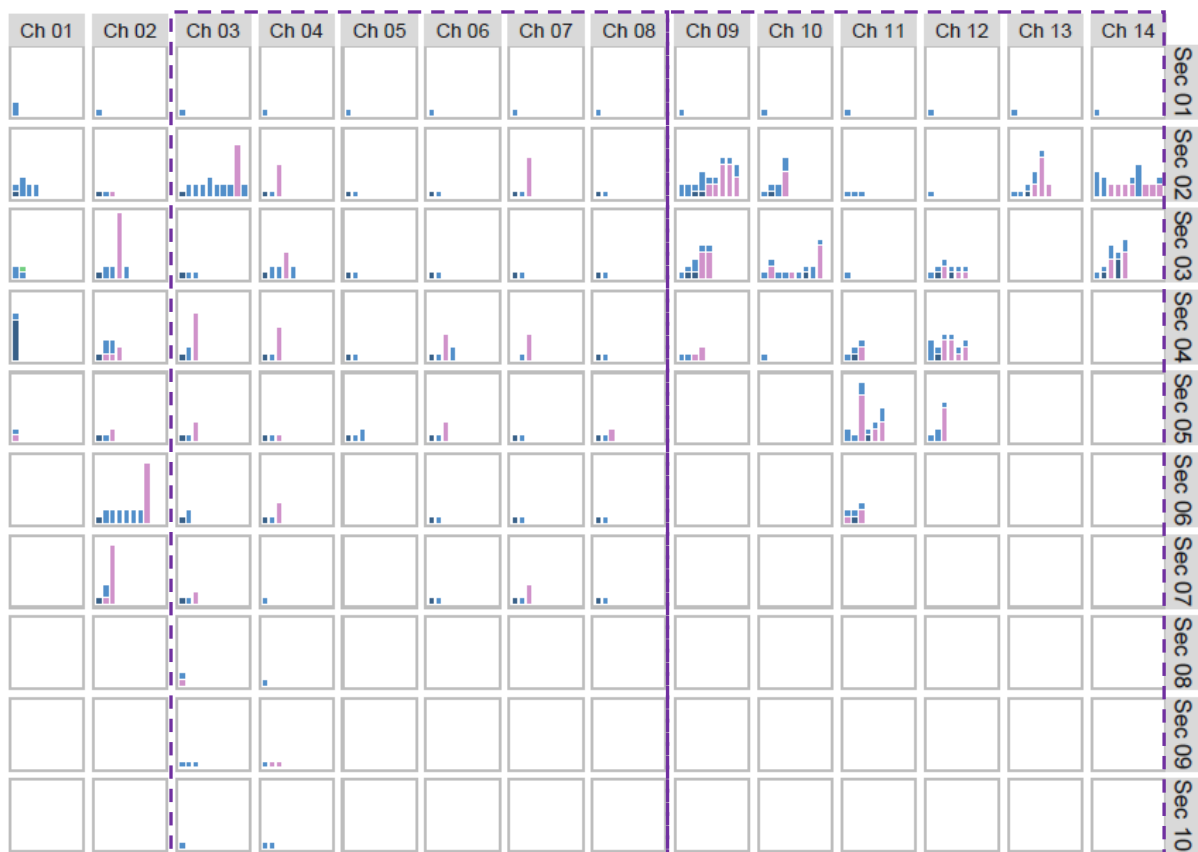


Figure 3: The verified MOOC design: course structure.

4.3 Analyzing the Learning Design of MOOCs and SPOCs

To furthermore illustrate the performance of the proposed mechanism, the course structure of another ten MOOCs and SPOCs were also analyzed. Table 3 describes the analyzed results. In summary, MOOCs tend to have more learning components (e.g. forum discussions and assessments). This is because blended SPOC learners usually have both online and on-campus

learning and assessment experiences, but not for MOOC learners. In total, there are with 393 faulty and 2731 at-risk instructional design warnings on these courses, which may be difficult and effort-demanding to be identified manually. This indicates the needs for automatic verifications. For example, pages in HKU03x contain insecure links to external resources, leading to a high number of faulty warnings. It also does not design questions with correct pedagogical settings, leading to a high number of at-risk warnings. The tool provides an efficient way for estimating learners' workload.

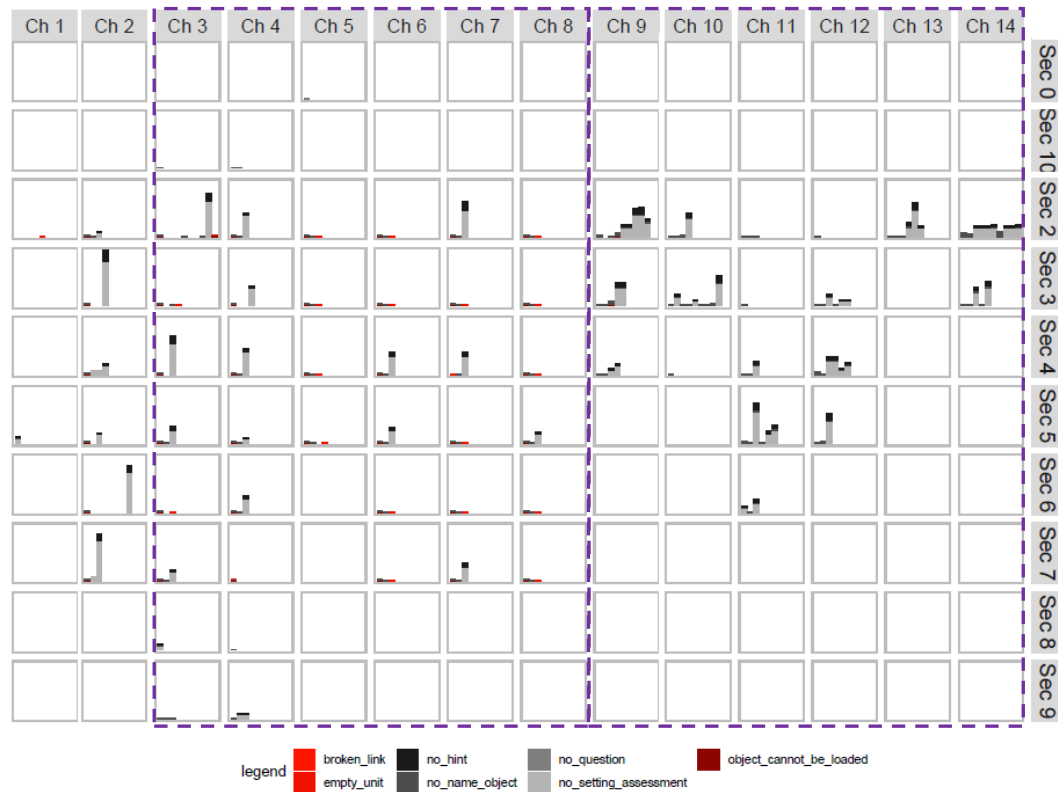


Figure 4: The verified MOOC design: faulty/at-risk components.

Table 3: Summary of the analysis

Course Code	Course Type	Number of Chapters / Sections / Sub-sections	Number of HTML / Video / Problem / Forum / Other components	Number of faulty/at-risk warnings
ARCHx	MOOC	7/59/115	50/80/94/25/9	81/381
HKU01x	MOOC	7/33/131	51/97/124/11/2	24/538
HKU03X	MOOC	14/150/245	117/126/238/8/2	201/961
HKU04x	MOOC	7/36/84	128/63/31/16/3	25/129
HKU06x	MOOC	10/125/219	80/98/78/18/4	13/332
ELEC3542	SPOC (On-campus)	12/28/31	20/26/0/0/4	0/16
CCHU9001	SPOC (On-campus)	13/64/65	83/19/0/0/4	4/32
NURS	SPOC (Training)	7/27/91	147/7/20/0	23/92
IOL	SPOC (Training)	5/9/14	9/7/19/0/0	0/108
ILT	SPOC (Training)	6/23/57	78/35/25/7/0	22/142

Table 4 shows some of the popular issues identified by the mechanism. Most issues are related to URL links shown on the courseware as well as inappropriate design of assessment questions, which may be effort-demanding to be identified manually. This tool can quickly identify issues for revisions, which leads to a more efficient development progress and eventually a more effective learning.

Table 4: Major issues identified by the mechanism (10 MOOCs).

Issues	Severity of the issues	Number of issues
The provided link cannot be loaded.	Faulty	305
The provided link is broken.	Faulty	80
There is no pedagogical settings for the question.	At-risk	1889
There is no hint/feedback for questions.	At-risk	748
The component has not been named.	At-risk	31
There is no problem in a Section.	At-risk	48

5 CONCLUSIONS

An automated course learning design verification mechanism, based on instructional design theories, has been proposed in this paper. Through the mechanism, problematic designs can be identified and revised immediately, for preventing undesirable outcomes (learning obstacles). The mechanism has been evaluated through identifying faulty and at-risk designs of twelve MOOCs and SPOCs. Possible extensions for the mechanism include i) verifying MOOCs/SPOCs implemented in other LMSs, ii) auto-correcting learning design with faults, and iii) auto-identifying good teaching design patterns. Natural language processing techniques could also be introduced to understand the contents in learning objects for a more informative instructional design analysis.

REFERENCES

- Alhadad, S. S., & Thompson, K. (2017). Understanding the mediating role of teacher inquiry when connecting learning analytics with design for learning. *Interaction, Design, & Architecture (s)*, 33, 54-74.
- Davies, P. (1999). What is evidence-based education?. *British journal of educational studies*, 47(2), 108-121.
- Davis, D., Seaton, D., Hauff, C., & Houben, G. J. (2018, June). Toward large-scale learning design: categorizing course designs in service of supporting learning outcomes. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (p. 4). ACM
- Dijkstra, S., Schott, F., Seel, N., Tennyson, R. D., & Seel, N. M. (2012). *Instructional Design: International Perspectives: Volume I: Theory, Research, and Models; Volume II: Solving Instructional Design Problems*. Routledge.
- Ferguson, R., & Clow, D. (2017, March). Where is the evidence?: a call to action for learning analytics. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 56-65). ACM.

- Gamage, D., Fernando, S., & Perera, I. (2015, August). Quality of MOOCs: A review of literature on effectiveness and quality aspects. In *Ubi-Media Computing (UMEDIA), 2015 8th International Conference on* (pp. 224-229). IEEE.
- Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: an empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@scale conference* (pp. 41-50). ACM.
- Hollands, F. M., & Tirthali, D. (2014). Resource requirements and costs of developing and delivering MOOCs. *The International Review of Research in Open and Distributed Learning*, 15(5).
- Laurillard, D. (2002). *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge.
- Laurillard, D. (2013). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge.
- Law, N., Li, L., Herrera, L. F., Chan, A., & Pong, T. C. (2017). A pattern language based learning design studio for an analytics informed inter-professional design community. *Interaction Design and Architecture (s)*, 92.
- Lei, C. U., Hou, X., Kwok, T. T., Chan, T. S., Lee, J., Oh, E., ... & Lai, C. (2015, December). Advancing MOOC and SPOC development via a learner decision journey analytic framework. In *Proceedings of the International Conference on Teaching, Assessment, and Learning for Engineering*, (pp. 149-156). IEEE.
- Lowenthal, P., & Hodges, C. (2015). In search of quality: Using Quality Matters to analyze the quality of massive, open, online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, 16(5).
- Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation* (Vol. 41, pp. 85-139). Academic Press.
- Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education*, 80, 77-83.
- Merrill, M. D. (2002). First principles of instruction. *Educational technology research and development*, 50(3), 43-59.
- Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, Brendan. (2018). Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning. In *Proceedings of the 26th International Conference on Computers in Education*.
- Stracke, C. M. (2017, July). The Quality of MOOCs: How to improve the design of open education and online courses for learners? In *International Conference on Learning and Collaboration Technologies* (pp. 285-293). Springer, Cham.
- Van Merriënboer, J. J. (1997). Training complex cognitive skills: A four-component instructional design model for technical training. *Educational Technology*.

CLEAR: Cohort-Level Evidence Analysis and Reflection Process as a methodology to assist MOOC Providers and Adopters for effective teaching-learning using MOOCs

Jayakrishnan Madathil Warriem*, Bharathi Balaji

Institution: Indian Institute of Technology Madras, Chennai, India

Email: jkm@nptel.iitm.ac.in

ABSTRACT: Persistent engagement of learners with Massive Open Online Courses (MOOCs) has been a consistent challenge faced by MOOC providers over the years. There are multi-pronged strategies applied to solve this problem like refining the pedagogy, giving more instructor presence through periodic virtual interactions etc. Gathering of evidences of learner actions in MOOCs and providing insights to take quick remedial actions are important to ensure effectiveness of learner participation. The National Programme on Technology Enhanced Learning (NPTEL), a national MOOC initiative from India, is trying to solve this unique challenge by adopting a multi-pronged strategy by collaborating with local institutions and providing mechanisms for supporting learners from these institutions. In this paper, we look at the existing evidence gathering mechanisms used by NPTEL team and explain one of the follow-up action based on this evidences. Based on the insights from this experience we propose the Cohort-Level Evidence Analysis and Reflection (CLEAR) process flow that will assist the institutions signed up in NPTEL MOOCs to utilize the evidences available and take meaningful actions on it. We expect that this process will be helpful for both MOOC providers as well as institutions adopting hybrid online learning practices for making the teaching-learning process more effective for the learners.

Keywords: Cohort-level Analysis, MOOCs, Hybrid Online Learning, Evidence-based Practice, Reflection, Learning Analytics

1 INTRODUCTION

With questions being raised about democratization of education through MOOCs (Hansen & Reich, 2015) and the increasing disparity in achievement gaps (Kizilec, Saltarelli, Reich, & Cohen, 2017), the challenges faced by MOOC providers are no longer just about solving the completion rate problem. Mounting research evidence have indicated the need for improvement in the existing pedagogical designs (Bali, 2014; Hew, 2016) and the need for leveraging local expertise to promote social learning experiences around MOOCs (Anders, 2015). While there are several research based models to assist MOOC creators (Murthy, Warriem, Iyer, & Sahasrabudhe, 2018; Fassbinder, Barbosa, & Magoulas, 2017) to improve design of their MOOCs and instructors integrating MOOCs through flipped classroom designs (Rodríguez, Correa, Pérez-Sanagustín, Pertuze, & Alario-Hoyos, 2017), there aren't any specific orchestration guidelines that will enable either MOOC providers or the large-scale MOOC adopters like educational institutions to improve MOOC orchestration. The solutions to tackle this gap should go beyond providing analytics dashboards (for individual courses) to a more actionable set of methodologies or processes that can be adopted at the institutional level.

The DAPER model (Majumdar, Akçapinar, Hasnine, Flanagan, & Ogata, 2018) provides a clear framework for setting up these methodologies at an individual course level. In this paper we look at how the DAPER model can be made relevant for MOOC providers and adopters. We propose the Cohort-Level Evidence Analysis and Reflection (CLEAR) Process, developed on the basis of DAPER model, and elaborate the actions that have been taken by the MOOC providers to increase the effectiveness of MOOC integration in regular teaching-learning process. The CLEAR Process looks at evidences gathered at a cohort-level and provide probing questions that assist the MOOC providers and adopters to reflect on the impacts. This further leads to actionable items that can generate more evidences and thus providing a positive feedback loop.

A major difference of the CLEAR Process from the DAPER model is that it is aimed to directly support the operational activities of the MOOC provider and adopter.

2 NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

2.1 Background

The National Programme on Technology Enhanced Learning (NPTEL) is an initiative by Government of India to provide access to quality higher education content for the learners. From its inception in 2003, the NPTEL initiative has organically evolved and currently utilizes the modalities of Massive Open Online Courses to achieve the larger objective of providing access to quality learning for learners. This is a joint initiative of 8 premier Science and Technology institutions in India (7 IITs and IISc). The learners of the NPTEL MOOCs, also called NPTEL Online Courses (NOCs), have the opportunity to get certified through a proctored certification exam at the end of the course. Over the last four years, NPTEL has offered 1294 courses in MOOC format and has seen 4.84 million enrollments and 0.32 million certifications (NPTEL, 2013). Table 1 below shows data related to the evolution of NPTEL Online Courses over the past few years.

Table 1: Growth of NPTEL Online Courses (NPTEL, 2003)

Offering Period	Courses	Enrollment	Exam Takers		Certified
			Overall	From Local Chapter * Reg. Count & Percentage	
Jan-June 2014	1	53807	1182	NA	546
July-Dec 2014	2	58947	1549	NA	1526
Jan-June 2015	18	89045	2113	NA	1931
July-Dec 2015	36	160819	6006	2127 (29.17%)	3165
Jan-June 2016	64	241691	15310	12296 (71.10%)	10331
Jul-Dec 2016	104	401176	26544	22047 (61.67%)	19595
Jan-June 2017	130	535223	38405	35942 (77.52%)	31117
Jul-Dec 2017	159	1049265	63398	50042 (71.17%)	54092
Jan-June 2018	226	934182	76125	64567 (74.25%)	66167
July-Dec 2018	269	1330816	161256	132875(82.4%)	124270

With its context firmly rooted in the higher education setting within India, NPTEL has adopted a unique strategy to further scale-up and strengthen the MOOC initiative. The strategy used is the formation of NPTEL-Local Chapters (LC) among institutions that want to take up courses from NPTEL. These institutions belong to the category of MOOC adopters and they formalize local strategies to integrate MOOC in regular academic curriculum. NPTEL Local Chapter (LC) is defined as a “focused teaching-learning community within an educational institution whose primary aim is to support the MOOC learners throughout the entire MOOC cycle” (Warriem, 2018). NPTEL has established more than 2000 such LCs across the country till now (NPTEL, 2016).

The activities in the LC is coordinated by a Single Point of Contact (SPoC) who will be either a faculty or non-teaching staff from the institution. SPoCs can nominate a set of mentor faculty to help the learners from the LC better understand the concepts within the online course. The learners from the LC have the freedom to chose whether they need the support of the mentor or not. The strategies adopted by the mentor can vary from procedural tasks like emailing reminders for assignment submission and other MOOC weekly learning activity to more advanced pedagogic strategies like Flipped Classroom to engage learners. Figure 2 shows an overview of the way in which LCs function within the NPTEL Online Course system.

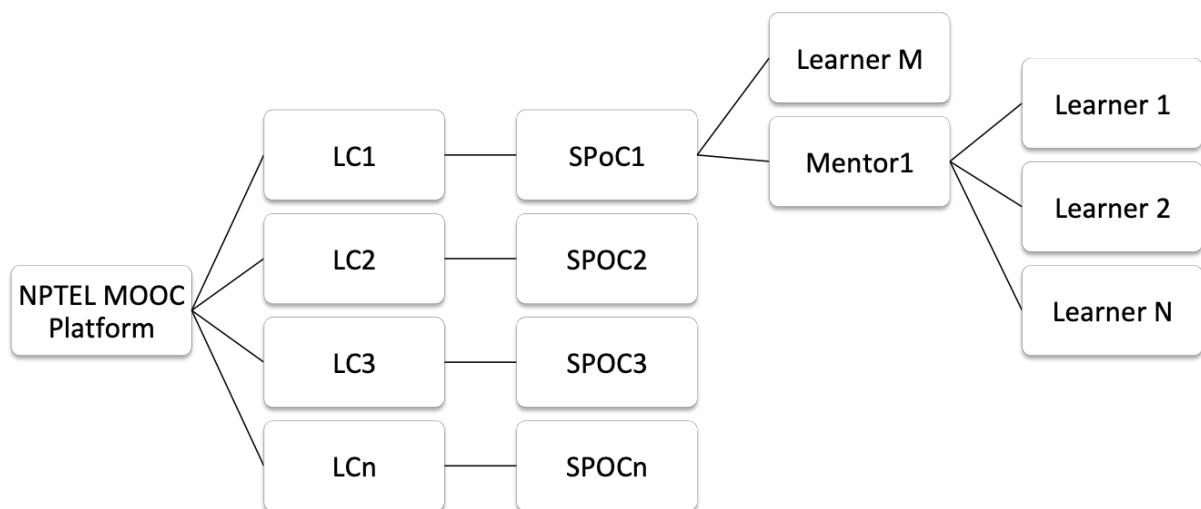


Figure 1: Overview of roles within a Local Chapter

Figure 2 shows the interface of NPTEL Online courses (NOCs) as visible to a learner. NOCs use the Google Course Builder platform as the Learning Management System and any learner having a Google Authenticated email address can freely enroll in the course and access the course resources. The course content (videos and solved assignments) is separately archived in NPTEL Course repository (NPTEL, 2003) and is openly accessible to anyone. Thus, NPTEL remains truly massive, open and online for the learners

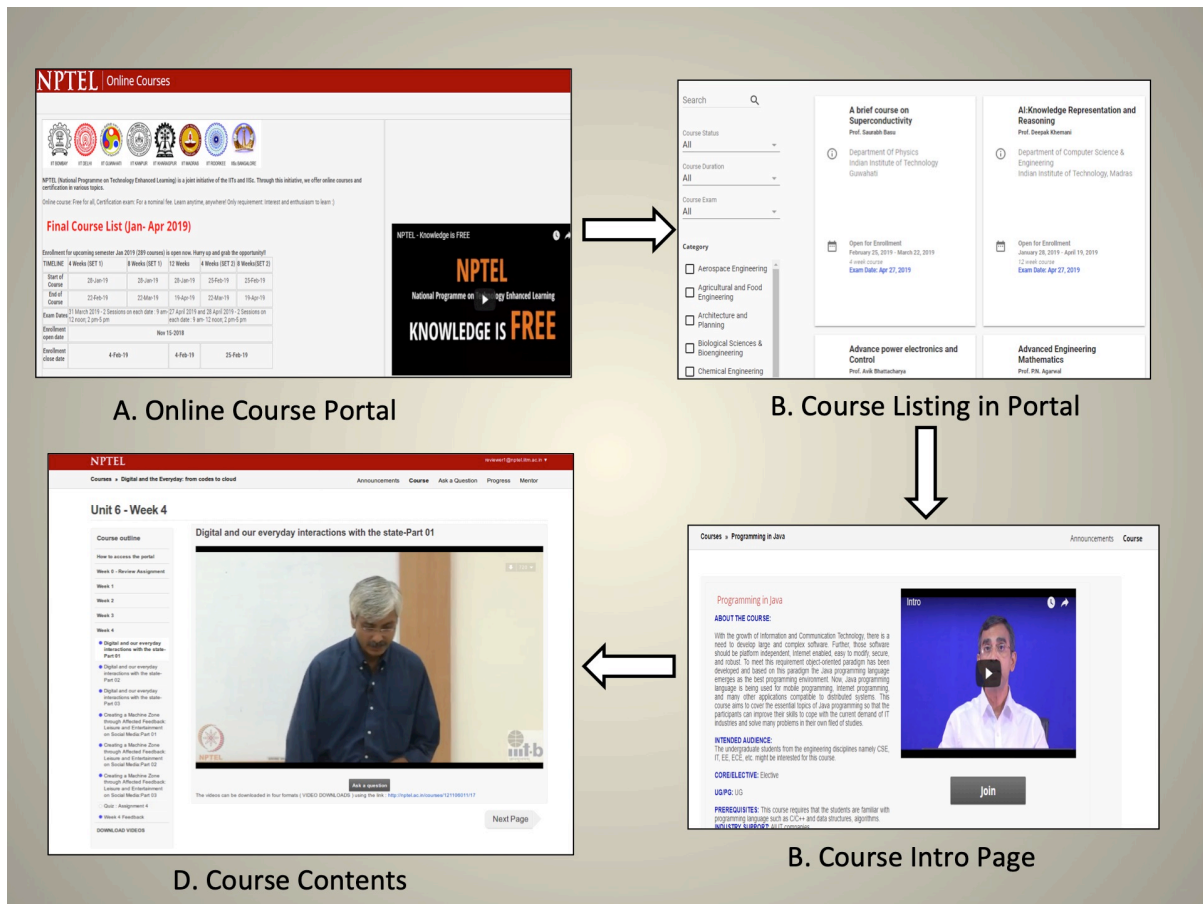


Figure 2: NPTEL Online Course Interface (From course selection till learning)

2.2 Data provided for SPoCs and Mentors

To monitor and review the activities of the learners within the Local Chapter, the SPoC is provided with a separate dashboard within the Learning Management System as well as within the main NPTEL website that will inform them of the learner status and performance. A sample screenshot of the SPoC's view inside the MOOC is shown in Figure 3. Figure 4 shows screenshot of a SPoC public webpage available within the NPTEL website where performance of the learners from the LC in the NPTEL Online Certification exam are archived.

COURSE BUILDER

Courses

Course

Steps

Publish

Manage

All Local Chapter Members

Appraisals

Settings

Help

Programming Assignments

Course Staff

Manage > All Local Chapter Members

kamala@npTEL.ilm.ac.in | Logout

Members for All Local Chapter Members

Number of rows to show in a page: all Order By: Email Invert Clear

Search by email: Email Search Clear

User ID	Email	Name	Date of Birth	Mobile	Enrolled Courses	Country	State	City	College	Roll Number	Employer	Age
1	11250502468585862093				• Enhancing Soft Skills and Personality	IN	Tamil_Nadu	VIRUDHUNAGAR				25
2	114376217909338355371				• Programming in Java	IN	Tamil_Nadu	VIRUDHUNAGAR				12
3	112987285379331622334				• Programming in Java	IN	Tamil_Nadu	VIRUDHUNAGAR				12
4	11768984849443655449				• Fundamentals of Power Electronics	IN	Tamil_Nadu	VIRUDHUNAGAR				25

Figure 3: A sample SPoC View inside the NPTEL MOOC Portal

Jan- Apr 2018 CONGRATS! You are one of the top 100 Local Chapters. Your college is hereby recognized as an ACTIVE Local Chapter.						
Course Run	Present	Gold	Elite	Successful	Participation	Topper
Jan-Apr 2018	2441	1	227	1843	370	40
Jul-Dec 2017	1286	8	449	588	241	28
Jan-June 2017	1478	1	253	1066	158	15
Jul-Dec 2016	790	1	400	370	19	24
Jan-June 2016	879	0	136	635	108	12

Figure 4: Archived statistics of performance of learners from an LC in NPTEL Certification Exam

The mentors are also provided with a page within the MOOC portal to track the progress of the mentee in terms of their weekly assignment scores (see fig 5). By providing these dashboards, the NPTEL project is allowing SPOCs and Mentors (who are the MOOC adopters) within the LC to reflect both during and after the MOOC offering on possible strategies that they can plan around it.

NPTEL

Courses » Announcements Course Ask a Question Progress Mentor **Mentee List** FAQ

Mentee Details

Name	Email	Assignment 0	Assignment: Introduction to Online Portal	Programming Assignment-1: Difference	Programming Assignment-2: Difference
<input type="text"/>	<input type="text"/>	✓			
<input type="text"/>	<input type="text"/>	✓			
<input type="text"/>	<input type="text"/>	✓			

* ✓ implies the student has submitted but the assessment due date has not passed yet.

Register for Certification exam

Course outline

How to access the portal

Week 1: Introduction

DOWNLOAD VIDEOS

TEXT TRANSCRIPTS

Figure 5: Data that mentor will see in his login regarding course progress of mentees

2.3 Need for cohort-level evidence gathering

From the previous sections, it is evident that NPTEL has setup a support structure to scale up and sustain its efforts. Such a structure enables MOOC providers to look at aggregate data and focus on creating facilitating interventions for the support group below it. However, much like the analysis of learning data from an individual class, the challenge still remains to interpret the cohort-level data against the intended goal and the local context to evaluate the effectiveness (Dawson, Bakharia, Lockyer, & Heathcote, 2010) of targeted intervention. The DAPER model (Majumdar, Akçapinar, Hasnine, Flanagan, & Ogata, 2018) provides a possible framework through which we can perform the data collection and analysis to find the existing problems and then designing intervention strategies to mitigate these problems. Though the model focuses on the indicators at individual learner level, it can also be adapted to a cohort-level data within a MOOC setting where the teacher gets replaced by either the MOOC providers or adopters (See Figure 5).

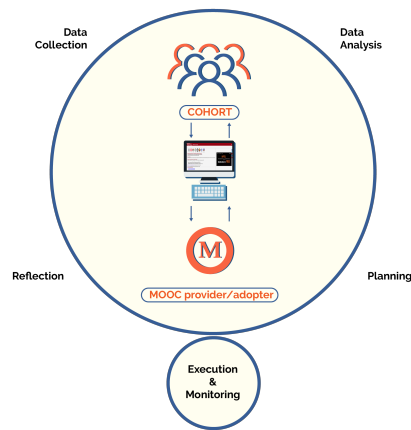


Figure 5: Adapting DAPER Model (Majumdar et. al., 2018) for cohorts

In the next section, we present a case study of an ongoing intervention designed for increasing effectiveness of NPTEL MOOCs.

3 CASE STUDY OF LOCAL CHAPTER INCENTIVIZATION

The broad goal of NPTEL initiative is to make quality higher education accessible for all. The evidence of increasing numbers of courses and learners within NPTEL MOOCs over the last four years (see Table 1, page 2) provides us an insight that diffusion of MOOCs in India have moved from an early adoption stage to an early majority stage (Rogers & Shoemaker, 1971). Thus, the goals of NPTEL have to slowly shift from ‘accessibility to learning content’ to ‘effectiveness in learning’ for the majority. When we specifically take the case of Local Chapters, this goal will translate to “Is MOOC adoption effective at the Local Chapter?” We now explain a specific case study that explains how the MOOC provider, based on the evidences available, is changing an existing policy. We then explain the impact of this decision for the MOOC adopters.

3.1 Context

To build the network of Local Chapters, initially it was important that NPTEL team looked into incentivizing the institution for their participation as a Local Chapter. Thus, every local chapter is assigned a rating point based on the number of learners (students and faculty) who participated in the certification exam. The top 100 Local Chapters with maximum rating will then be publicly acknowledged and felicitated. The rationale of such an incentivization is to ensure that other Local Chapters will work on improving their rating points in the subsequent offering and thereby facilitate better learning from NPTEL MOOCs. The existing rating point formula is:

*$R = 0.1 * \text{Number of learners from LC writing exam (Capped at 10)} + 1 * \text{No of people getting marks in between 40 and 59 (Successfully Completed)} + 2 * \text{No of people getting marks in between 60 and 89 (Elite)} + 10 * \text{No of people getting marks greater than or equal to 90 (Gold)} + 20 * \text{No of toppers in a course (Toppers)}$*

This formula was being used from till January 2018 offering and this helps in identifying existing evidences from the data about effectiveness of MOOC adoption at the local chapter. We use the CLEAR process to come up with alternate recommendations for the rating policy so as to achieve the broader goal of effectiveness of MOOC adoption at Local Chapter.

3.2 Applying CLEAR process

The Cohort-Level Evidence Analysis and Reflection (CLEAR) process is built on the DAPER model by adding two phases before the start of data collection and after data analysis (see figure 6). Though these processes are implicit in the DAPER model, it is important for making the goal setting and reflection explicit to the practitioners (MOOC providers and Adopters) so that they are not drowned in the data that they obtain from the MOOC.

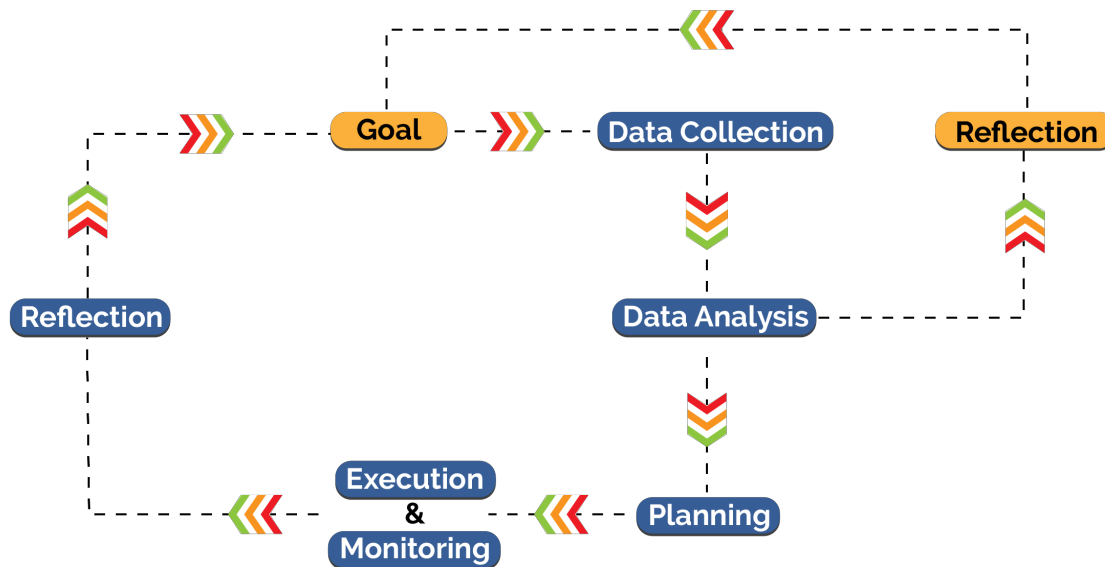


Figure 6: Overview of applying CLEAR Process on DAPER Model

In the case of Local Chapter rating problem, the broad question that drove the entire CLEAR process was “Does the current rating mechanism for local chapters align with the larger goal of NPTEL to provide greater accessibility to quality learning?” Thus, the broad goals were identified as increasing “Number of learners” (Accessibility) and “Learning Scores”(Quality). This enabled the data collection to be focused on getting the count of learners within each mark range used for local chapter rating. A sample set of data for five LCs is given in table 2. The reflection process in this phase is guided by the question that compares these two goals - “Does greater access lead to better quality?”

Table 2: Rating Points and the Mark distribution of learners

Rating	Local Chapter	Exam Takers	Total Certified	Toppers	Distribution of Marks				
					< 40	40-59	60-74	75-89	90+
1	LC1	2535	2457	61	78	1254	897	293	13
6	LC2	1847	1403	27	444	876	342	136	49
10	LC3	1334	1213	31	121	485	553	160	15
21	LC4	1309	1034	16	275	687	267	69	11
32	LC5	483	472	29	11	69	260	131	12

As seen from the table, the LCs with a higher number of exam takers have larger proportion in the lower scoring bins (<60). Thus, it necessitated revisiting the rating criteria and providing an upper

limit to the number of learners being counted in the 40-59 bin for rating calculation. Additionally, the weights for bins were incremented progressively by introducing a new mark range of 75-89. Thus the new rating criteria decided was:

$R = 0.1 * \text{Number of learners from LC writing exam (Capped at 10)} + 1 * \text{No of people getting marks in between 40 and 59 (Capped at 100)} + 2 * \text{No of people getting marks in between 60 and 74} + 5 * \text{No of people getting marks in between 75 and 89} + 8 * \text{No of people getting marks greater than or equal to 90 (Gold)} + 10 * \text{No of toppers in a course}$

By applying this new rating, the table has been modified as given in Table 3 below:

New Rating	Local Chapter	Exam Takers	Total Certified	Toppers	Distribution of Marks				
					< 40	40-59	60-74	75-89	90+
1	LC1	2535	2457	61	78	1254	897	293	13
10	LC2	1847	1403	27	444	876	342	136	49
6	LC3	1334	1213	31	121	485	553	160	15
35	LC4	1309	1034	16	275	687	267	69	11
20	LC5	483	472	29	11	69	260	131	12

Once this analysis is done, we reflect on the impact of this new rating on the existing rating system. We see that only 12% of the Local Chapters move out of the top 100 due to the rating change, which is not a significant impact.

3.2.1 Impact of CLEAR Process

To enable the MOOC adopters to subsequently take action on achieving the goals of accessibility and quality of learning, NPTEL will be implementing one additional tab in the SPoC portal that will allow them to reflect on the existing results and plan for the upcoming iteration (see fig 7 below). By clicking on each of the action buttons, they will be taken to a form that will elicit the reflection on current results and plan for the upcoming offering. At the semester end, we will be analyzing the plans created by each Local Chapter and the corresponding results to identify a set of best practices that can be adopted by other Local Chapters.

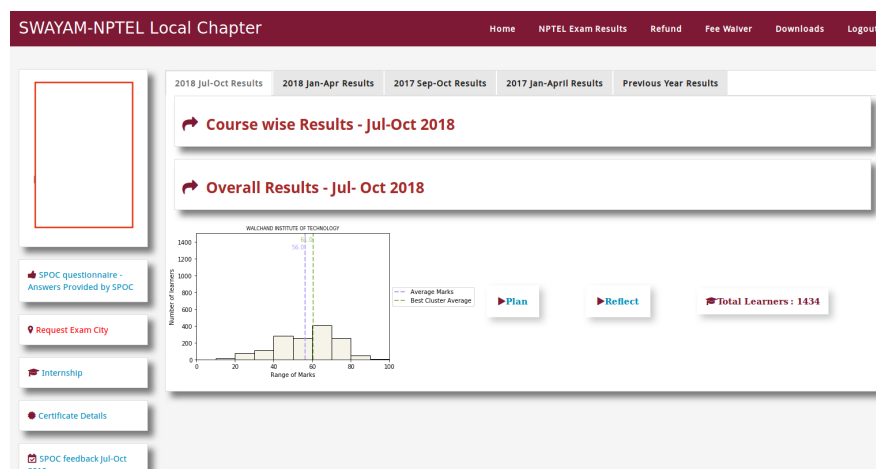


Figure 7: The modified SPoC Login Interface with buttons for planning and reflection

4 FUTURE WORK

The Cohort Level Evidence Analysis and Reflection process is shown to help MOOC providers and adopters by aggregating the analysis and presenting evidences directly. By aligning the entire process to a starting goal, we are able to collect minimum information required to answer the larger questions in the first iteration. The reflection questions at the end of analysis are then helping us to go deeper and get clarity in our understanding of the larger goal. In this paper, we presented an example, Local Chapter incentivization formula that was done using this process. This has resulted in follow-up actions at different levels of the operation of NPTEL MOOC, specifically in our communications with local chapter. At the local chapter level, this is expected to inform the local actions.

In the current paper, we stop at explanation of goal setting and reflection question formulation for CLEAR process. We have not detailed the complete iteration and the final rating point formula that was being adopted. The insights from the data analysis are being incorporated in the planning activity for the current semester. We believe that leveraging such local actions are critical for democratization of knowledge through MOOC and CLEAR helps in making this a more realistic goal.

ACKNOWLEDGEMENT

The authors would like to thank the entire NPTEL Team for providing us the support during the execution of the study. We also acknowledge the support provided by Mr. Yogesh and Mr Rajavel for processing the data and the insights and guidance provided by Prof. Prathap Haridoss, Prof. Niket Kaisare and Prof. Andrew Thangaraj during our effort.

REFERENCES

- Anders, A. (2015). Theories and Applications of Massive Online Open Courses (MOOCs): The Case for Hybrid Design. *International Review of Research in Open and Distributed Learning* , 16 (6), 39-61.
- Bali, M. (2014). MOOC pedagogy: gleaning good practice from existing MOOCs. *Journal of Online Learning and Teaching* , 10 (1), 44.
- Dawson, S., Bakharia, A., Lockyer, L., & Heathcote, E. (2010). 'Seeing' networks: visualising and evaluating student learning networks. Retrieved December 6, 2018, from Learning and Teaching Repository: https://ltr.edu.au/resources/CG9_994_Lockyer_Report_2011.pdf
- Fassbinder, A., Barbosa, E. F., & Magoulas, G. D. (2017). owards an educational design pattern language for massive open online courses (MOOCs). 24th Conference on Pattern Languages of Programs (p. 1). The Hillside Group.
- Hansen, J. D., & Reich, J. (2015). Democratizing Education? Examining Access and Usage Patterns in Massive Open Online Courses. *Science* , 350 (6265), 1245-1248.
- Hew, F. K. (2016). Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCs. *British Journal of Educational Technology* , 47 (2), 320-341.
- Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning* , 15 (1), 133-160.
- Jordan, K. (2011). MOOC Project. Retrieved May 5, 2013, from www.katyjordan.com/MOOCProject.html

- Kizilec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355 (6322), 251-252.
- Ogata H., Majumdar R., Akçapınar G., Hasnaine N.H, and Flanagan B. (2018) Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning, In *Proceedings of 26th ICCE, Manila, Philippines, Nov 2018*
- Murthy, S., Warriem, J. M., Iyer, S., & Sahasrabudhe, S. (2018). LCM: A model for planning, designing and conducting Learner-Centric MOOCs. *International Conference on Technology for Education*. Chennai: IEEE.
- NPTEL. (2016). Retrieved Aug 11, 2018, from NPTEL Local Chapter: https://nptel.ac.in/LocalChapter/active_clg.php
- NPTEL. (2003). NPTEL Courseware. Retrieved December 6, 2018, from NPTEL: <https://nptel.ac.in/course.php>
- NPTEL. (2013). NPTEL Statistics. Retrieved Aug 11, 2018, from NPTEL: https://nptel.ac.in/statistics_mar13.php
- Rodríguez, M. F., Correa, J. H., Pérez-Sanagustín, M., Pertuze, J. A., & Alario-Hoyos, C. (2017). A MOOC-based flipped class: Lessons learned from the orchestration perspective. *European Conference on Massive Open Online Courses* (pp. 102-112). Cham: Springer.
- Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of Innovations: A cross-cultural approach*.
- Warriem, J. M. (2018). NPTEL Local Chapters: Facilitating Mainstreaming of MOOCs in Higher Education. *9th International Conference on Technology for Education* (p. in press). Chennai: IEEE.

Extracting Self-Direction Strategies and Representing Practices in GOAL System

Huiyong Li

Graduate School of Informatics
Kyoto University
lihuiyong123@gmail.com

Rwitajit Majumdar

Academic Center for Computing and
Media Studies, Kyoto University
majumdar.rwitajit.4a@kyoto-u.ac.jp

Yuan Yuan Yang

Graduate School of Informatics
Kyoto University
44.yangoo@gmail.com

Brendan Flanagan

Academic Center for Computing and Media Studies
Kyoto University
flanagan.brendanjohn.4n@kyoto-u.ac.jp

Hiroaki Ogata

Academic Center for Computing and Media Studies
Kyoto University
ogata.hiroaki.3e@kyoto-u.ac.jp

ABSTRACT: To enable students to engage in lifelong learning, it would require developing self-direction skills (SDS). The use of technology in education is common, however, the available software still offers poor support from a self-direction point of view. From the theories of self-directed learning (SDL) and evidence-based practice (EBP), we proposed a framework to track self-directed actions and represent strategies of practice by the learner. This framework is one of main components of the GOAL (Goal Oriented Active Learner) system. The GOAL system is built to support for acquisition of SDS in the context of learning and health. In this paper we describe how learner interactions in the GOAL system are captured as eXperience API statements and later visualized to enable learners reflect on their strategies.

Keywords: Self-direction skills, evidence-based practice, learning analytics, DAPER model, GOAL system

1 INTRODUCTION

With the growing trend of preparing students for lifelong learning, the theory of self-directed learning (SDL) has been increasingly applied in the context of higher education. Being self-directed would help students to prepare them for success in their future careers, and enables them to engage in lifelong learning. Since it's a cognitively and behaviorally complex task during executing SDL, the ongoing diagnosis of learners in underdeveloped skills and instructional design of environment are essential.

We developed the GOAL (Goal Oriented Active Learner) system, where learner engage with their own data from learning and physical activities context to foster their skills of being self-directed (Majumdar et al., 2018). The idea is to support students for acquisition of self-direction skills (SDS) through everyday activities. Since the learning logs and health records could be automatically integrated into our support system, students are given more opportunities to engage in self-direction.

In this paper, we propose a framework to address the challenge of tracking self-direction practices of the learners. We capture the student actions as eXperience API statements. Utilizing those action statements, first we extract strategies of self-directedness. Then students' self-directedness practices could be represented in a simple format to support self-assessment and self-reflection.

2 RELATED WORK

2.1 Self-Direction Models

SDL is primarily studied in the context of adult education and covers the following processes: learning needs or learning motivation, learning resources, learning goals, learning plans and activities, learning evaluation, and communication skills.

Three main models have been proposed to study SDL: Candy's four-dimensional model (Candy, 1991), Brockett and Hiemstra's personal responsibility orientation model (Brockett & Hiemstra, 1991) and Garrison's three-dimensional model (Garrison, 1997). Candy (1991) concluded that SDL encompasses four dimensions: personal autonomy, self-management, learner-control, and autodidaxy. Brockett and Hiemstra (1991) provided a rationale for two primary orientations in developing an understanding of SDL: process and goal. Garrison's model of SDL includes three dimensions interacting with each other: self-management, self-monitoring, and motivation.

For our work, we proposed a process model, DAPER (Majumdar et al., 2018) which synthesizes the SDL model for data driven activities. The initial phase of data collection which gives learners the initiative, followed by four key phases (data analysis, goal setting and planning, executing monitoring, reflection). Section 3.1 presents the details on those five phases of the model.

2.2 Measuring Self-Direction

Mostly in the context of learning, learners rely on their own memory and notes to define their goals and plans, and then monitor and evaluate their own progress and performance. The researchers commonly assess learners' SDS using self-reported questionnaires, like PRO-SDLS (Stockdale & Brockett, 2011), SRSSDL (Williamson, 2007) or SDLI (Cheng et al., 2010). While these instruments provide a picture of each learner's skills at a certain moment in time, they do not continuously track learner's skills. Also, these instruments are intrusive and time consuming.

However, the assessments could be supported through tracking interactions with software, especially in online learning environment (Li et al., 2018). The key interactions related to metacognition of self-direction should be extracted, like goal setting, planning, reflection, etc. Moreover, since a wide variety of self-direction interactions could be recognized, the definition of self-direction actions and strategies should be identified.

2.3 Evidence-Based Practice

In epistemology, evidence is that which serves to confirm or disconfirm a hypothesis (claim, belief, theory; Achinstein, 2001). It can perform a support function, including all sorts of data, facts, and personal experiences. Evidence-based practice (EBP) involves the use of the best available evidence to bring about desirable outcomes, or conversely, to prevent undesirable outcomes (Kvernbekk,

2016). Moving toward more EBP has the potential to improve the quality of learning, especially the acquisition of SDS.

Because of the complexity in the self-direction cycle, more high level data need to be provided for learners. The learners need reliable, revealing and relevant data that support decision-making. To support it, the five phases of DAPER model, activity model, strategies extraction, and practice representation are described in the following section. Previous studies of self-direction and self-regulation has highlighted learner agency regarding how they learn and the superiority of autonomous motivation for learning (Stockdale & Brockett, 2011; Greene & Azevedo, 2007). We follow that paradigm and let students choose their own goal and direct their own plan.

3 MODELING PROCESS AND ACTIVITY OF SDS

3.1 DAPER Model

DAPER model is a five-phase process model to conceptualize data driven self-direction skill execution and acquisition (Majumdar et al., 2018). Figure 1 shows the DAPER model and its five phases.

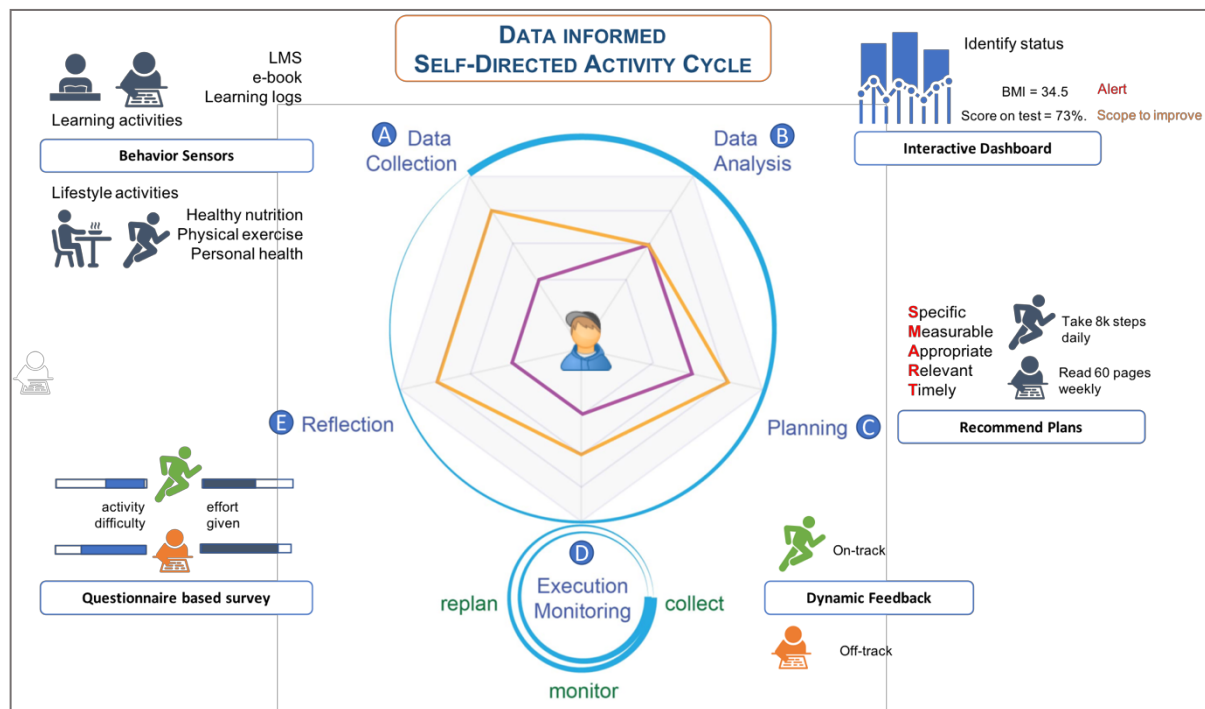


Figure 1: DAPER model and its five phases of SDS execution and acquisition (Majumdar et.al. 2018)

A. Data Collection. Most of activity data will be automatically collected from activity sensors. For the health activities, the raw data is from smartphones and wearable devices. For the learning activities, the raw data is collected from an e-book system BookRoll and an e-learning system Moodle. Also, learners can add their records manually if the records were not automatically collected, revise their records and delete certain records.

B. Data Analysis. The trend of one activity will be showed by chart graph, and the average value of one activity will be compared with the standard level or the average value of group. Based on the trend and compared results, the status of each activity will be easily identified.

C. Goal setting and Planning. After being aware of their status of each activity, learner set goals regarding any activities whose data was analyzed. The goals could be specific with a target value and expected date, or not specific just with a description. Under one goal, multiple plans could be created. The plans are with different frequencies, target values and duration.

D. Execution monitoring. The progress of each plan will be shown by chart graph since the activity data will be continuously collected and be compared with the target value. For example, in the health scenario, learners may monitor their heart rate during a specific physical exercise. In the learning scenario, learners might monitor the completion of their course content before an upcoming assessment. This phase often includes multiple cycles of other phases, including data collection, analysis, re-planning, reflection.

E. Reflection. During the process of self-directed, learners could write daily reflection journal for their goals or plans with self-rated items and notes. The self-rated items include the evaluation of task performance and their efforts given for the chosen task. The note form is a single text field which is organized by learners. The information in the notes could be current problems, specific strategies, or further actions.

GOAL system is based on the described DAPER model. Learners can build their personal goals and continuously improve them in the context of learning and health. The phases of DAPER model are weakly sequenced so currently the learner can openly navigate in the GOAL system and access functions of any phase.

3.2 Activity Model

Activity model provides a context of self-direction in the Goal system. It has two elements: *Activity* and *Milestone*.

The *Activity* is learning logs or health records automatically collected from activity sensors, such as smartphones. Learning logs are tracked by the e-book and e-learning system. They contain digitized reading logs, status of course assignments, and answers of quizzes. The health records are collected through Apple Health application or Google Fit platform. They include steps taken, runs, walks, workouts, biking, sleep, weight, heart rate, and food. For example, an *Activity* could be reading 50 pages or running 3 kilometers.

The *Milestone* is an accumulated value from the *Activity*. It's as an indicator of the activity achievement. A *Milestone* could be the first try, completed 25%, completed 50%, or completed 100%.

4 FRAMEWORK FOR EXTRACTING AND REPRESENTING SELF-DIRECTED PRACTICES

First, we extract strategies from interactions between learners and the GOAL system and activity logs. The definition of strategies is from the five phases of DAPER model. Next, we integrate these strategies into practices of self-direction and represent the practices to support self-assessment and self-evaluation for each individual users.

4.1 Extracting Strategies

Following xAPI structure we define strategies of self-direction. Table 1 shows a list of definition of self-direction strategies. The self-direction strategies are from five phases of DAPER model. They consist of activity log management, activity log analysis, goal management, planning, self-monitoring, and self-evaluation. Each strategy includes multiple actions. An action is defined by the verbs and the objects in the GOAL system. For instance, *John created a plan "Running at weekdays"* is an action which contains a verb, *created*, and an object, *a plan "Running at weekdays"*.

Table 1: Definition of self-direction strategies

<i>DAPER Phase</i>	<i>Strategy</i>	<i>Verb</i>	<i>Object</i>	<i>Example</i>
Data collection	Activity log management	<i>added</i> <i>edited</i> <i>deleted</i>	activity log	John added an e-book reading log
Data analysis	Activity log analysis	<i>checked</i>	activity log	John checked the activity "Running"
Goal setting and planning	Goal management	<i>created</i> <i>edited</i> <i>deleted</i> <i>achieved</i> <i>discarded</i>	goal	John edited the goal "Get A+ Grade" with a new description "Complete all reports"
	Planning	<i>created</i> <i>edited</i> <i>deleted</i>	plan	John created a plan "Running at weekdays"
Executing monitoring	self-monitoring	<i>checked</i>	plan	John checked "Plan 3" at 2:00 pm
Reflection	self-evaluation	<i>noted</i> <i>scored</i>	goal plan	John scored the effort to the plan "Running at weekdays" with "Much"

4.2 Utilizing Strategies to Represent Practice

After extracting self-direction strategies, the practice will be generated and represented for learners. It's a key component to support decision-making when learners reflect their practices or identify obstacles.

The components of practice are *Activity*, *Milestone*, *Decision* and *Achievement*. The *Activity* and *Milestone* are from the activity model. The *Decision* means key interactions between learners and

the GOAL system. It's related to manage goals and plans: created a goal, edited a goal, deleted a goal, created a plan, edited a plan, deleted a plan, noted a goal, scored a goal, noted a plan, scored a plan. The *Achievement* means that a goal has been achieved or discarded.

We chose a tree and a timeline structure to represent practices. An example of practice representation with an editable tree and a visual timeline is given in Figure 3. The tree of practice has three columns: activity & action, date and description. The activity & action column contains *Decision*, *Milestone* and *Activity*. As noted before, the name of *Decision* is generated from the action between learners and the GOAL system. The default descriptions are from inputs when learners manage goals or plans. For instance, the description of "Created a goal" is the input description of the new goal, the description of "Edited a plan" is the target value and frequency value of the updated plan. Moreover, each branch of practice tree could be edited by learners.

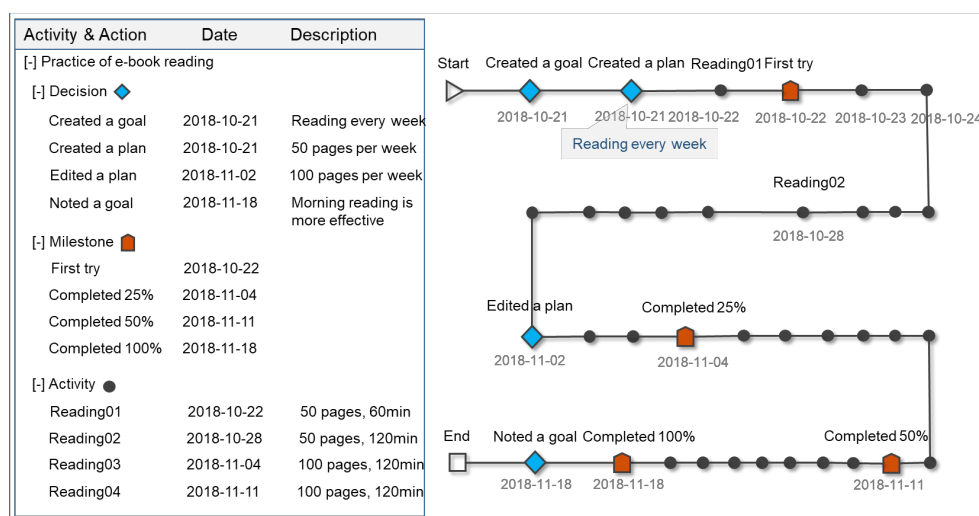


Figure 3: An example of practice representation with an editable tree and a visual timeline

We also generated a timeline to represent practices. The timeline of practice is from the left tree data but with a user-friendly visual format. It also contains *Decision*, *Milestone* and *Activity*, which are shown with blue diamond icons, red arrow icons and black dot icons, respectively. It also has start and end date of one goal. It will be generated when one goal was achieved or discarded. The description of one element will be shown when the learner tries to click it. For example, learner will see "Reading every week" when click the *Decision* element, *Created a plan*.

Thus, our developed framework (shown in Figure 2) contains two steps: self-direction strategy extraction and self-direction practice representation. The basic structure of strategy combines information from the DAPER model and the activity model. Practice is represented for each individual based on their own activity trace data and GOAL system interaction data.

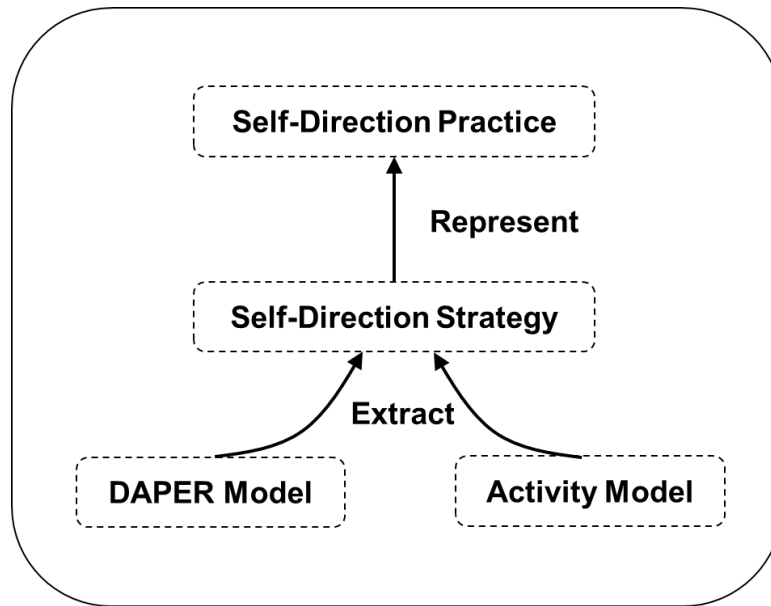


Figure 2: Framework for extracting and representing self-directed practices

5 DISCUSSION

5.1 Measurement of Self-Direction

What information should be extracted during self-direction process? In our work we use two kinds of trace data to answer the question: activity trace data, from the activity logs in the context of self-direction and interaction trace data, from the interaction logs in the GOAL system.

Since the rapid development of smartphones and wearable devices, tracking fine-grained, time-stamped data from learning and health activities is more common (Ogata et al., 2017; Case et al., 2015; Hekler et al., 2015). In contrast to self-report data, trace data is immediately collected within actual environment and could not be degraded the accuracy and completeness of learners' recall, perceptions and interpretations about how they learn.

The versatility and openness of xAPI makes us to define a wide and comprehensive selection of self-direction actions, directly related to the selection of the most relevant self-direction strategies (Manso-Vázquez et al., 2018). We start a simple definition of strategies from actions and activities since it could be part of complex strategies. For instance, a complex strategy called goal-oriented planning, could be formed by two simple strategies: goal management, planning. The simple definition is suitable to represent the complexity of self-direction strategies.

Combine activity trace data and interaction trace data, not only activity status but also strategy selection could be measured. These simple but important activities and interactions can be the foundation for learning analytics and evidence-based analytics in the context of data-driven self-directed activities.

5.2 Feedback for Supporting SDS Development

What information should be presented to support SDS development? We proposed a practice-based feedback to facilitate the selection of strategies.

Feedback is a powerful influence on learning, especially on SDL. It's not easy for novice SDL learners to select, monitor and evaluate their strategies independently. We offer learners feedback with practice trees and practice timelines. The *Decision* and *Milestone* on them are strategic level information, which contain a format of knowledge. Other format of feedback could also be considered based on self-direction strategies, such as strategies time distribution, strategies preference with a radar graph and so on.

6 CONCLUSION

This paper proposed a framework which could extract self-direction strategies and represent practices with editable trees and visual timelines. The actions and activities of self-direction process are captured to the strategies as eXperience API statements and then those strategies are presented with practice information. The framework is built on the DAPER model with five phases of self-direction process. The activity data and interaction data are tracked and therefore important activities and interactions related to strategies could be represented, like goal management, planning, self-evaluation. The framework provides reliable, revealing and relevant data and practice-based representation that support making valid inferences, which is essential for acquiring and promoting SDS.

ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304 and JSPS KAKENHI Research Activity Start-up Grant Number 18H05746.

REFERENCES

- Achinstein, P. (2001). *The book of evidence*. Oxford: Oxford University Press.
- Bjorklund, D. F., Miller, P. H., Coyle, T. R., & Slawinski, J. L. (1997). Instructing children to use memory strategies: Evidence of utilization deficiencies in memory training studies. *Developmental Review*, 17(4), 411-441.
- Brockett, R. G., & Hiemstra, R. (2018). *Self-direction in adult learning: Perspectives on theory, research and practice*. New York: Routledge.
- Candy, P. C. (1991). *Self-Direction for Lifelong Learning. A Comprehensive Guide to Theory and Practice*. Jossey-Bass, 350 Sansome Street, San Francisco, CA 94104-1310.
- Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama*, 313(6), 625-626.
- Cheng, S. F., Kuo, C. L., Lin, K. C., & Lee-Hsieh, J. (2010). Development and preliminary testing of a self-rating instrument to measure self-directed learning ability of nursing students. *International journal of nursing studies*, 47(9), 1152-1158.
- Davies, P. (1999). What is evidence-based education?. *British journal of educational studies*, 47(2), 108-121.

- Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. *Adult education quarterly*, 48(1), 18-33.
- Greene, J. A., & Azvedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, 77(3), 54–372.
- Hadwin, A., Oshige, M., Miller, M., & Wild, P. (2009, July). Examining student and instructor task perceptions in a complex engineering design task. In *international conference on innovation and practices in engineering design and engineering education*. McMaster University, Hamilton, ON, Canada.
- Hekler, E. B., Buman, M. P., Grieco, L., Rosenberger, M., Winter, S. J., Haskell, W., & King, A. C. (2015). Validation of physical activity tracking via android smartphones compared to ActiGraph accelerometer: laboratory-based and free-living validation studies. *JMIR mHealth and uHealth*, 3(2), e36.
- Kvernbekk, T. (2016). *Evidence-based practice in education: Functions of evidence and causal presuppositions*. London: Routledge.
- Li, H., Flanagan, B., Konomi, S. I., & Ogata, H. (2018). Measuring Behaviors and Identifying Indicators of Self-Regulation in Computer-Assisted Language Learning Courses. *Research and Practice in Technology Enhanced Learning*, 13(1), 19.
- Majumdar R., Yang Y.Y., Li H., Akçapınar G., Flanagan B., & Ogata H. (2018) GOAL: A System to Support Learner's Acquisition of Self Direction Skills, *26th ICCE*, Manila, Philippines, Nov 2018
- Manso-Vázquez, M., Caeiro-Rodríguez, M., & Llamas-Nistal, M. (2018). An xAPI Application Profile to Monitor Self-Regulated Learning Strategies. *IEEE Access*, 6, 42467-42481.
- Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., & Hirokawa, S. (2017) Learning Analytics for E-Book- Based Educational Big Data in Higher Education. In *Smart Sensors at the IoT Frontier* (pp. 327-350). Springer, Cham.
- Oppezzo, M., & Schwartz, D. L. (2014). Give your ideas some legs: The positive effect of walking on creative thinking. *Journal of experimental psychology: learning, memory, and cognition*, 40(4), 1142.
- Saks, K., & Leijen, Ä. (2014). Distinguishing self-directed and self-regulated learning and measuring them in the e-learning context. *Procedia-Social and Behavioral Sciences*, 112, 190-198.
- Stockdale, S. L., & Brockett, R. G. (2011). Development of the PRO-SDLS: A measure of self-direction in learning based on the personal responsibility orientation model. *Adult Education Quarterly*, 61(2), 161-180.
- Williamson, S. N. (2007). Development of a self-rating scale of self-directed learning. *Nurse researcher*, 14(2).
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American educational research journal*, 45(1), 166–183.

Workshop on Educational Data Visualization

Nirmal Patel

Playpower Labs

nirmal@playpowerlabs.com

Derek Lomas

Delft University of Technology

j.d.lomas@tudelft.nl

Collin Sellman

Arizona State University

collin.sellman@asu.edu

ABSTRACT: The primary goal of this workshop is to produce open source data visualizations that help communicate results of learning analytics (LA) research to educators. Instructors are increasing their use of data to drive instruction, and various results of LA research are useful towards this end. However, the actionable insights discovered by the LA community are often inaccessible to educators due to their relative complexity. In this case, it is possible to use data visualization to communicate actionable insights about learners to optimize instruction effectively. Visualizations of learner data can make it easy for teachers and other education stakeholders to take evidence-based action. Organizers of the workshop intend to invite authors to describe and implement educational data visualizations that can aid decision making in online and offline classrooms. The workshop will result in a gallery of open source educational data visualizations that can be freely used by the LA community.

Keywords: Data Visualization, Data-Driven Instruction, Information Communication

1 INTRODUCTION

A recent survey in the United States found that 95% of the K-12 teachers use a combination of academic data and non-academic data to understand their students' performance. However, 34% of the surveyed teachers also reported that there was too much data for them to look at. How can we help educators make sense of large amounts of student data? Data visualization is one of the most widely used techniques that help people make sense of large amounts of numerical information. Graphical representations of data can be used very effectively to communicate context-specific information.

Reporting of learner data is one of the cornerstones of LA research, and the LA community has developed domain-specific data visualizations to show student learning in different contexts. Some of these visualizations emerged from the Learning Science research community (e.g., Learning Curves,) while other visualizations have a close affinity with classroom practice (e.g., Curriculum Pacing Plots.) Although these visualizations are slowly making their way into the hands of educators, many of these visualizations are not readily available for reproduction in the open source data

analysis environments such as R and Python. This workshop aims to produce an open source gallery of education data visualizations that are easily reusable by LA researchers and practitioners.

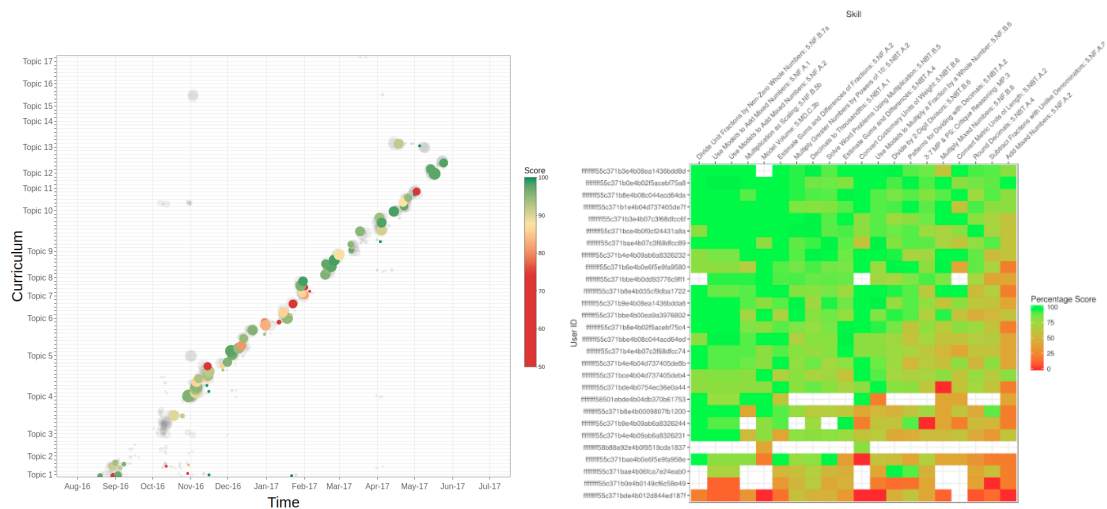


Figure 1a (left) and 1b (right): The figure on the left shows a Curriculum Pacing Plot, showing the progression of a classroom through a yearlong period. The figure on the right shows a Mastery Matrix for a classroom, which allows easy identification of struggling students and difficult topics.

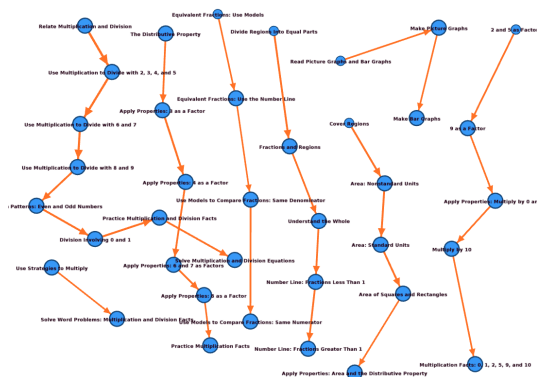


Figure 2: Learning Pathway Visualization, showing frequently taken learning paths taken by students in an online curriculum.

To make data visualizations a tool for knowledge discovery, they have to be made domain specific. A scatter plot, when adapted to a context of classroom, can become a Curriculum Pacing Plot (figure 1a,) a heatmap with student and skill data can be turned into a Mastery Matrix (figure 1b,) and a Learning Pathway Visualization with educational activities as nodes can display highways of student learning (figure 2.) All these visualizations are based upon commonly used graphical representations of data, but with a little tweaking, they turn into tools that let us extract meaningful information from educational data. Moreover, useful information visualizations can help support and engage a range of different stakeholders. For example, learner data visualizations can help curriculum coordinators in schools understand student behavior in a manner that can inform instruction. Learning Pathways of a MOOC can help instructional designer discover whether students are progressing through the curriculum as expected or not. Whereas many data mining techniques can produce “black boxes,” visualizations are human readable. This readability allows direct stakeholders

(e.g., teachers and instructional designers) to critique assumptions made by data scientists, who are often removed from the context of data production.

LA research has the potential to inform real-world instruction, but bringing research findings into real-world requires effective communication to stakeholders outside the LA community. This can be difficult because educators usually find results of learning analytics challenging to understand. We believe that graphics can help us bridge this communication gap and make the results of LA research more fruitful.

2 ORGANIZATION

The workshop will be organized as a half-day event, as this is the first data visualization workshop being organized in the community. The organizers will invite authors to develop and implement open source data visualizations and describe them in posters, short papers, or full papers. Posters will describe visualizations that are relevant to the community but are still being developed. Short papers will describe works that are sufficiently mature but haven't been tested in the field, and full papers will describe novel visualizations that are mature and have been tested with the stakeholders in the field. The participation of the workshop will be mixed, and delegates other than the authors will be invited to register and take benefit of the workshop. The workshop will contain a series of visualization demos, and each presenter will show how everyone in the audience can use the open source visualization. Authors will be highly encouraged to develop their visualizations in R and Python, and all of the visualization programs will be uploaded in a GitHub repository that will remain freely accessible to the LA community. We expect 15 to 30 participants to attend our workshop. No special equipment other than a projector will be required.

3 OBJECTIVES

The goal of this workshop is to adapt existing visualizations or product novel visualizations of educational data that can communicate actionable information to educators. We will invite authors to describe and implement educational data visualizations relevant to a range of educational contexts such as various types of digital learning environments such as MOOCs, virtual schools, K-12 and university classrooms, and other exploratory learning environments like as educational games. We will encourage authors to produce interactive visualizations so that the users can actively engage with the data and use the graphs as tools to explore data and understand students rather than a snapshot of data that they can look at and reflect on. As the open-source code is an emphasis of the workshop, authors will also be asked to write programs that are easy to use, e.g., providing functions that take well-defined data structures (e.g., data tables, graphs) as input and produce desired visualizations as output. Topics for the data visualizations will be:

- Student learning and behavior
- Student knowledge and mastery
- Student learning trajectories and processes
- Student misconceptions
- Problem-solving strategies
- Teaching strategies
- Collaborative learning

- Classroom learning
- Emotional states
- Clickstream data
- Student groups and their differences
- Comparison of observed and ideal student behavior
- Usage and efficacy of educational content
- Demystifying “black box” machine learning models

Any other topics that are relevant to educational data and environments will also be included. We hope that the visualizations produced during this workshop can act as worked examples of visual design patterns that can be applied to educational data from a range of different sources and serve as a quick reference guide for LA community.

REFERENCES

- Data Quality Campaign. (2018). *What Parents and Teachers Think About Educational Data*. Retrieved from <https://dataqualitycampaign.org/resource/what-parents-and-teachers-think-about-education-data/>
- Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, 43, 43-56.
- Patel, N., Sharma, A., Sellman, C., & Lomas, D. (2018). Curriculum Pacing: A New Approach to Discover Instructional Practices in Classrooms. In *International Conference on Intelligent Tutoring Systems* (pp. 345-351). Springer, Cham.

Implications of Instructor Analytics Use Patterns for the Design of Actionable Educational Data Visualizations

Alyssa Friend Wise
New York University
alyssa.wise@nyu.edu

Yeonji Jung
New York University
yeonji.jung@nyu.edu

ABSTRACT: This paper offers insights to inform evidence-based learning analytics design through the presentation of an empirically-derived model of instructor analytics use. The model represents key elements of the ways in which instructors make sense of and respond to the analytics data as part of their pedagogical decision-making process, which can assist educational visualization designers in choosing among the myriad data representations possible to produce interpretable and actionable learning analytics systems. Instructor analytics use is shown as a multi-phase process divided across the larger activities of sense-making and pedagogical response. Sense-making process moves from a general area of curiosity which instructors can develop into more specific questions through interaction with the data, to reading the data to identify noteworthy patterns and appraising the patterns' importance in the course. Pedagogical responses involve taking the form of actions (whole-class scaffolding, targeted scaffolding, and revising course design) or adopting a wait-and-see holding pattern, and/or deeply reflecting on pedagogy. Drawing on this model, specific recommendations are made for how learning analytics design can align information presentation with core instructional practices and embed features to support processes of use in order to be most impactful on teaching and learning.

Keywords: Learning analytics use, Data-informed decision-making, Teaching analytics, Learning analytics design, Educational data visualization

1 INTRODUCTION

While the process of using analytics data to inform pedagogical decisions is acknowledged to be complex (Herodotou et al., 2017), little is known about the details of how it occurs in authentic teaching contexts. Data-informed pedagogical decision-making process involves more than just instructors' uptake of learning analytics tools; rather, it entails instructors' translation of tool-provided information into locally-meaningful knowledge and subsequently use of it to guide their pedagogical actions (Molenaar & van Campen, 2018). Examining such information use is critical to impact educational practice and inform the design of learning analytics in interpretable and actionable visualization. This paper fills a gap in the information available to educational visualization designers to make evidence-based design decisions by presenting an empirically-derived model of instructor analytics use to guide the design and implementation for learning analytics.

2 A BRIEF OVERVIEW OF THE MODEL DEVELOPMENT

The model was developed based on empirical case studies conducted with five university instructors who used a learning analytics dashboard in their teaching during the course of a semester. In-depth interviews were guided by the (limited) existing literature on the topic (Molenaar & van Campen, 2018; van Leeuwen, van Wermeskerken, Erkens, & Rummel, 2017; Verbert et al., 2013) and involved instructors' think-aloud walk-through of their analytics use, showing the relevant visualizations to concretize their responses. The dashboard used by these instructors was developed and rolled out by the university's Instructional Technology team based on consultations with each instructor about the kinds of student activity and performance information they would like to see for one of their courses. The personalized dashboard involved three to four distinct views (e.g. student access of course site and resources, video viewership, online quiz results or survey responses), each which acted as an independent overview into the data (see Figure 1 and 2). An inductive qualitative analysis was conducted on the interview transcripts and surfaced twenty emergent themes related to instructors' practices of analytics use. The themes were then synthesized into a situated model of instructor analytics use that is presented in the next section. The full study, including details and evidence for each theme and the model development process, is reported in Wise and Jung (in review).

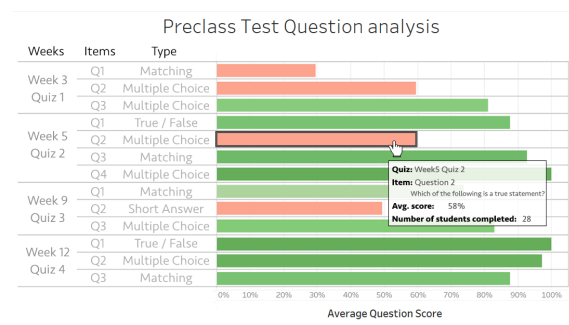
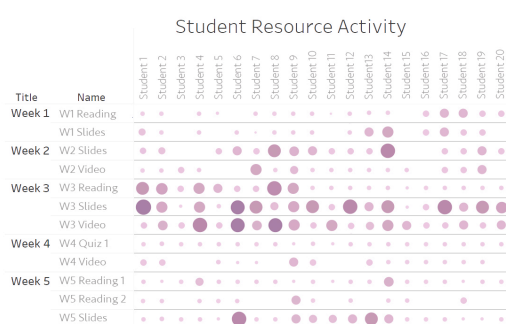


Figure 1. Example Dashboard View of Resource Access Figure 2. Example Dashboard View of Quiz Results

3 A SITUATED MODEL OF INSTRUCTOR ANALYTICS USE

The model consists of multiple phases of activity embedded in a two-part structure with sense-making and pedagogical response as distinct aspects of practices (see Figure 3). Such structure aligns with the majority of prior studies which have described instructor analytics use as first determining an understanding of what the information tells about the course and then considering potential actions in response to it (Herodotou et al., 2017; van Leeuwen et al., 2015).

Looking inside each part of the process, sense-making begins with an instructor's general area of curiosity (e.g. class-level or individual-level engagement, usefulness of course materials). While prior studies suggest that instructors can come to analytics with fully-formed questions (Dyckhoff et al., 2012) either based on prior analytics use or their own methods of data collection and analysis, or that they may just respond to the data as presented (Herodotou et al., 2017), this model offers a third possible path: that instructors may start their analytics use with a general area of curiosity. Areas of curiosity can develop into more specific questions through interaction with the data, which

can then be answered with more careful examination (e.g. identifying potential relationships in the data that instructors hope to explore further) (Molenaar & van Campen, 2018).

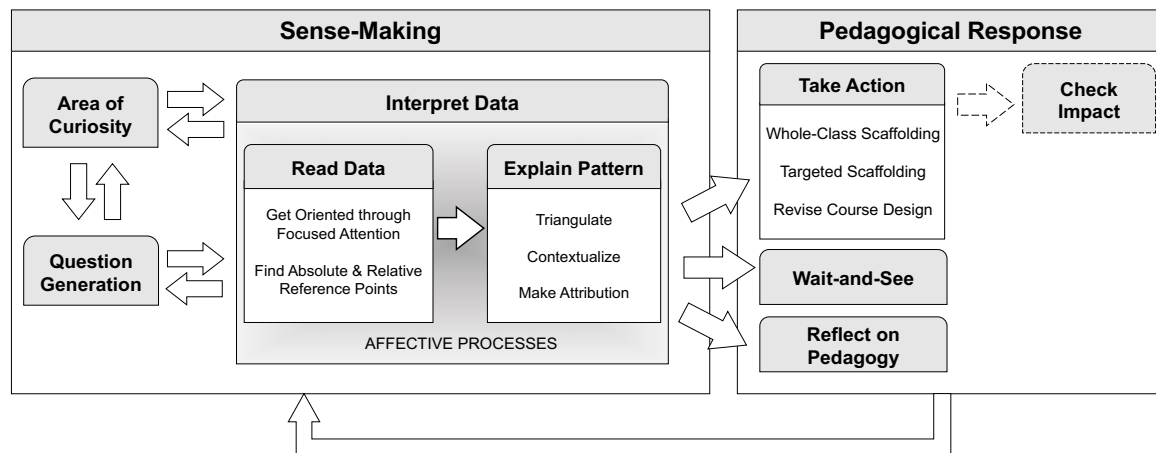


Figure 3. A Situated Model of Instructor Analytics Use

Data interpretation occurs through two distinct and equally important phases of activity: reading the data to identify meaningful patterns and then generating explanations that address the patterns' importance for the course. In reading the data, instructors get oriented to visualizations through paying attention to a specific piece of information or a noticeable pattern as initial anchors, and then expanding their view outwards to explore the different kinds of information offered by the overall analytics. This differs from previous studies which suggested a sequence of first getting oriented, then applying focused attention to specific data (Van Leeuwen et al., 2017). Such difference may depend on how the analytics view is organized as certain kinds of layout design and information arrangement can allow instructors to get oriented first; thus both activity sequences should be considered as possibilities. In reading the data, there is a need for reference points of comparison which instructors can feel confused about what to use as; either absolute (e.g. do students engage with at least 75% of the provided materials) or relative (e.g. does student engagement change during the course, how does this year's pacing trajectory compare to last year's). Consideration of how to best provide explicit reference points to aid users in navigating through data is a growing area of attention in the literature (for example see Patel, Sharma, Sellman, & Lomas, 2018). In explaining patterns, instructors extend the meaning of patterns identified by explaining (or questioning) their implications as related to their course. Instructors often try to triangulate the patterns with additional information (e.g. class observation) to confirm their interpretation. When this supports the interpretations, instructors may use their contextual knowledge of the course and students to explain what the results might mean and make attributions to potential causes (Molenaar & van Campen, 2018). When the external information and analytics data do not align, it can lead instructors to question the analytics (Dazo, Stepanek, Chauhan, & Dorn, 2017) and/or hesitate to take action (Herodotou et al., 2017). In addition to cognitive processing of patterns, data interpretation can provoke affective responses such as surprise, disappointment, or joy as reported in the literature (Wise, Zhao, & Hausknecht, 2014).

Following sense-making, pedagogical response is instructors' decisions and/or changes in thinking based on the analytics. The most common response type is taking action towards (1) the whole-

class, (2) particular students, or (3) course materials (e.g. Herodotou et al., 2017; Molenaar & van Campen, 2018; Xhakaj, Aleven, & McLaren, 2017). Checking whether the actions taken have achieved the intended impact is a final phase to close the loop; however this does not always occur (Dazo et al., 2017). Another common response to analytics is to adopt a holding pattern of waiting to see what will happen as more data is made available (Herodotou et al., 2017). Deep reflection and shifts in how an instructor conceptualizes their pedagogy is a new, interesting response type that has not received much attention in the literature (c.f. Molenaar & van Campen, 2018) but may have greater and longer-lasting effects than simple adjustments to teaching.

In addition to the unidirectional path through the different phases described above, instructor analytics use may also occur iteratively within and across each of the two larger parts (e.g. access to data leads to new questions; actions taken to test an initial interpretation influence back to the interpretation itself).

4 IMPLICATIONS FOR LEARNING ANALYTICS DESIGN

This situated model offers a clear starting place for efforts to design learning analytics to support instructors' pedagogical decision-making practices, which can guide designers in thinking ahead to instructors' analytics use during the design process (Xhakaj et al., 2017). In making evidence-based design decisions, it is critical to work directly with educational stakeholders throughout the development process. This process, however, requires more than just asking stakeholders what information they would like to look at and use, since it is a difficult question to answer in the absence of prior experience working with such data. More details should be considered in this process, including how gaining access to the data contributes to shifts in instructors' understanding or new question generated to the data. This highlights the need for evidence-based design to be attentive to actionability where "analytics connect with education and the changes that administrators, teachers and students want the tools to make in order to support their everyday learning, teaching and assessment work" (Ferguson et al., 2016, p. 9). The core areas for attention are presented in the following sets of design recommendations based on the situated model.

4.1 Learning analytics design should align information structures with instructors' pedagogical concerns.

Organize information from the perspective of instructors, not data structures. In analytics development, it is easy (and often necessary) to start thinking about the data in the form in which it is made available (e.g. organized alphabetically, by type of interaction, or by system-time). But instructors often think in different categories: weeks or units of a class, sets of associated course activities. This disconnection creates a critical barrier for effective analytics use, but can be addressed by explicitly eliciting instructor conceptualizations of how they think about their course and the different elements that compose it (which is a quite different kind of question to ask instructors than what kinds of information they would like to know). Attention to this issue may also raise the need for (re)considering learning design before analytics are built so that the two can be in alignment (Lockyer et al., 2013).

Align the timing of system and instructors' practices. Similar to the prior consideration on the organization of information, the timing of access to information needs to be considered from the

perspective of the instructor. The deferred update of data refresh can limit the usefulness of the analytics for instructors who want to access the dashboard immediately prior to a class as reported in Wise and Jung (in review). While constant data updating for the entire system may not be realistic, allowing instructors to update their data on demand in situations of need could be one way to address the situation.

4.2 Learning analytics should be designed to support processes of use.

Embed support for question generation and maintenance. A key element of the value proposition for instructor analytics use is that analytics data responds to important questions on which instructors can take action. As an iterative process within and across each of the two larger parts of the model, questions often emerge or are refined through further examination of the data. Importantly, amidst hectic teaching schedules such questions may not be formed or remembered across analytics use sessions. Analytics tools can be designed to support this process by including features which support the generation and maintenance of questions (and perhaps answers). For example, a question area associated with a visualization could offer a set of editable, tailorable questions with add, delete and edit functions that provide flexibility of use. Questions can be maintained across sessions of use and annotated with answers instructors find in the data or tags for future follow-up.

Incorporate visual aids to find entry points to the analytics. Another important consideration is how to facilitate instructors to find entry points with which they can get oriented to the analytics. Rather than beginning with an overview and then digging in a particular part, instructors may often begin with some part of the analytics that they can make sense of and expand it outwards (Wise & Jung, in review). Visual aids can be incorporated into analytics to support this process; for example, in a large matrix of data toggleable tools which highlight information by rows or columns could help users focus their attention on finding certain kinds of patterns or extended sequences of visual information which can allow grouping to let users attend to individual weeks or quizzes in turn.

Help instructors to find and work with reference points for data interpretation. A further consideration to support instructors in making use of analytics is offering support for finding reference points with which to make sense of the data. Providing access to similar data from prior terms or overarching trends from similar courses or tools for making comparisons across time can provide high level relative reference points. Absolute reference points may be explicitly elicited through a process of guided reflection through which instructors articulate their expectations for class activity, engagement or performance in terms of the metrics available.

Embed flags for later decisions to take action and check impact. One of the pedagogical responses that instructors can commonly make is wait-and-see approach that delays action till further data is available. Effectiveness of this strategy requires the instructor to remember the situation they are waiting to know more about and return to reexamine it at some future date. However, it is quite possible that this never occurs. Rather than relying on instructor memory, analytics design can support this process by offering features that let instructors mark and/or annotate a pattern they observe in the data for future follow-up. Similarly, when action is taken, analytics features can be used to create externalized reminders to check the impact of the action.

4.3 Learning analytics should be designed to support sharing and conversations.

Help instructors to share the analytics by offering de-identified views. When instructors make sense of the analytics or take actions based on it, instructors may need or want to share analytics with other instructors to engage in a process of collaborative interpretation or with students as an object for discussion and reflection in the class. However, this practice may raise potential privacy concerns regarding the use of analytics-as-evidence. One way to facilitate the use of analytics as a mediational object (Wise, 2014) is to make it easy for instructors to switch to a view in which student identities have been removed or hidden.

5 IMPLICATIONS FOR LEARNING ANALYTICS IMPLEMENTATION

In addition to rethinking learning analytics design in the context of instructional practices, it is also important to consider ways in which analytics use can be facilitated through pedagogical support for the process of use itself. This is critical to facilitate instructors' translation of information into actionable insights which can feed back to the evidence-based design processes. One potential way is to consider ways to educate instructors in how to work with data to inform their teaching. This can be done upfront through structured instructor data training or in-situ with the introduction of a pedagogical analytics coach who creates a series of scaffolds to support instructors in this translation process. Analytics coach supports can take the form of periodic emails, one-on-one coaching sessions, department-based workshops and the cultivation of local instructor communities. For example, email messages can highlight particular pedagogical questions (e.g. how can I find and help students who seem unengaged in the first few weeks?), explaining how to use the dashboard to answer it (e.g. open the student-course interaction grid and look across the rows for consistently light colored cells), and then discussing actions that could be taken in response (e.g. speak with them individually to find out what is going on and make them aware that you are invested in their success, highlight for the whole class habits of successful students). In this way, pedagogically meaningful questions, answers, and actions are linked together to frame analytics use, rather than starting with a data-centric orientation. The same issues can be engaged with on a broader scale through one-on-one coaching or workshops in which instructors are supported in working through these sequences using data from their own classes. Collaborative interpretation with a sample data set can be also implemented to discuss common challenges in analytics use and disambiguate the meaning of analytics through dialogue with other participants. In the long term, local instructor communities of practice around analytics use can be cultivated where such contextualized, embedded, ongoing support networks are more effective than short-term information delivery (Darling-Hammond & Richardson, 2009). Put together, the sets of recommendations highlight the importance of establishing opportunities for future efforts in both learning analytics design and implementation based on empirical findings for supporting instructors' situated use of analytics.

6 CONCLUSION

This paper offers insights to inform evidence-based learning analytics design decisions through the presentation of an empirically-derived model of instructor analytics use. An understanding of the practices instructors engage in when using analytics in their teaching can guide educational visualization designers in choosing among the myriad data representations possible to produce interpretable and actionable learning analytics systems. Future work can investigate the impact of

the decisions taken to validate or refine the recommendations made above. In addition, broader data collection on instructor analytics use including log-file records, experience sampling data, and classroom observations can further reveal how analytics use occurs and feeds back into instructors' teaching practices. Together these efforts help strengthen the lines of communication between stakeholders and designers, and help us move as a field towards evidence-based learning analytics design as a practice to support teaching and learning, fostering educational success.

REFERENCES

- Darling-Hammond, L., & Richardson, N. (2009). Research review/teacher learning: What matters. *Educational leadership*, 66(5), 46-53.
- Dazo, S. L., Stepanek, N. R., Chauhan, A., & Dorn, B. (2017, May). Examining instructor use of learning analytics. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2504-2510). New York, NY: ACM.
- Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Journal of Educational Technology & Society*, 15(3), 58-76.
- Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittlemeier, J., Rienties, B., Ullman, T., & Vuorikari, R. (2016). Research evidence on the use of learning analytics: Implications for education policy. In R. Vuorikari, & J. Castano-Munoz (Eds.), *A European framework for action on learning analytics*. Luxembourg: Joint Research Centre Science for Policy Report.
- Herodotou, C., Rienties, B., Borooowa, A., Zdrahal, Z., Hlosta, M., & Naydenova, G. (2017, March). Implementing predictive learning analytics on a large scale: the teacher's perspective. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 267-271). New York, NY: ACM.
- Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, 57(10), 1439-1459.
- Molenaar, I., & van Campen, C. (2018). How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*.
- Patel, N., Sharma, A., Sellman, C., & Lomas, D. (2018, June). Curriculum Pacing: A New Approach to Discover Instructional Practices in Classrooms. In *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 345-351). Cham, Switzerland: Springer.
- van Leeuwen, A., Janssen, J., Erkens, G., & Brekelmans, M. (2015). Teacher regulation of cognitive activities during student collaboration: Effects of learning analytics. *Computers & Education*, 90, 80-94.
- van Leeuwen, A., van Wermeskerken, M., Erkens, G., & Rummel, N. (2017). Measuring teacher sense making strategies of learning analytics: a case study. *Learning: Research and Practice*, 3(1), 42-58.
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10), 1500-1509.
- Wise, A. F. (2014, March). Designing pedagogical interventions to support student use of learning analytics. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 203-211). New York, NY: ACM.
- Wise, A. F., & Jung, Y. (in review). Teaching with analytics: Towards a situated model of instructional decision-making.

- Wise, A. F., Zhao, Y., & Hausknecht, S. (2014). Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics*, 1(2), 48-71.
- Khakaj, F., Aleven, V., & McLaren, B. M. (2017, September). Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In *Proceedings of the 2017 International Conference on Artificial Intelligence in Education* (pp. 315-329). Cham, Switzerland: Springer.

Visualizing the Solution Space of Educational Games using TRESTLE

Erik Harpstead

Carnegie Mellon University
harpstead@cmu.edu

Christopher J. MacLellan

Soar Technology, Inc.
chris.maclellan@soartech.com

ABSTRACT: When designing open-ended educational games and other creative instructional environments it is important for designers to understand what learners can do within the space their games afford and whether behaviors across that space are supporting their instructional goals. In this demo we will present several prototype visualization concepts based on the TRESTLE concept formation algorithm to organize data from student solutions into a tree structure amenable to several kinds of visualization.

Keywords: Visualization, concept formation, alignment, solution space

1 INTRODUCTION

Exploratory data analysis is an important step within the educational data mining process, particularly so in the context of educational games, which generate large amounts player data. Being able to visually inspect trends in data can provide context and perspective on complex statistical analyses and can help guide educational technology design. Unfortunately, hand conducting such analyses on larger data sets is often too unwieldy and time consuming to be practical. To overcome this barrier, we developed the TRESTLE algorithm and an accompanying set of visualizations to help designers and researchers hierarchically organize and explore structured data, such as the kind generated by educational games and other open-ended, creative instructional environments.

Understanding the breadth of approaches that learners can take in these environments and, more importantly, how the game reacts to those approaches, is essential for ensuring effective instruction. In this paper we briefly describe the TRESTLE approach and describe two examples of how it can support the organization, exploration, and interpretation of structured educational game data.

2 TRESTLE

TRESTLE is a concept formation algorithm that incrementally learns conceptual hierarchies given structured examples as training data (MacLellan, Harpstead, Aleven, & Koedinger, 2016). Unlike most learning systems that only support a vector of feature values, TRESTLE supports hierarchical attribute-value lists (represented as Python dictionaries) that contain both nominal (e.g., discrete object types) and numeric (e.g., x and y position) attributes as well as attributes that refer to nested attribute-value lists, which we refer to as structural attributes (e.g., "block1" might have a nested set

of attribute-values that describes its location and type). It also supports relational attributes that can describe relationships between other attributes, such as specifying that "block1" is on top of "block2", e.g., `on(block1, block2)`¹. The variety of attribute types that TRESTLE can handle makes it broadly applicable to wide range of potential data sets.

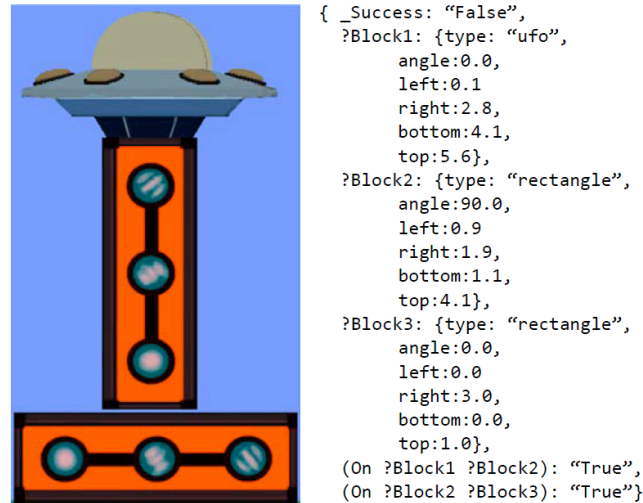


Figure 1. An example game state from *RumbleBlocks* and how it would be described to TRESTLE.

Given structured examples described in this representation, TRESTLE can engage in both supervised and unsupervised learning or a combination of the two. Specifically, the system can learn a shared hierarchical organization of both labeled and unlabeled data that enables learning from one kind of data to benefit the other kind. Learning within TRESTLE is incremental, meaning that it is presented with a sequence of examples. Upon receiving each example, TRESTLE sorts each new example into its hierarchy, updating it to reflect the new training data. To guide this learning process TRESTLE uses an objective function called category utility, which is derived from psychological studies of human concept formation (Fisher, 1987). This objective function is similar to the information-gain metric used in decision tree learning but supports the ability to predict arbitrary attributes of examples.

The hierarchical knowledge structure learned by TRESTLE supports two key capabilities: prediction and clustering. Prediction within TRESTLE operates similarly to learning. The system accepts as input examples with some of their attribute values missing. The system sorts these partial examples into its current organization using the available features and the resulting cluster it is sorted into is used to predict the values of any missing attributes. In this regard, TRESTLE is similar to other instance-based learning approaches, such as k-nearest neighbor, but it automatically determines—based on the data—how many examples (the k) to use for prediction. Previous work suggests that TRESTLE's prediction performance is similar to humans on the task of labeling the stability of block structures generated by students in an educational game (MacLellan et al., 2016).

¹ Additional specifics on the semantics of this instance representation are available from TRESTLE's online documentation: https://concept-formation.readthedocs.io/en/latest/instance_representation.html

In addition to prediction, TRESTLE can cluster examples both hierarchically and into flat clusters (e.g., into k groups). To generate clusterings, users present TRESTLE with a sequence of complete or partial examples (labeled or unlabeled), which it organizes into a conceptual hierarchy using its learning mechanisms. The resulting hierarchy can be directly returned as a clustering of the data. Additionally, TRESTLE has multiple post-processing routines that can translate these hierarchies into flat clusterings of the examples. For example, it can start at the root of the hierarchy (which contains all examples) and progressively break this top-level clustering into progressively smaller and smaller groups, stopping when it has optimized one of a range of metrics, such as Category Utility, AIC, or BIC. Our analysis of TRESTLE's ability to cluster block structures suggests that it produces groupings of examples that have reasonably high agreement with human-generated clusterings of the same blocks structures (MacLellan et al., 2018), which suggests that TRESTLE might be used to organize large volumes of examples in a way that aligns with how humans would organize the same data.

3 VISUALIZATION USE CASES

We have designed several visualizations to facilitate interpretation of TRESTLE's outputs, though few have been empirically validated with target users. These visualizations have been built to be interactive using D3.js (Bostock, Ogievetsky, & Heer, 2011) in order to enable an analyst to explore the information and make their own reflective judgements about their design.

Each of the visualizations presented here was developed as part of the analysis of *RumbleBlocks* (Christel et al., 2012) an educational game designed to teach based concepts of structural stability and balance to young children. In this game players build block towers that must survive an earthquake to succeed. Given that the game relied on a real time physics engine to generate its feedback it was not always clear if the game was providing clear guidance to players that would help them learn its targeted concepts.

3.1 TRESTLE Tree Visualization

The core visualization of TRESTLE is a visual representation of its hierarchical concept tree. Figure 1 shows this visualization of examples of player solutions to one level in *RumbleBlocks*. In this visualization each circle represents a collection of student solutions to a game level. The leaf concepts (filled in circles) represent specific instances within the dataset while the transparent enclosing circles represent higher-order clusters containing subgroups. The size of each cluster represents how many instances are grouped within it. To the right of the tree is a control panel showing various options as well as an Attribute-Value Table that shows the distributions of each attribute-value within the tree. When a concept or instance is clicked on, the view zooms to focus on that concept and the Attribute-Value Table updates shows the distribution of properties within the selected concept/instance.

Nodes in the visualization can be colored according to their different attributes to highlight trends within the concept tree. In main example in Figure 1, the nodes are colored based on whether a solution succeeded or failed the level in question according to the game's log data. Solutions are colored yellow if they are more likely to pass the level, and purple if they are more likely to fail. In this case there is a clear successful cluster (bottom of left branch), and a mostly clear negative cluster (right branch). The outcomes of the other two main branches of the tree (left and right of the

larger left branch) are less clear and would potentially warrant further investigation. An analyst can re-color the same visualization according to different properties of solutions to see if there are any correlations in trends that might warrant investigation as design issues. The smaller examples in Figure 1 are each colored according to a different physical property of the solution tower (e.g., symmetry, base width, center of mass height). In principle these properties should roughly correspond with succeeding at the level but there is not a clear correspondence to the coloring based on success, suggesting there might be an issue with how feedback is assigned in the game.

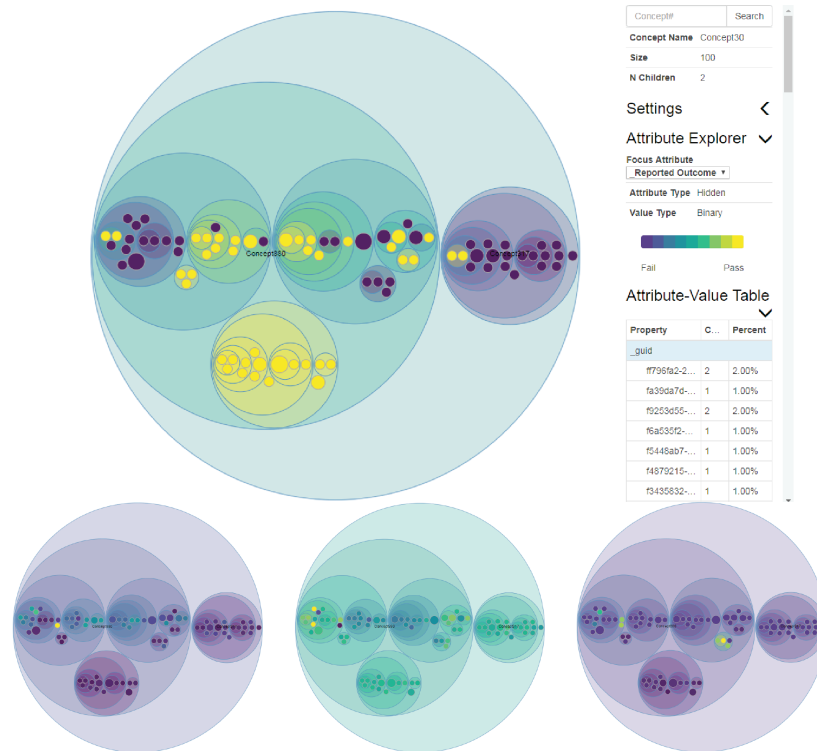


Figure 2. A visualized TRESTLE tree of a sample of 100 solutions to a level in RumbleBlocks. In the top visualization clusters are colored by their likelihood of succeeding on the level (yellow for passing, purple for failing). In the lower three visualizations the same tree is re-colored according to different physical properties of game solutions.

3.2 TRESTLE Alignment Visualization

While the hierarchical tree is currently the default output for visualization in TRESTLE, we have also developed more advanced forms of visualization that make use of the TRESTLE data structure in a flattened form. The alignment visualization shown in Figure 3 is based on analyses we have done of the solution space of *RumbleBlocks* (Harpstead, MacLellan, Aleven, & Myers, 2014). This visualization breaks up the TRESTLE tree into representative clusters that can be plotted according to their properties to look for correlations in data that might be indicative of problems. In the case shown in Figure 3 a principle relevant metric (in this case the symmetry of the tower) should be predictive of successful performance in the game. Within the visualization this is denoted by the red and green shaded regions where it would be desirable for solution clusters to fall in the green areas while it would be a problem if they mainly fell into red regions.

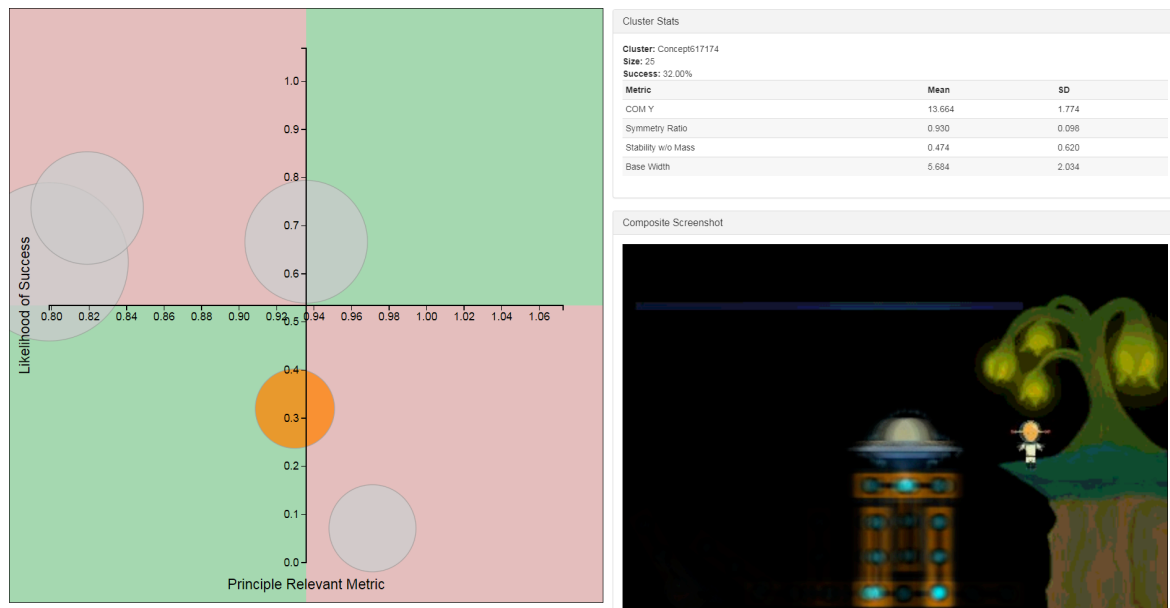


Figure 3. An example of employing TRESTLE to visualize the alignment of clusters of player solutions to a game. Solutions are plotted along an axis for principle relevant metrics (x-axis) and likelihood of success (y-axis). Ideally there would be a relationship between principle relevant metrics and success, denoted by the shaded areas. ²

To support digging further into apparent trends the visualization supports the option of including screenshots that can be linked to instances within the clustering. These screenshots can be composited together (lower right of Figure 3) to allow an analyst to more quickly examine why a particular trend might be happening within their instructional environment and what actions they may explore to fix the problem.

4 CONCLUSION

Our goal in designing visualizations for TRESTLE is to help analysts to organize their data in way that can support intuitive exploration. We have found this to be particularly useful within datasets from complex instructional environments such as educational games. We hope that others can see utility in this approach and can find a way to apply it to their own contexts.

REFERENCES

- Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309. <https://doi.org/10.1109/TVCG.2011.185>
- Christel, M. G., Stevens, S. M., Maher, B. S., Brice, S., Champer, M., Jayapalan, L., ... Lomas, D. (2012). RumbleBlocks: Teaching science concepts to young children through a unity game. In *Proc CGAMES 2012* (pp. 162–166). IEEE. <https://doi.org/10.1109/CGames.2012.6314570>
- Fisher, D. H. (1987). Knowledge Acquisition Via Incremental Conceptual Clustering Learning ~ Element Performance Element. *Machine Learning*, 2, 139–172.

² A live version of this visualization is available at <http://erikharpstead.net/alignment/visualization.html>

- Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2014). Using extracted features to inform alignment-driven design ideas in an educational game. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 3329–3338). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2556288.2557393>
- MacLellan, C. J., Harpstead, E., Aleven, V., & Koedinger, K. R. (2016). TRESTLE: A Model of Concept Formation in Structured Domains. *Advances in Cognitive Systems*, 4, 131–150. Retrieved from <http://www.cogsys.org/papers/ACSvol4/paper10.pdf>

“What do students know, how long does it take them to know?” at a Glance for Teachers and Instructional Designers

Truong-Sinh An, Majd Edriss, Agathe Merceron, Thuy-Anh Nguyen

Beuth University of Applied Sciences

truong-sinh.an@outlook.com, edriss.majd@gmail.com, merceron@beuth-hochschule.de,
thuyanh.nguyen.de@gmail.com

ABSTRACT: In this paper, we present the visualizations realized in the learning analytics service for the Learning Companion Application. User interactions stored by the application form the evidence for these visualizations. The diagrams for teachers enable them to grasp at a glance which topics their students master or not so that they can prepare their next class accordingly. The same diagram offers additional options, like the total number of attempts for instructional designers so that they can reflect on the difficulty level of the exercises. An additional visualization for instructional designers shows the time students spend on each learning object. The LA service is being realized as an LTI-Tool.

Keywords: traffic-light diagrams, xAPI statements, learning locker, elasticsearch, grafana

1 INTRODUCTION

The Learning Companion Application (LCA) is developed in the smart learning project¹ to fit the needs of full-time employees who take part in an Energy Consultant training in a Chamber of Crafts. LCA can be thought of as a learning management system (LMS) with two distinctive features. First, the digital learning resources are stored centrally in a repository and can be accessed without replication when a course is taught in different institutions. Second, it includes a recommendation service for learners which selects appropriate contents, as well as a learning analytics service to different stakeholders, in particular to teachers and instructional designers. LCA is independent of any topic and any institution and, therefore, can be used in other contexts and for other courses (Krauss et al. 2017).

A course in LCA as in many LMS can be divided into sections, which can be divided into learning units. A learning unit contains different learning resources also called learning objects (LO) such as texts, videos, animations, PDF files, other media-types, and exercises. These learning objects can be reused in other courses. To support the pedagogical concept adopted in LCA as well as to implement the recommendation and learning analytics services, metadata are associated with any learning object. These metadata contain among others at least one learning objective and a typical learning time. A learning unit is rendered as an accordion with a specific sequential structure, see Figure 1. The top item is the list of the learning objectives of all learning objects of that unit. A learner can rate each learning objective and so reflect on how much s/he knows already on that topic, from 1

¹ <https://projekt.beuth-hochschule.de/smartlearning/>

(know nothing) to 5 (expert). This item is followed by the sequence of the LOs of that unit. In Figure 1, this list includes a set of exercises shown in orange. Following all LOs, the next item in the accordion-view is again the list of learning objectives. By rating them, a student can reflect on how much s/he knows after learning the unit. The follower item allows students to provide feedback on the typical learning time for that unit (from 1, way too little time to 5, way too much) and give comments. The last item in the list opens a discussion thread on that unit. These last two items are marked a Communication-tools in Figure 1.

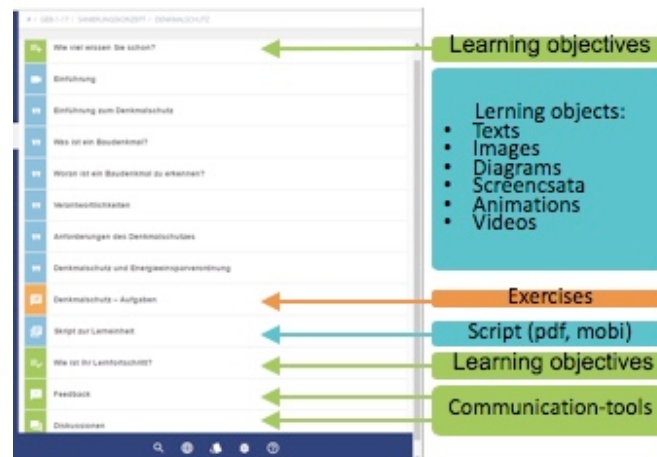


Figure 1: A learning unit in LCA

The first aim of the learning analytics (LA) service is to support teachers and instructional designers. During the project, three meetings with three teachers (N=3) of the Chamber of Crafts have taken place to sense their needs and to discuss proposed solutions. The outcome stressed the importance of a simple and unambiguous visualization: teachers should clearly understand what they see and not be overwhelmed with too much. Therefore, we have opted for well-known diagrams that teachers are familiar with. The LA service should enable teachers to be aware of how many students are mastering, are in the process of mastering or do not master at all the topics of a learning unit, so that they can prepare their next class according to the learning needs of their students. The LA service should enable instructional designers to improve the learning objects they create in cooperation with the teachers. For them too, it is important to understand what they see. However, they may need to explore more students' interactions to be aware of whether the resources they develop have the right length or the right level of difficulty.

In the next section, we describe the interactions data stored by the system. In the follower section, the diagrams for teachers and instructional designers are presented. This paper ends with a conclusion and future works.

2 DATA AND TOOLS

Comprehensive user interactions are stored as xAPI² statements in Learning Locker³. Examples of stored interactions include the opening of a learning unit, opening and closing of every single learning object, self-assessment of each learning objective, attempt in solving an exercise, starting, pausing or quitting a video etc. As an example, consider the following xAPI statement:

```
{ "actor": { "mbox_sha1sum": "13648454125cf6ef31a9e632389c9a806316c9ad" }, (1)
  "verb": { "id": "http://adlnet.gov/expapi/verbs/answered" }, (2)
  "object": { (3)
    "id": "https://vfh143.beuth-hochschule.de/...?itemID=U05LX0ZUU19BRkdfRmV1Y2h0ZXNjaHV0e18wMV9NQw",
    "definition": { "type": "http://adlnet.gov/expapi/activities/question",
      "name": { "de-DE": "Bauphysikalische Grundlagen" } }
  },
  "result": { (4)
    "score": { "scaled": 0.5, "min": -1, "max": 1 },
    "response": "[\\"Die Wasserdampfsättigungsmenge ist die Höchstmenge an Wasserdampf die Luft bei einer bestimmten Temperatur aufnehmen kann.\"]",
    "duration": "PT0H1M11S",
    "extensions": {
      "https://slehwr&46;beuth-hochschule&46;de/xapi/extensions/questionType": "choiceMultiple",
      "https://slehwr&46;beuth-hochschule&46;de/xapi/extensions/correctResponsePattern": [
        "Die Wasserdampfsättigungsmenge ist die Höchstmenge an Wasserdampf die Luft bei einer bestimmten Temperatur aufnehmen kann.",
        "Der Wasserdampfdruck ist abhängig von der relativen Luftfeuchtigkeit und der Lufttemperatur."
      ]
    }
  },
  "context": { (5)
    "platform": "moodle.hwk-berlin.de", (6)
    "statement": { "id": "db36072e-6759-401a-bc7b-daad0677b683" }, (7)
    "contextActivities": { (8)
      "parent": [
        { "id": "https://vfh143.beuth-hochschule.de/...?itemID=U05LX0ZUU19BRkdfRmV1Y2h0ZXNjaHV0eg",
          "definition": { "type": "http://adlnet.gov/expapi/activities/interaction" } }
      ],
      "grouping": [
        { "id": "https://vfh143.beuth-hochschule.de/api/lcms/courses/GEB",
          "definition": { "type": "http://adlnet.gov/expapi/activities/course" } },
        { "id": "https://vfh143.beuth-hochschule.de/...?itemID=U05LX0ZUUw==",
          "definition": { "type": "http://adlnet.gov/expapi/activities/module" } },
        { "id": "https://vfh143.beuth-hochschule.de/...?itemID=U05LX0ZUU19BRkdfRmV1Y2h0ZXNjaHV0eg",
          "definition": { "type": "http://adlnet.gov/expapi/activities/interaction" } }
      ]
    },
    "extensions": {
      "http://adlnet&46;gov/expapi/activities/course": [ "GEB-1-17#GEB" ] (9)
    }
  },
  "timestamp": "2017-03-30T13:30:15.152500+00:00", (10)
  "id": "01a785e0-d77f-4268-860b-2b32883d6c7e" (11) }
```

The xAPI statement with the given *id* (11) above contains the information that a *specific actor* (1) did *answer* (2) a *specific question* (3) with the shown *result* (4) on a specific *timestamp* (10). The *scaled*

² <https://github.com/adlnet/xAPI-Spec>

³ <https://github.com/LearningLocker/learninglocker>

score of 0.5 indicates that the given solution is partially correct; the question was displayed for a *duration* of 1 minute and 11 seconds. For further analysis, the given *response* and the *correct response pattern* are also stored. For the purpose of better understanding the data, further information is bundled in the *context* (5). The statement reference (7) links to the prior stored xAPI statement on a higher level which allows to build a graph of the learning behavior; xAPI statements on the same level, e.g. multiple attempts of *answering* the same question, will refer to the same higher statement (7) within this learning session – their *timestamps* (10) help to order the attempts. Information about the parent (8), like the exercise this question belongs to, and grouping helps to distinguish if a learning object, here the given question (3), is used in several learning units or courses. As the same course can run several times, the *platform* of the host institution (6) and the internal *course short name/id* in that platform (9) help to distinguish between the instances.

The visualizations are realized in the LA service as a plug-in of the Grafana⁴ framework. We use the *statement-forwarding* feature of learning locker since version 2 to sync the statements with elasticsearch⁵ from which Grafana reads the data. Initial import is realized using an own tool.

3 VISUALIZATION

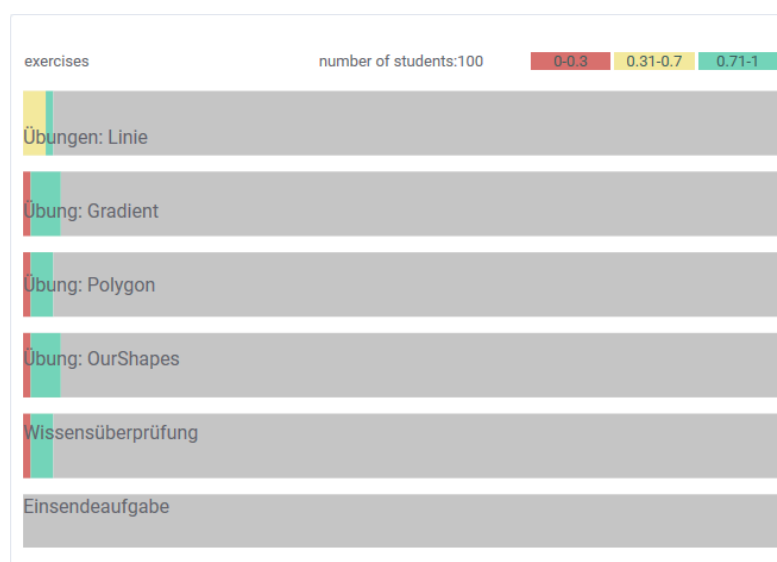


Figure 2: Many of the 100 students did not attempt any exercise (long grey bars). The exercises Gradient and OurShapes were attempted the most and mostly correctly solved (large green area)

The visualization depicted in Figure 2 enables teachers to grasp at a glance the performance level of their students at the level of a learning unit. It uses the well-known traffic-light metaphor used in other works as well, for example in (Dollár & Steif 2012). Teachers see for each exercise of the unit how many students are in green – correct solution –, yellow – solution partially correct –, red – wrong solution –, or grey – no attempt. The time span and the threshold values from red to yellow

⁴ <https://github.com/grafana/grafana>

⁵ <https://github.com/elastic/elasticsearch>

and from yellow to green can be chosen by the user. Figure 1, top corner right, shows the threshold values 0.7 and 0.3: if an exercise has got between 31% and 70% of the maximal score, it counts as partially correct and is color-coded in yellow. There are several options to calculate the performance of a student on an exercise. The default value set for teachers is simply the score of the last attempt as it reflects the best the current knowledge of students so that teachers can adjust their next lesson accordingly. Other metrics are available and can be chosen in a drop-down list. They are primarily for instructional designers to give them awareness on how difficult or easy it was for the students to solve that exercise. As an action, instructional designers in cooperation with teachers might rework that exercise to make it easier or more difficult to solve, or, leave it as is. These metrics are the average, minimal and maximal score on all attempts, as well as the total number of attempts. A general large number of attempts might indicate that the difficulty level is not appropriate.

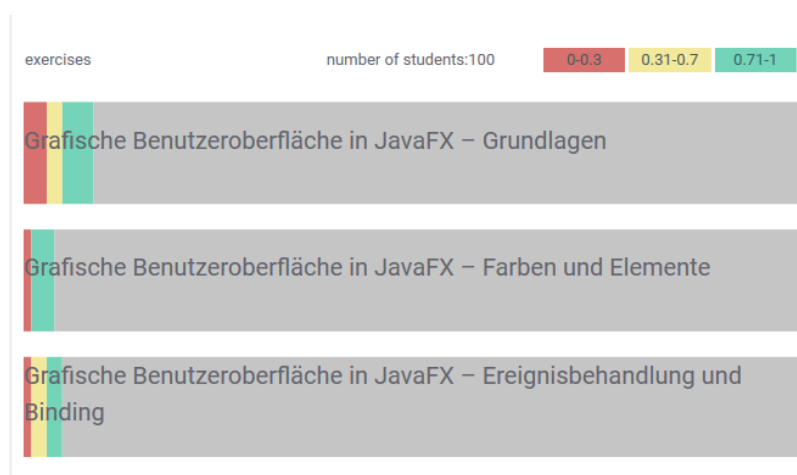


Figure 3: Overview of three learning units

Discussions with teachers have shown that they also need a general overview to plan remediation classes like at the end of a course to tackle again questions that have not been well understood by their students. They need an overview at the course level; if they detect an important part of students in the red or yellow area, they might want to spot the problematic topics, drilling down at the section level, then at the learning unit level as shown in Figure 3, and then into the unit itself and get the visualization presented in Figure 2. For this situation, we have developed a similar visualization: green, yellow, red and grey. Starting from the visualization depicted in Figure 2, the aggregation at the learning unit level uses the well-known method of mapping the values of an ordered categorical variable to ordinal numbers, as for example explained in (Han, Kamber & Pei 2012) p. 74. The values grey, red, yellow and green in this order are mapped to 0, 1, 2 and 3 respectively. Take the example of a unit with three exercises. Consider a student who solved correctly two exercises – green color code – and did not attempt the third exercise – grey color code. The aggregation value at the unit level for this student is $(3+3+0)/3 = 2$, which is color-coded in yellow. If the third exercise was wrongly solved, the aggregation value will be $(3+3+1)/3 = 2.3$, which is also color-coded in yellow. However, if the third exercise was partially correct, the aggregation value will be $(3+3+2)/3 = 2.6$ and color-coded in green. The same procedure is used to produce the visualizations at the section or course level. This procedure can be applied whatever metrics have

been used to produce the visualization of all the exercises of a learning unit. The same kind of visualization has been implemented for the self-assessments on the learning objectives. Teachers can see at a glance how their students assess their own knowledge on each learning objective of a learning unit, and, as above, obtain an overview at the course, section and learning unit level.

For instructional designers as well as for the recommender service, it is important to know whether the typical learning time indicated in the metadata for each learning object is realistic. To this end, we propose a visualization that shows not only the central tendency but also the dispersion of the overall time students spend on a learning object. The visualization is a sequence of simplified box-plots; each box represents a session with the bottom of the box being the minimum time spent by a student on that object in that session and the top of the box the maximal time; average time is drawn as a line in the box; the typical learning time as given in the meta-data is also represented, see Figure 4. On higher levels, such as a learning unit, each box represents the time spent on all learning objects and each student is considered by the overall time spent (sum of all sessions).

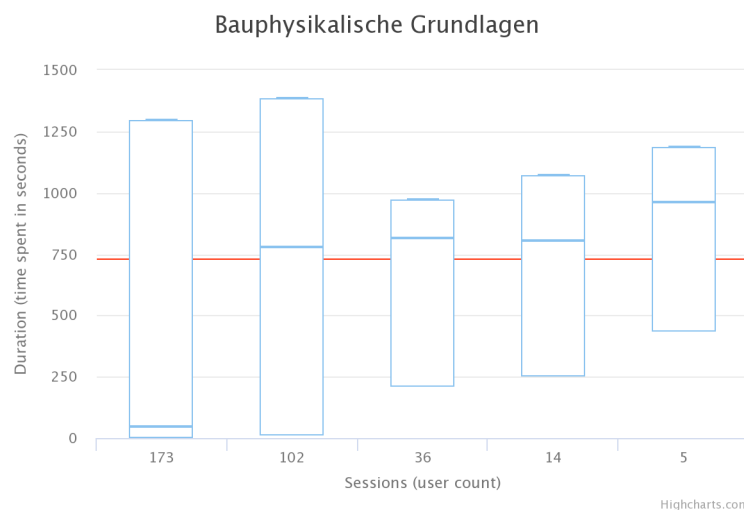


Figure 4: In the first session, 173 students accessed the object. The minimal time spent was about 1 second, the maximum above 1250 seconds and the average about 20 seconds. In a second session, 102 students accessed that object. The typical learning time is given by the red line

4 CONCLUSION AND FUTURE WORKS

The visualizations presented in this paper are for teachers and instructional designers. By showing them how many students are in the green, yellow, red and grey areas, teachers can prepare their next class according to the needs of their students. By showing them evidence of how long students spend on learning resources or how many attempts they make on exercises, instructional designers can reflect on the length and difficulty level of the resources and adapt them. Students can see the same diagrams as teachers do, but with their own data instead of the full class.

An obvious prerequisite for the diagrams to be useful is that the exercises have a high quality and learning objectives are well formulated. This requires some effort. The diagrams support a more

active pedagogy style like the inverted classroom. Experience shows that it requires some training for teachers to integrate the diagrams in their daily routine. Further works include diagrams showing to instructional designers the paths that students follow while navigating through the learning objects and the learning units. This learning analytics service is being realized as an LTI-Tool so that it can be used with any LTI-compliant learning system.

Acknowledgements: The authors thank the whole Smart Learning team. This work is partially supported by the German Federal Ministry of Education and Research grant number 01PD17002B.

REFERENCES

- Dollár, A., Steif, P. (2012). *Web-based statics course with learning dashboard for instructors*, Proceedings of Computers and Advanced Technology in Education (CATE 2012), Napoli
- Han, H., Kamber, M., Pei, J. (2012): *Data Mining Concepts and Techniques* (3rd. Ed.). Elsevier.
- Krauss, C., Merceron, A., An, T.-S., Zwicklbauer, M., Steglich, S., Arbanowski, S. (2017). *Teaching Advanced Web Technologies with a Mobile Learning Companion Application*. Proceedings of the 16th ACM World Conference on Mobile and Contextual Learning (mlearn 2017), ACM, Ont. 30 - Nov. 1, 2017, Larnaca, Cyprus, doi>10.1145/3136907.3136937

Visualizing Cronbach's Alpha for a Large Number of Assessments

Aditya Sharma

Playpower Labs

aditya.sharma@playpowerlabs.com

Nirmal Patel

Playpower Labs

nirmal@playpowerlabs.com

ABSTRACT: In this paper, we present a novel data visualization that shows the distribution of Cronbach's Alpha for a large number of assessments and their items. Cronbach's alpha measure can be used to improve assessments by removing items that are not consistent with other items of the test. The exclusion of items can affect the alpha of the assessment in a different way. The proposed visualization makes it easy to identify assessments where the removal of some items can lead to a significant gain in the internal consistency of the assessment items. The visualization is particularly useful when the number of assessments being analyzed is large. The visualization presented is also open source and reusable.

Keywords: Interactive Data Visualization, Assessment Data, Cronbach's Alpha

1 INTRODUCTION

Quality assessments are central to the design of a good curriculum. They provide us with measurements that help us gauge how well students understand the instruction of the course. Assessments become even more important in the context of online curricula, because the instructor is physically separated from her students, and assessment data is one of the few ways to gain an insight into the progress of the cohort. Online learning platforms are now able to employ various kinds of assessments, ranging from in-video quizzes to auto-graded programming assignments to know how well students are moving towards the goals of the course. But the measures of student knowledge provided by these quizzes and tests are directly related to the quality of the assessments. Good quality assessments can tell us right things about students' progress, but bad quality assessments can give us ambiguous information about student knowledge.

There are many different metrics to assess the quality of the tests. Some of these metrics give us information about the quality of test items, while other metrics give us information about the quality of the overall test. Moreover, different test theories such as classical test theory and item response theory give us different metrics to measure how effective the items and tests are to measure student knowledge. The typical measures for items from classical test theory are percent correct and point biserial (or item-total correlation,) while item response theory has its own procedures to calculate item difficulty and discrimination. For tests, classical test theory provides Cronbach's alpha (Cronbach, 1951) that measures the internal consistency of the items, while item response theory uses a measure called test information function that relates test information with

latent student ability. All of these measures give us actionable information about how we can improve the test so as to improve our estimate of student ability.

2 METHOD

In this paper, we focus on Cronbach's alpha, which is one of the most widely used metrics to measure the quality of tests. Cronbach's alpha tells how closely items within a test are related to each other. It is a measure of internal consistency. If Cronbach's alpha is very high for a test, we can assume that all of the items in the test are measuring a similar knowledge construct. If Cronbach's alpha is very low for a test, we can assume that the questions of an assessment are trying to measure very different or unrelated constructs. Ideally, for any curriculum, we would want the alpha to be high for every assessment. But often times, we can run into assessments with low Cronbach's alpha. One reason for low alpha is that all of the items are measuring different knowledge constructs, but another reason for low alpha is that only one or some items of the test are outliers, while other items of the test are correlated. If we can remove these outlier items, we can make the test more consistent.

One way to identify items that are not consistent with other items of the test is to look at how Cronbach's Alpha is affected when an item is dropped out of the test. If the alpha of the test increases after removal of an item, we can surmise that the item that was removed helped to increase the internal consistency of the test items. A test with high internal consistency can provide us with a more reliable measure of student knowledge. Imagine a test about for loops in programming that contains some items on if/else statements that students haven't learned about. Results of this test are more reliable if we remove all of the items testing if/else statements, because then, the scores will tell us more about the student knowledge of for loops.

To detect outlier items in a test, we can use item response data of students to calculate the alpha for the test, and alpha for the test after removal of each of the test item. The metrics obtained thereafter can help us decide which items from the test can be removed to improve the quality of the test. This gives us a way to use data to improve the design of assessment, which, in turn, might improve the quality of data collected later.

A Cronbach's alpha value of 0.70 is considered acceptable (Cortina, 1993). For a small number of assessments, we can calculate Cronbach's alpha and look at a table of values to find out removal of which increases the alpha of the test. But, when the number of assessments from a curriculum is large (e.g. $N > 100$), we might find it very difficult to identify assessments that can be improved the most. We can look at sorted tabular data to find outlier items that can lead to the most improvement in alpha, but we would not get much insight into the distribution of the rest of the data. In this case, it is possible to use a data visualization that helps us identify both the outlier items and overall pattern of the assessment quality. This is the motivation behind the visualization described in this paper.

The visualization presented in this paper is also open source¹ and is written in R, and it can take data in a specific format and generate the visual for any set of input data.

3 DATA

To generate the visualization using the provided open source R code, we need item response data of students for different assessments in a single table. Table 1 describes the data format and shows an example table that can be ingested by the R program to generate the visualization.

The Assessment ID, Question ID, and Student ID column can be any possible unique identifiers for assessments, questions within those assessments, and students who attempted those assessments. Assessment IDs will appear on the X-axis of the plot, so it is suggested to use more meaningful names in that column. For every Question ID in the data, a dot will be made in the visualization and hovering over that dot will reveal back the Question ID. So interpretable values in the Question ID can also be very useful. Student ID column can contain either anonymized IDs or real student names, it will not make a difference in any aspect of the visualization.

Table 1: Data format and example data required to generate the visualization

Assessment ID (String)	Question ID (String)	Student ID (String)	Correct (1/0)	Time (YYYY-MM-DD HH:MM:SS)
Test1	Q1	Anon1	1	2018-01-01 10:00:00
Test 1	Q2	Anon1	0	2018-01-01 10:01:00
Test 1	Q3	Anon1	1	2018-01-01 10:02:20
Test 2	Q1	Anon2	1	2018-01-02 11:30:00
Test2	Q2	Anon2	0	2018-01-02 11:30:55
Test3	Q1	Anon3	1	2018-01-04 15:20:12

4 VISUALIZATION

The visualization is generated by using the data described in Table 1. The visualization program uses the function `alpha()` from R package `psych` (Revelle, 2017) to calculate the Cronbach's alpha. The alpha value is calculated for each assessment inclusive of all items, and the alpha value is re-calculated by removing each item of the assessments. So if an assessment has n test items, $n + 1$ alpha values will be calculated for that assessment. If there are a total of m unique assessment IDs in the data, $m \times (n + 1)$ alpha values will be calculated. Once the calculation is finished, all of these values will be visualized together.

An example of the visualization is shown in Figure 1. The X-axis of the plot shows the Assessment ID. The Y-axis of the plot ranges from 0 to 1 and points the Cronbach's alpha for the assessment on the X-axis. In the visualization, every assessment is represented as a vertical line with multiple dots. The dots are of two colors: blue and black. The blue dots show the alpha of assessment without dropping any item, while the different black dots show alpha of the assessment after dropping one of the

¹ <http://github.com/nirmalpatel/edviz-2019>

items. A red horizontal line at $Y = 0.7$ is drawn as a reference alpha value that is generally seen as acceptable. The assessments on the X-axis are sorted from the highest to lowest alpha values that were calculated by using all of the assessment items. The visualization is also interactive, and as shown in Figure 2, hovering over any dot will tell us which item the dot refers to, and how the alpha of the assessment is affected by removing the selected item.

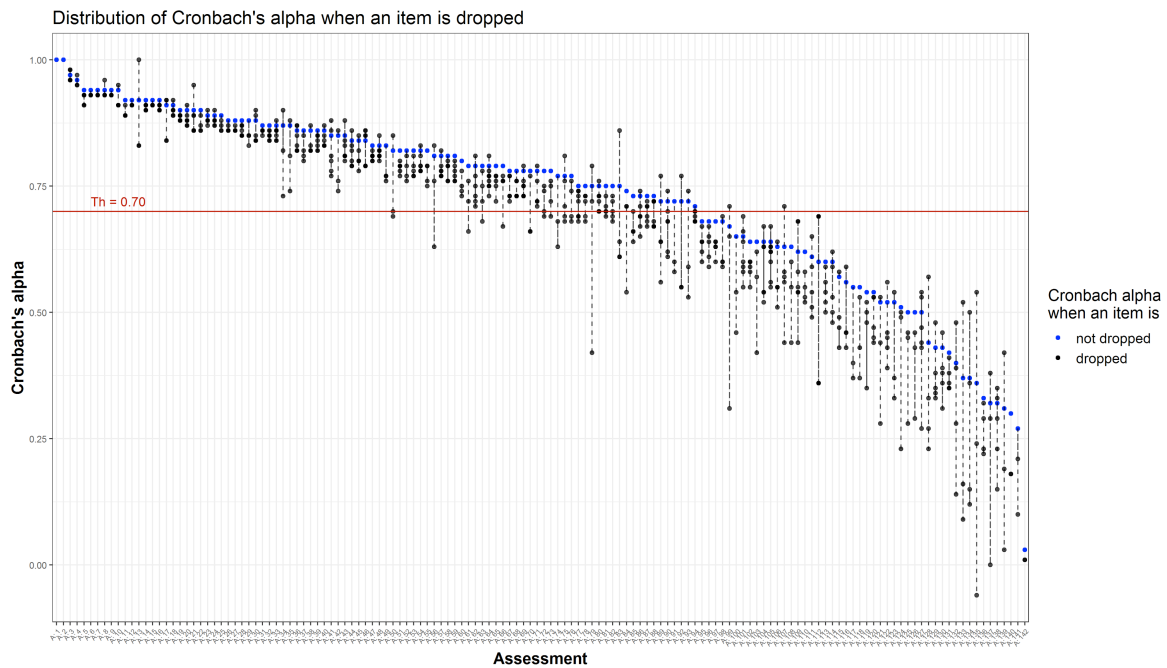


Figure 1: Distribution of Cronbach's alpha when an item within an assessment is dropped. The points in blue color signify Cronbach's alpha for the overall assessment and the points in black signify Cronbach's alpha when one of the assessment items is dropped.

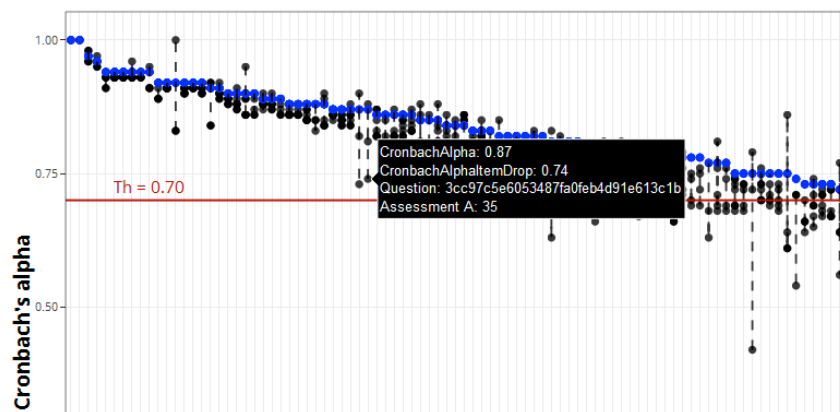


Figure 2: The assessment and item information show up when we hover over the dots of the plot.

5 DISCUSSION

The visualization of Cronbach's alpha for many assessments at once provides us a very easy to grasp overview of the assessment quality data. Rather than just seeing a distribution (or histogram) of Cronbach's alpha for a set of assessments, the visualization enables us to see how different items are impacting the overall alphas of the tests.

The first feature to notice is where the blue dot or the overall alpha of the assessment lies. If this is below 0.7 or another expert defined threshold, the test may be of concern. Another noticeable feature of the visualization is how close all of the black dots are for an assessment. If the dots are very close together, this that the removal of a single item may not affect the alpha of the test significantly. If the black dots of multiple items coincide, it means that the coinciding items contribute similarly to the scale in terms of consistency.

The next important feature of the plot is assessments with wide variation in their alpha values when items are removed. Some black dots in the plot goes high above the blue dots, while some black dots go well down below. The black dots going above the blue dot tell us that removal of an item can increase the alpha of the assessment, while the black dots going below the blue dot tell us that if we remove some items of the assessment, the alpha value can decrease.

In Figure 1, we can see that on the top left the 13th assessment from the left has one item that can be removed to increase the overall alpha, while removal of another item can decrease the alpha. There are a few more assessments between an alpha value of 0.5 and 0.7 that can be improved by removing certain items.

6 CONCLUSION

This described visualization can be used by instructional designers to identify assessments that need attention. Quality tests are a very important tool to gauge student understanding of the learning material, and having a test with high consistency can provide instructors with more reliable estimates of student mastery. This reliability can lead to a better ability to use data to drive instruction.

7 FUTURE WORK

The visualization is still limited in its ability to show other related item analysis metrics such as percent correct and point biserial. These can provide users with more information about how well items in the plot are performing. Another feature that is missing in this visualization is showing how the removal of multiple items can affect the overall alpha. Although this is more complex as there are combinatorial possibilities when removing a combination of items. But adding these features to the plot might help assessment designers track down problematic items in a more better way.

REFERENCES

- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, 78(1), 98.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Revelle, W. (2017) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psych> Version = 1.7.8.

3rd Annual Workshop of the Methodology in Learning Analytics Bloc (LAKMLA19)

Yoav Bergner
New York University, USA
yoav.bergner@nyu.edu

Charles Lang
Columbia University, USA
cl3584@tc.columbia.edu

Geraldine Gray
Technological University
Dublin, Ireland
geraldine.gray@itb.ie

ABSTRACT: Learning analytics is an interdisciplinary and inclusive field, a fact which makes the establishment of methodological norms both challenging and important. Building on the success of the LAK17 and LAK18 workshops on methodology, this community-building workshop intends to convene methodology-focused researchers to discuss new and established approaches and co-develop guidelines to help move the field forward with quality and rigor.

Keywords: Models, Methodology, Measurement, Statistics, Evaluation

1 WORKSHOP BACKGROUND

Learning analytics is an interdisciplinary and inclusive field that brings together educational technologists, psychologists, data scientists, learning scientists, substantive experts in various domains, and measurement specialists (Siemens and Gašević, 2012; Bergner, Gray, and Lang, 2018). For all of the strength that comes from such diversity, there are also potential pitfalls when it comes to establishing norms for methodological work. For example, Clow (2013) described learning analytics as, “a ‘jackdaw’ field of enquiry, picking up ‘shiny’ techniques, tools and methodologies... This eclectic approach is both a strength and a weakness: it facilitates rapid development and the ability to build on established practice and findings, but it—to date—lacks a coherent, articulated epistemology of its own.” (p. 686).

In the years since this observation, the learning analytics community has grown rapidly, and the number of shiny techniques has grown as well. Looking just at recent proceedings of the International Conference on Learning Analytics and Knowledge (LAK), the variety is staggering. Methods range from descriptive statistics to correlation analyses, classification, clustering, regression, (M)AN(C)OVA, structural equation modeling, item response theory, hidden Markov models, time-series analysis, latent semantic analysis, social network analysis, and the list goes on. It is understandable and even expected that reviewers and readers of learning analytics manuscripts are unlikely to be expert evaluators of the methodological rigor in all of these cases. There is a naturally occurring process of specialization in any academic field. However, if growth of adoption outpaces systematic specialization then there is a risk that methodological errors will proliferate and that quality of community products will suffer.

To make matters even more complex, a number of recent papers have emphasized the sensitivity of quantitative analyses to data collection and variable operationalization choices, for example with

regard to effects of selection bias (Brooks, Chavez, Tritz & Teasley, 2013), results of time-on-task analyses (Kovanović et al., 2015), studies of discussion forum usage (Bergner, Kerr, & Pritchard, 2015), evaluation of student models (Pelánek, Rihák, & Papoušek, 2016) and the definition of social ties in social network analysis (Fincham, Gasevic, & Pardo, 2018). In addition, learning analytics models often incorporate a selection of proxy variables as indicators of latent constructs, such as learning and engagement. What proxy variables actually measure is less clear. For example, measures of engagement may be influenced by instructional conditions (Gašević, Dawson, & Siemens, 2015; Motz, Carvalho, de Leeuw, and Goldstone, 2018), adding ambiguity, and a lack of consistency, to our interpretation of models of learning.

In short, methodological concerns can arise from a range of practices including but not limited to selecting inappropriate methods, misusing methods, inadequate model evaluation or model comparison, sensitivity to operationalization, and over-reliance on proxy variables (Bergner, 2017). As the learning analytics community matures, it is particularly important to establish standards for good practice and to educate new students in accordance with these standards. Clear methodological guidelines increase the quality of work and facilitate communication not only within the community but also with practitioners in other research communities, where norms may be clearer. This is a challenging problem in large part because of the aforementioned diversity of approaches. The present workshop seeks to build a community of researchers with an interest in methodology and its rigorous application and development to the field of learning analytics.

There have been several previous LAK workshops and tutorials that have focused on specific methodologies—a limited set of examples includes the tutorials for classification and clustering using Weka (2014, 2016), special topics in discourse analysis (2013-2014), writing analytics (2016-2018), multimodal learning analytics (2016-2018), assessment design (2016-2017), and temporal analysis (2012-2016). In contrast, this workshop series focuses on cross-cutting methodological issues such as developing methodological frameworks within learning analytics, framing and prioritizing methodological issues for the community, and providing resources to move the field forward.

1.1 Building on the LAK17 and LAK18 Workshops

The LAK17 and LAK18 Methodology Workshops received substantial interest from a variety of LAK participants, from seasoned computer scientists to people who were entering the field of learning analytics for the first time. This interest led to the Journal of Learning Analytics' recent publication of a special section on methodological choices, encompassing a range of relevant approaches including sensitivity analysis, model evaluation, model fit, causal inference, and visualization as a methodology (Bergner et al., 2018). Together, these events have served to seed a community that is interested in having both high level discussions of what methodology means in learning analytics and specific methodological issues that can arise in both quantitative and qualitative investigations. We plan to continue to build this community at LAK19 with an eye to developing a taxonomy of methods in Learning Analytics. The LAK17 event began this process by defining possible projects, such as “cheat sheets” for relevant methods and the publication of methodological problems specific to the field. This was continued in both LAK18, which started to describe the range of methodologies prevalent in learning analytics, and the publication of the JLA special section on methodological choices. We would like the opportunity to follow up on the progress made so far by exploring what a taxonomy

of methods in learning analytics might look like, and explore other ways of ensuring a continued focus on methodology in a way that is accessible to the field as a whole.

1.2 Relevance to the Theme

We plan to incorporate the LAK19 theme into the workshop in various ways. Specifically, ways in which learning analytics can be used to promote inclusion and success as a methodology in itself and the importance of methodology in ensuring results from learning analytics represent valid actionable intelligence. A focus on methodological rigor also supports evidence-based learning analytics practice.

2 ORGANISATIONAL DETAILS

The proposal is for a half day, open workshop covering introductions (15 mins) and a series of individual and group based activities. Including breaks, the session will last 4 hours in total. A call to participate will be disseminated through relevant listservs, our network of contacts that have expressed an interest in methodology in learning analytics, and a workshop website. Participants are welcome submit their areas of interest in advance of the workshop to facilitate birds of a feather discussion groups during workshop activities. The expected participant number is approximately twenty.

3 WORKSHOP OBJECTIVES

Methodology Guidelines Posters

The first objective of the present workshop is to advance the development of guidelines on methods relevant to learning analytics, and promote critical review of methodological choice. Several participants in our previous two workshops confirmed a need for products which provide compact, substantive guidance on methodological issues. Thus we continue with work started at the LAK18 workshop to cooperatively develop community guidelines regarding the uses of various methods including data acquisition, data analysis and evaluation of results in conference and journal publications. For example, these could take the form of Methodology Guideline Posters (MGPs) that are infographic representations of decision flows in learning analytics methodology, working backwards from the ultimate goals. An MGP will emphasize how operational decisions are guided not only by the goals but also by the types and properties of available data and by problems of statistical inference. We do not imagine that MGPs will be instructional with regard to how to carry out analyses but rather will rather point the reader to appropriate references. The emphasis of MGPs will be interrogating the methodological choices. As such they should describe alternate case scenarios, explain pitfalls, and suggest options for sensitivity and goodness-of-fit tests.

A Taxonomy for Learning Analytics Methods

Group based activities will also address a second objective of the workshop, the development of a taxonomy of learning analytics methods to serve as a framework for the development of methodology guidelines.

Provide Expertise for Review Panels

A third objective of the workshop is to take responsibility for maintaining a database of methodology experts who are active in the learning analytics community. The expert listing is by no means intended to be exclusionary or to promote certain researchers over others but rather to help community members and editorial committees find methodology experts who are willing to consult and/or review relevant work.

Community Building

Last but not least, an objective of the workshop is to provide a meeting place for researchers who take a special interest in methodological issues. We anticipate that a concentrated meeting will promote continuing collaboration on this important topic.

4 REFERENCES

- Bergner, Y., Gray, G., Lang, C., (2018) What Does Methodology Mean for Learning Analytics? *Journal of Learning Analytics* 5(2), 1-8.
- Bergner, Y. (2017) Measurement and its uses learning analytics. In Lang, C., Siemens, G., Wise, A., & Gašević, D. (Eds.), *Handbook of Learning Analytics* (pp. 34-48). Society for Learning Analytics Research.
- Bergner, Y., Kerr, D., and Pritchard, D. E. (2015) Methodological Challenges in the Analysis of MOOC Data for Exploring the Relationship between Discussion Forum Views and Learning Outcomes. *Proceedings of 8th International Conference on Educational Data Mining*.
- Brooks, C., Chavez, O., Tritz, J., and Teasley, S. (2015) Reducing selection bias in quasi-experimental educational studies. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp 295–299), ACM.
- Clow, D. (2013) An overview of learning analytics. *Teaching in Higher Education* 18(6), 683–695.
- Fincham, E., Gašević, D., & Pardo, A. (2018) [From Social Ties to Network Processes: Do Tie Definitions Matter?](#) *Journal of Learning Analytics* 5(2), 9-28.
- Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R.S.J.D., and Hatala, M. (2015) Penetrating the black box of time-on-task estimation. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (pp. 184-193), ACM.
- Gašević, D., Dawson, S., and Siemens, G. (2015) Let's not forget: Learning analytics are about learning. *TechTrends* 59(1), 64-71.
- Pelánek, R., Rihák, J., and Papoušek, J. (2016) Impact of data collection on interpretation and evaluation of student models. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, (pp 40–47), ACM.
- Siemens, G. and Gašević, D. (2012) Guest Editorial-Learning and Knowledge Analytics. *Educational Technology & Society* 15(3), 1–2.
- Motz, B. A., Carvalho, P. F., de Leeuw, J. R., Goldstone, R. L. (2018) Embedding Experiments: Staking Causal Inference in Authentic Educational Contexts. *Journal of Learning Analytics* 5(2), 9-28.

Can Learning Analytics Generate Trustworthy Assessment on a Geometry Game?

José A. Ruipérez-Valiente, Yoon Jeon Kim, Louisa Rosenheck, Philip Tan and Eric Klopfer

Massachusetts Institute of Technology
{jruipere, yjk7, louisa, philip, klopfer}@mit.edu

ABSTRACT: There are many benefits to using learning analytics and educational data mining approaches for measuring learning and behavior in educational games. In this LAK Hackathon proposal, we present a challenge to build learning analytics for the assessment machinery of Shadowspect, a game-based assessment system game about 3D geometry.

Keywords: Game-based Assessment, Learning Games, Learning Analytics

1 BACKGROUND

There are numerous reports that indicate how traditional assessment can become extenuating and stressful. One alternative method that has gained importance in recent years is the use of games and less intrusive and indirect assessment methods that do not interrupt the flow experience. We have developed Shadowspect, an educational game that aims to explicitly measure common core geometry standards (e.g. visualize relationships between 2D and 3D objects) and relevant reasoning skills (e.g. spatial reasoning). Shadowspect sessions consist of a series of puzzles, where each one is composed of three orthogonal views of a figure, where each figure is composed of a series of 3D geometric primitives. Participants build a 3D figure by using the 3D game environment prototype to solve the puzzles, or to create imaginative structures in the game's sandbox mode ([see a video online](#)). Shadowspect is able to collect very rich data that allows us to reconstruct in deep detail the students' interactions with the game. We will present for this challenge a dataset from Shadowspect from both puzzle and sandbox modes, and prepare a playable demo of the game as well.

2 RESEARCH QUESTIONS

There are two main research questions, one specific and the second one more open:

1. Can we implement learning analytics that can generate reliable assessment of 9th grade geometry standards and spatial reasoning skills based on the data from the learning game?
2. Can we implement behavioral algorithms to detect interesting cognitive skills or behaviors for lifelong learning like creativity, experimentation, productive struggle or strategy?

3 EXPECTED OUTCOMES

The ideal outcome of this challenge would be a set of algorithms applicable to Shadowspect data that could provide a response to the research questions, and a batch of interesting results that could provide the basis for a potential joint publication between the Shadowspect team and the LAK Hackathon team that worked on this challenge.

Learning Analytics in Open and Distributed Knowledge Infrastructures

Atezaz Ahmad

DIPF | Leibniz Institute for Research and Information in Education
ahmad@dipf.de

ABSTRACT: In recent years, the field of learning and education has transcended from the traditional brick and mortar institutions to a more open and global scope. Learning analytics plays a vital role in supporting researchers and teachers to improve learning. With the help of learning analytics methods, we are able to analyze and extract information from educational data sets. In this study, we will be analyzing metadata from different learning platforms and probably with different data sources. We will try to figure out the limitations (like data security, privacy, compatibility etc.) and find out its solutions. Then we will analyze metadata (e.g. course presentations, research papers, and other learning resources) with the help of text data mining and apply learning analytic methods to make sense out of the data.

Keywords: learning analytics, educational data mining, open educational resources, recommender systems, metadata

1 BACKGROUND

In recent years, the field of learning and education has transcended from the traditional brick and mortar institutions to a more open and global scope. Learning analytics plays a vital role in supporting researchers and teachers to improve learning. With the help of learning analytics methods, we are able to analyze and extract information from educational data sets. Usually, learning analytics deals with the development of methods that harness educational data sets to support the learning process. Learning analytics has opened up new ways of learning, which have led to learner-centered, open, and networked learning models, e.g. Personal Learning Environments (PLEs), Open Educational Resources (OER), Massive Open Online Courses (MOOCs) etc [1], [2].

Currently, we have metadata from different educational institutions across Europe stored in different databases. Extracting the data from metadata of different resources is always problematic. It has many restrictions like data security, privacy and of course compatibility.

In order to start, it is a good practice to use a reference model by MA Chatti [3] for learning analytics based on four dimensions, i.e. *data*, *environments*, *context* (what?), *stakeholders* (who?), *objectives* (why?), and *methods* (how?), as shown in Figure 1. The dimensions are as follows:

What? What kind of data does the system gather, manage, and use for the analysis?

Who? Who is targeted by the analysis?

Why? Why does the system analyze the collected data?

How? How does the system perform the analysis of the collected data?

This reference model has led us to a few research questions.

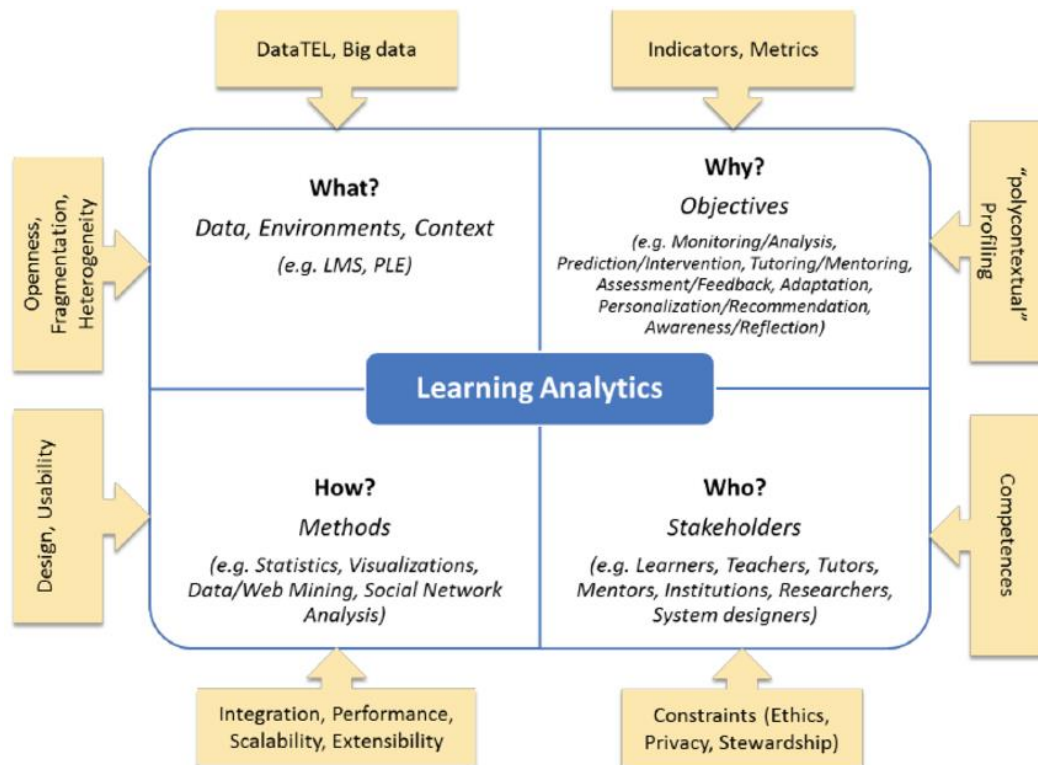


Figure 1: Learning Analytics Reference Model [3]

2 RESEARCH QUESTIONS

- 1) How to extract the data from the metadata of different open and distributed educational resources? How data mining will help in the extraction of meaningful data?
- 2) What to extract from the metadata? And which type of data should be used for learning analytics?
- 3) Which learning analytics methods will be beneficial to use?
- 4) Will it help the stockholders to use recommender system?

3 EXPECTED OUTCOMES

We aim to use the metadata from different sources (learning platforms) keeping in mind the different limitations and apply data mining techniques to extract the required and meaningful data. This data will be processed with learning analytics methods and will be presented on a dashboard in the form of statistics or visualizations. The dashboard may also contain processed data by the recommender system methods. We will further identify various challenges and research opportunities in the area of learning analytics in open and distributed knowledge infrastructures.

REFERENCES

- [1] Chatti, MA, Lukarov, V., Thus, H., Muslim, A., Yousef, AMF, Wahid, U., ... & Schroeder, U. (2014). *"Learning analytics: Challenges and future research directions"* eled , 10 (1).
<https://eled.campussource.de/archive/10/4035>
- [2] Verbert, Katrien, et al. *"Learning analytics dashboard applications"* American Behavioral Scientist 57.10 (2013): 1500-1509.
<https://journals.sagepub.com/doi/abs/10.1177/0002764213479363>
- [3] Chatti, Mohamed Amine, et al. *"A reference model for learning analytics."* International Journal of Technology Enhanced Learning 4.5-6 (2012): 318-331.
<https://www.inderscienceonline.com/doi/abs/10.1504/IJTEL.2012.051815>

Role of Learning Analytics in Individual, Goal Driven Person – Job Matching

Gábor Kismihók

Leibniz Information Centre for Science and Technology
Gabor.Kismihok@tib.eu

Stefan Mol

University of Amsterdam
s.t.mol@uva.nl

ABSTRACT: Goal-setting is thought to enable students to become skilled self-regulated learners, and in the long-term to become self-directed lifelong learners. Therefore, this hackathon topic aims to create an opportunity for educational researchers, computer scientists and practitioners to explore ways to support university students to set learning goals in the light of labour market data.

Keywords: Goal-setting, labour market, dashboard, recommendation system

1 BACKGROUND

Goal-setting (GS), as an important dimension of Self-Regulated Learning (SRL) forehead stage, was the discussion focus in the Goal-setting workshop at Learning Analytics and Knowledge Conference (LAK) 2016. Participants suggested that GS should be an integral part of designing learning interventions (Wise et al., 2014). They also discussed the limited organizational uptake of GS, despite its demonstrated effects on study success. There is also evidence that learning analytics dashboards aid the visualization and internalization of learning goals and objectives (Scheffel et al., 2014; Verbert et al., 2014). Following the GS workshop and subsequent research work, this hackathon topic aims at continuing the conceptualization of GS and Learning Analytics (LA) interface (Mol et al., 2016).

2 RESEARCH QUESTION

During this session we will build on the LAK16 Goal setting workshop (Mol et al., 2016), the LSAC 2018 Hackathon, available open source applications (Goal Setting app), and a large amount of labour market data (vacancy announcements) to formulate personal goals beyond the frames of formal education. More precisely the work will focus on designing a dashboard, which provides learning recommendations to students based on labour market data.

3 EXPECTED OUTCOMES

As a result we expect 1) to fine tune the requirements for the dashboard and 2) to provide a prototype for the algorithm, which will recommend learning tasks for the learner.

REFERENCES

- Mol, S.T., Kobayashi, V.B., Kismihók, G. & Zhao, C. (2016). Learning through goal setting. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 512–513
- Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Journal of Educational Technology & Society* 17, 4, 117.
- Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F. Parra, G. & Klerkx, J. (2014). Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing* 18, 6, 1499–1514.
- Wise, A., Zhao, Y., & Hausknecht, S. (2014). Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics* 1, 2, 48–71.

Multimodal Time-Series analysis with the Multimodal Tutor for CPR

Daniele Di Mitri

Open University of the Netherlands

daniele.dimitri@ou.nl

Jan Schneider

DIPF - Frankfurt

schneider.jan@dipf.de

ABSTRACT: This Hackathon challenge opens up the analysis of a multimodal data corpus which will be used to train the Multimodal Tutor for CPR. A dataset of 6000 chest compressions from 14 participants was collected. In this hackathon challenge, the participants are asked to provide suggestion of the methodology for the data analysis of this multimodal corpus. This challenge belongs to the Multimodal Learning Analytics theme described in the Hackathon proposal.

Keywords: Multimodal Learning Analytics, Hackathon challenge, Machine Learning

1 BACKGROUND

The Multimodal Tutor for CPR is an intelligent system that supports people to learn cardiopulmonary resuscitation using patient manikins. The tutor uses a multi-sensor setup for tracking the CPR execution and generating personalised feedback. The multimodal data are the trainee's body position collected with with Microsoft Kinect, electromyogram recorded with Myo armband, and performance metrics derived from the Laerdal ResusciAnne QCPR manikin. This specific manikin can be linked with the SimpPad reporter, a device capable of calculating reliable CPR performance metrics, such as *CompressionRate*, *CompressionDepth* and *CompressionRelease*. To train the models used by the CPR tutor, we run an expert pilot study collecting around 6000 chest compressions from 14 different participants. Each chest compression is composed of 200 attributes, which are 0.5 seconds long time-series. The aim of this Hackathon challenge is to find the best way to compare the chest compressions.

2 RESEARCH QUESTION(S)

- How to best represent the time series to be used in a classifier?
- Can we use deep neural networks and feed the entire signals instead of extracting features?
- Is it better to train a classifier for each participant individually or one using the data from all the participants?
- How can we improve the overall prediction accuracy of the classification models?

3 EXPECTED OUTCOMES

In this Multimodal Hackathon Challenge we expect to make progress on the analysis of the multi-dimensional data. The expected outcome is to find a mechanism to best represent the data without losing precious details but still be able to allow machine learning algorithms to learn how to distinguish the classes accurately. One possibility is the use deep neural networks as classifier to avoid manual feature extraction, which can cause missing relevant information.

Open and Reproducible Learning Analytics through the development of an R or Python Library

Alan Mark Berg

ICTS, University of Amsterdam

a.m.berg@uva.nl

ABSTRACT: This orientation paper asks the question: How do we ease the programmatic barriers for data scientists who wish to enter the field of Learning Analytics?

Keywords: R package, reproducible science, data scientist

1 BACKGROUND

Data Scientists use tools to make their tasks easier. These tools may include GUI based systems such as SPSS and programmatic languages such as R and Python. To ease the learning curve for the implementation of Learning Analytic workflows, often known as pipelines researchers can develop example library(s), that provide a fully functioning pipeline. The purpose of the libraries being first to entice data scientists into the field and secondly to provide functionality that supports reproducible science.

2 RESEARCH QUESTIONS

- *What is a Learning Analytic workflow in the context of programmatic Datascience?*
- *Which functions are necessary to decrease the number of lines of code to achieve the full workflow?*
- *Which freely available LA related data sources already exist that libraries can pull into the developer's environment via helper functions?*
- *What is the relationship between an LA targeted library and commonly used libraries in the Tidyverse and Caret?*

3 EXPECTED OUTCOMES

Consider initially developing dummy functions with documentation so as to scope out the core features of the library(s). Secondary, code the core feature set.

The main outcome is an initial R or Python library(s) with many dummy functions scoping the broad range of functionality needed by data scientists to easily transition to coding Learning Analytic pipelines. Each function dummy or functional being fully documented so that in follow-on

Hackathons the participants understand the intent. The initial package(s) should contain fully descriptive documentation and stored in the following GitHub location¹.

¹<https://github.com/AlanBerg/Package-Hackathon-LAK19>

Diving in to Educational Experiments: Process, Evaluation, and Reasoning in Support of Learning (DEEPER Support of Learning)

ABSTRACT: The path to improving educational efficacy, equity, and experience is paved with questions of cause and effect. This workshop focuses on the most reliable approach for estimating causal effects: randomized experiments. The workshop's aim is to promote and strengthen the use of online experiments within the learning analytics community. To this end, we will review best practices for the design, implementation, and analysis of experiments in educational settings and learn about recent innovations in the field.

Keywords: Experimentation, causation, reasoning

1 BACKGROUND

Randomized experiments in educational settings are an accepted standard for making claims about causal effects about learning and pedagogical decisions. Widespread adoption of online learning has opened new opportunities to conduct large-scale experimentation at lower costs (Kizilcec and Brooks 2017). In light of these unique affordances, Stamper and colleagues proclaimed a new era in experimentation putting forward a Super Experiment Framework (SEF), where multiple experiments can be conducted online at the same time in authentic learning contexts (Stamper et al. 2012). Besides the obvious evolution of the infrastructure underpinning large-scale experimentation, new exciting developments, such as micro-randomized trials in health, have taken place in how experiments could be designed. Finally, extension of mainstream statistical toolkits to Bayesian statistics and the development of machine learning algorithms offered new avenues to more rigorous analysis of data collected through experimental research.

Despite the promise to strengthen the quality of insights gleaned by learning analytics through experimentation, the uptake of educational research in our community is not high. Such in part could be due to the tension existing in educational research around the evidence-based approaches (Nelson and Campbell 2017). Despite its prominence as a method for claims around cause and effect in science, a counter-narrative challenges the relevance and application of experimentation when it comes to education. For instance, Biesta (2010) critiqued key assumptions of evidence-based education (including experiments) as limiting the scope of educational effectiveness and restricting opportunities for participation in educational decision making. Besides, the issues of whose evidence counts and how professional practices interact with evidence, highlighted by Biesta, educational researchers also raise concerns about ethical considerations underpinning experimentation in authentic learning settings. The issues around adoption and use of evidence from experimentation in authentic educational settings are best demonstrated through the recent controversial study by Pearson presented at the AERA, which resulted in significant press coverage (Strauss 2018; Herold 2018) which ended up with Pearson disavowing the inquiry as experiments, instead arguing that the “messages weren’t psychological experiments, but product tests.” (Fussell 2018)

The Learning Analytics and Knowledge (LAK) conference offers a unique venue where researchers interested in informing action to improve learner experiences through evidence can engage in dialogue and use of experimental research in education. This workshop is envisioned as a stepping

stone towards a stronger use of experimental research within learning analytics community. The workshop will also set tone for a broader discussion around issues associated with participatory and rigorous experimental research in educational settings. The workshop aims to bridge the knowledge gap about experimental thinking and help broker connections between those in the learning analytics community open and interested in doing experiments. The workshop will briefly cover the fundamental concepts required for experimentation but will focus on introducing innovative experimental approaches. During the workshop we will create opportunities for researchers and practitioners to partner and design experiments.

2 ORGANIZATIONAL DETAILS

We propose to run this workshop as a half day event. This workshop will have three stages: pre-workshop activities, short lectures, and breakout groups.

2.1 Pre-Workshop Activities

On signing up, participants will be asked to indicate on a google sheet whether they fit in the category of a (i) learning context or system provider, (ii) experimental methods researcher, or (iii) a learning theory researcher. The intent of the workshop is to match people from each group together to design experiments to collect evidence about a given learning theory in a learning context or system. For instance, someone who teaches a course on Data Science on the Coursera platform might sign up in the first category, while someone who does sequential randomized trials might be in the second category, and someone who is interested in Growth Mindset (Dweck 2009) might be in the third category. We expect that matching these individuals together in the breakout groups (described below), will be powerful in generating new avenues of investigation to move the field forward.

Once the workshop is approved, the workshop proposal authors will all fill in this spreadsheet to start discussions, and this description of pre-workshop activities will be a key aspect of our recruitment plan.

2.2 Short Lectures (one hour)

On the day of the workshop we anticipate spending the first hour engaging in short lectures on experiments in education and learning analytics. These lectures will be intended as training on new methods for learning analytics researchers. Our draft schedule of lectures might include topics such as:

1. Common Ground in Experimental Research: Fundamentals, Sasha Poquet, National University of Singapore
2. Nuts and Bolts of Conducting Online Learning Experiments, Rene Kizilcec, Cornell
3. Sequential Randomized Trials for Developing Personalized Interventions, Timothy NeCamp, University of Michigan

4. Evaluating Randomized Trials Using Bayesian Analysis, Josh Gardner, The University of Washington
5. Lightning Talks about Experiments in MOOCs, Dan Davis, TU Delft, Christopher Brooks, University of Michigan, Rene Kizilcec, Cornell, Timothy NeCamp, University of Michigan
6. An Instructor Centered Approach to Practical Experimentation through Machine Learning, Joseph Jay Williams, University of Toronto

Each mini-lecture will be posted online (slides only, unless recording facilities are available) after the event, and we anticipate them being no more than (and perhaps shorter than, depending on the number of presenters) 15 minutes in length with 5 minutes for questions.

2.3 Breakout Groups (two hours)

The second part of the workshop program will see participants put into breakout groups for 60 minutes at a time. Each group will have at least one of each kind of participant as determined through the pre-workshop activities. Groups will have 45 minutes to develop their focal education research questions and design an experiment to answer those questions, and 15 sharing out with the larger group. We expect that there will be roughly three rounds of breakout groups, resulting in dozens of different experiments being shared using different methods and investigating different learning theories.

3 PARTICIPANTS AND RECRUITMENT

We intend to recruit no more than 20 participants to the workshop, including the authors of this proposal. Recruitment of participants will be done through social networks, and open advertisement on mailing lists of various communities (LAK, EDM, ICLS, AIED, etc.). Enrollment will be on a first come first served basis based on full registration in the LAK19 registration system, minus the 7 spots which will be reserved for the workshop proposers and released after the LAK19 registration early-bird deadline.

4 REQUIRED EQUIPMENT

For this workshop we will need projection facilities (e.g. computer and projector) and a small room (for the attendees). No other equipment is needed.

5 INTENDED OUTCOMES

Through the running of this workshop we are interested in:

1. Strengthening the methodological base of students and other researchers when conducting experiments in education (e.g. power analysis, causal inference, confounding factors, experimental design, and post-experimental data analysis);
2. Developing new educational experiments in a collaborative setting using next generation experimental techniques, such as sequential randomized trials (Murphy 2005) and adaptive

trial designs (Chow and Chang 2008), within the community of learning analytics researchers;

3. Increasing the social capital and connections between researchers and practitioners who are engaged in educational experiments, potentially leading to a community of practice or SOLAR special interest group (SIG);
4. Creating an environment by which joint research proposals can take shape.

In achieving these outcomes we anticipate creating open online resources based on the presentations given which might be suitable for future Learning Analytics Summer Institute (LASI) workshops, exploring the potential for a SOLAR SIG application to support the training and collaboration of researchers in the area, exploring the interest in creating other community spaces (discussion fora, mailing lists, slack channels, etc.) and preparing several joint research summary statements based on break-out activities.

REFERENCES

- Biesta, Gert J. J. 2010. "Why 'what Works' Still Won't Work: From Evidence-Based Education to Value-Based Education." *Studies in Philosophy and Education* 29 (5): 491–503.
- Chow, Shein-Chung, and Mark Chang. 2008. "Adaptive Design Methods in Clinical Trials--a Review." *Orphanet Journal of Rare Diseases* 3 (1): 11.
- Dweck, Carol S. 2009. "Mindsets: Developing Talent through a Growth Mindset." *Olympic Coach* 21 (1): 4–7.
- Fussell, Sidney. 2018. "Pearson Embedded a 'Social-Psychological' Experiment in Students' Educational Software [Updated]." *Gizmodo*. *Gizmodo*. April 18, 2018. <https://gizmodo.com/pearson-embedded-a-social-psychological-experiment-in-s-1825367784>.
- Herold, Benjamin. 2018. "Pearson Tested 'Social-Psychological' Messages in Learning Software, With Mixed Results." *Education Week - Digital Education*. April 17, 2018. https://blogs.edweek.org/edweek/DigitalEducation/2018/04/pearson_growth_mindset_software.html.
- Kizilcec, René F., and Christopher Brooks. 2017. "Diverse Big Data and Randomized Field Experiments in Massive Open Online Courses: Opportunities for Advancing Learning Research." G. Siemens & C. Lang (eds.), *Handbook on Learning Analytics & Educational Data Mining*.
- Murphy, S. A. 2005. "An Experimental Design for the Development of Adaptive Treatment Strategies." *Statistics in Medicine* 24 (10): 1455–81.
- Nelson, Julie, and Carol Campbell. 2017. "Evidence-Informed Practice in Education: Meanings and Applications." *Educational Research* 59 (2): 127–35.
- Stamper, John C., Derek Lomas, Dixie Ching, Steve Ritter, Kenneth R. Koedinger, and Jonathan Steinhart. 2012. "The Rise of the Super Experiment." *International Educational Data Mining Society*, June. <http://files.eric.ed.gov/fulltext/ED537230.pdf>.
- Strauss, Valerie. 2018. "Pearson Conducts Experiment on Thousands of College Students without Their Knowledge." *The Washington Post*, April 23, 2018. <https://www.washingtonpost.com/news/answer-sheet/wp/2018/04/23/pearson-conducts-experiment-on-thousands-of-college-students-without-their-knowledge/>.

Connectivism: Using learning analytics to operationalize a research agenda

Srećko Joksimović

Teaching Innovation Unit and School of Education
University of South Australia
srecko.joksimovic@unisa.edu.au

George Siemens

Center for Change and Complexity in Learning
University of South Australia
LINK Research Lab
University of Texas at Arlington
gsiemenes@gmail.com

Shane Dawson

Teaching Innovation Unit
University of South Australia
shane.dawson@unisa.edu.au

Vitomir Kovanović

Teaching Innovation Unit and School of Education
University of South Australia
vitomir.kovanovic@unisa.edu.au

ABSTRACT: Given the rapid changes confronting society, important questions remain regarding how theory influence the work of researchers. Within learning and knowledge building literature, cognitivism and constructivism have remained the primary theories. Connectivism learning theory has more recently been posited and has been heavily cited over the past 15 years. Unfortunately, it has not been explored empirically. In this full day workshop, we utilize learning analytics methods and techniques to operationalize a research agenda. Specifically, we will explore the core assertions of connectivism with the goal of fostering a research community.

Keywords: Connectivism, learning theory, learning analytics

1 WORKSHOP BACKGROUND

In 2005, an article titled “Connectivism: A learning theory for the digital age” was published in the International Journal of Instructional Technology and Distance Education (Siemens, 2005). The article gained significant attention and remains one of the most cited articles in educational technology over the past 15 years. Google Scholar indicates over 15,000 mentions of connectivism and the original article has over 5,000 citations. Unfortunately, the core assertions of connectivism have not been tested empirically and a research community has failed to develop to meet this challenge. Most citations refer to the paper as a new model of learning and despite numerous

explorations in books, critical articles, and special issues, the theory has not advanced beyond a series of assertions that have not been tested. This proposed workshop will gather researchers to focus on planning a research agenda to evaluate, extend, and validate the suitability of connectivism as a model of learning.

1.1 Exploring the Connectedness

Connectivism recognizes three domains of connectedness (Siemens, 2005). Specifically, knowledge can be observed at the *neuronal* (i.e., biological) domain that observes brain processes that occur as a result of learning. The *conceptual* domain, on the other hand explores the means of forming connections between concepts, as a basic learning activity. Through the process of learning, students are constantly adding new knowledge or re-evaluate existing and develop novel forms of existing knowledge. Finally, the *social and technological* domain is focused on examining importance of social (i.e., students) and technological (e.g., social media and increasingly technology agents, bots, and machine learning/artificial intelligence) factors and understanding their role in creating networked knowledge.

Contemporary connectivist research, however, primarily focuses on the social and technological aspect of learning in networks (e.g., Paredes & Chung, 2012; Skrypnyk, Joksimović, Kovanović, Gasšević, & Dawson, 2015) or observes the theory as a whole (e.g., Ozturk, 2015), without discussing any of the particular domains. Existing research, thus, fails to account for specificities of knowledge creation at neuronal or conceptual levels. This nuance stems from the fact that the operationalization of social or technological aspects of networked learning is straight forward compared to other domains of networked knowledge (i.e., neuronal and conceptual). It is also noteworthy that the studies conducted to date, explored the connectivism across different domains, applying wide range of qualitative (e.g., Mackness, Waite, Roberts, & Lovegrove, 2013), quantitative (e.g., Joksimović et al., 2015), or mixed-methods analysis (Skrypnyk et al., 2015).

1.2 Challenges and Critique

Although connectivist research attained significant attention in recent years, the theory has also been a subject to criticism. Those critiques addressed ontological and epistemological status of connectivism, as well as psychological contents of connectivist assumptions. Questioning scientific rigour and novelty that connectivism brings, critics suggest that connectivism should be abandoned as a learning theory (Kop & Hill, 2008). Most of the researchers, however, agree that the idea of connectivism is influential in practice.

Observing connectivism from the ontological and epistemological aspect, Kop and Hill (2008), and later Kop (2011), questioned to what extent connectivism as a learning theory brings novelty to teaching in distributed settings, with respect to the existing theories. Giving connectivism a credit for recognizing a paradigm shift, and for playing an important role in developing new theories, Kop and Hill (2008) posit that connectivism should be observed at the curriculum and pedagogy level, rather than a learning theory on its own. As the main argument, Kop and Hill (2008) indicate that learning theory should be based on the existing body of research, whereas connectivism was not developed on the existing studies that use scientific methods. Moreover, it is questionable, Kop and Hill argue, to what extent connectivism was logically constructed to allow for verification through testing.

Besides arguing that most of the connectivist postulates, such as the relationship between individual and external knowledge or learning in networks, already existed in learning theories, Kop and Hill also question how connectivism explains understanding as well as the internal processes that lead to deep thinking and creating understanding. Finally, highlighting Downes (2005) interpretation of learning as recognizing patterns shaped by complex networks, Kop and Hill (2008) posit that connectivism fails to explain the existence of those patterns. That is, how one knows what the pattern is in the first place.

Clara and Barbera (2014), on the other hand, adopted psychological point of view in analyzing main postulates of connectivism. Specifically, Clara and Barbera (2014) agree with Kop and Hill (2008) and posit that connectivism does not provide an "adequate explanation to learning phenomena" (p. 131), as it fails to account for some crucial aspects of learning, such as learning paradox, interaction and dialog, as well as concept development. According to Clara and Barbera (2014), connectivist explanation of knowledge and learning entails an epistemological contradiction. On one hand, Clara and Barbera continue, connectivism argues that the knowledge is formed in the network and can reside in external appliances. Whereas, on the other hand, knowing consists of individually recognizing a pattern of connections. Clara and Barbera, thus, proposed two potential solutions: connectivism should either "abandon its current explanation of knowing as individually recognizing connective patterns" (p. 201), or "explain the (innate) mechanism that makes it possible to recognize a set of connections as a pattern" (p.201).

According to Clara and Barbera (2014), connectivism also underconceptualizes the interaction in distributed educational settings. That is, connectivism ignores some of the important aspects of how human nodes and their connections are operationalized in a network of learners. They further detail that peers (or other as noted in the study) are "regarded as one of the nodes of the connective pattern, that is, part of what is learned, an object of learning" (Clara & Barbera, 2014, p.202). However, speaking of educational interaction in general, Clara and Barbera explain that peers (meaning other nodes in a network) are usually understood as fundamental to learning (e.g., being identified as assistant for learning), and not just as objects of learning. Moreover, connectivism further understands interaction as the binomial conceptualization of nodes (i.e., connection either exists or not), which tends to be an oversimplification of the complexity of potential connection states in network of learners. Finally, Clara and Barbera argue that connectivism tends to ignore "procedural nature of interaction" (p.203), failing to account for the development of learning processes.

The final criticism in Clara and Barbera (2014) work related to the issue of concept development. As argued in those studies, connectivism does not provide an explanation how concepts (i.e., knowledge) develop. This problem has been identified as a major challenge to the connectivist conceptualization of learning. Specifically, Clara and Barbera pose a question:

"if a concept consists of a specific pattern of associations, how can it be explained that the concept develops but the pattern of associations remains the same?"(p.203).

1.3 Future of Connectivism - Learning Analytics in Shaping Research

These critiques indicate that there are numerous attributes of Connectivism as a learning and knowledge building theory that remain incomplete. This workshop therefore aims at bringing together researchers from learning sciences, learning analytics, psychology, and computer sciences to discuss the main postulates of the connectivist theory, providing basics for future rigorous research agenda. The need for the development of a connected theory of learning is increasingly important as knowledge continues to be distributed across humans and technology devices, raising the need for researchers to better understand how knowledge is generated in these spaces.

Connectivism is primarily concerned with connections. These connections occur neuronally, conceptually, or socially. Learning analytics has progressed significantly in methods to evaluate both conceptual and social connections but has only minimally evaluated neuronal connections. Our proposed workshop will launch a research community to focus on evaluating social, technical, and distributed knowledge building and how these dynamics differ for existing theories such as behaviorism, cognitivism, and constructivism.

2 PROPOSED ORGANIZATION AND AGENDA

This will be a full day workshop with the following agenda:

8:00-8:30	Introduction and objectives
8:30-9:00	Analysis of existing themes in connectivism literature
9:00-9:30	Exploration of related theories, including Actor Network, Activity Theory
9:30-10:00	Analysis of themes of connectivism, empirical research conducted to date
10:00-10:30	Break
10:30-12:00	Accepted presentations
12:00-1:00	Lunch
1:00-2:30	Accepted presentations
2:30-3:00	Break
3:00-4:30	Group sessions planning research directions
4:30-5:00	Wrap up, planning for next steps

An open call for presentations related to the theme of “Connectivism: Using learning analytics to operationalize a research agenda” will be organized in fall 2018. Papers will be solicited that address the current state of research related to networked learning, SNA/ENA, neurosciences of learning, cognitive functions that underpin connectivism, and related concepts. Given the theme of LAK19, the focus will be on broadening the diversity of voices represented. The workshop itself will be open to all LAK19 delegates.

We expect about 50+ attendees for the workshop given the extensive citations in learning literature and the need to move the conversation from conceptual to pragmatic and from theoretical to research-based. Papers will be solicited through prominent learning analytics mailing lists, on social media, and through personal networks of the individuals involved in organizing and hosting the workshop.

Technology requirements: speakers, projectors, and internet connectivity.

3 WORKSHOP/TUTORIAL OBJECTIVES OR INTENDED OUTCOMES:

This is the first workshop in the learning analytics community to explore connectivism. As such, the intended outcomes of this workshop are to help develop the formation of both a shared research community and models of operationalizing and assessing effectiveness and impact of connectivism.

Specific outcomes of the workshop include addressing the following challenges:

- What are the trends of connectivism research citations in academic literature?
- How does Connectivism relate to other prominent theories of learning, including Community of Inquiry, Actor Network Theory, and Activity Theory?
- How can the core assertions of Connectivism be evaluated? What type of research is needed?
- What is the role of Connectivism in the larger LA and learning sciences communities?
- Which methodologies provide researchers with the greatest capacity to understand and evaluate learning in a digital age?
- What types of mindsets and skills should learners develop regarding connectivism?

The outcomes will be disseminated through a proposed special issue of Journal for Learning Analytics.

REFERENCES

- Clara, M., & Barbera, E. (2014). Three problems with the connectivist conception of learning. *Journal of Computer Assisted Learning*, 30(3), 197–206. <https://doi.org/10.1111/jcal.12040>
- Downes, S. (2005). *An Introduction to Connective Knowledge*. Retrieved from www.downes.ca/post/33034
- Joksimović, S., Dowell, N., Skrypnik, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2015). How do you connect? Analysis of Social Capital Accumulation in connectivist MOOCs (pp. 64–68). Presented at the Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK'15), New York, USA: ACM. <https://doi.org/10.1145/2723576.2723604>
- Kop, R. (2011). The Challenges to Connectivist Learning on Open Online Networks: Learning Experiences during a Massive Open Online Course. *The International Review of Research in Open and Distributed Learning*, 12(3), 19–38.
- Kop, R., & Hill, A. (2008). Connectivism: Learning Theory of the Future or Vestige of the Past?. *International Review of Research in Open and Distance Learning*, 9(3), 1–13.
- Mackness, J., Waite, M., Roberts, G., & Lovegrove, E. (2013). Learning in a Small, Task-Oriented, Connectivist MOOC: Pedagogical Issues and Implications for Higher Education. *International Review of Research in Open & Distance Learning*, 14(4), 140–159.
- Ozturk, H. T. (2015). Examining Value Change in MOOCs in the Scope of Connectivism and Open Educational Resources Movement. *International Review of Research in Open & Distance Learning*, 16(5), 1–25.
- Paredes, W. C., & Chung, K. S. K. (2012). Modelling learning & performance: A social networks perspective. In *ACM International Conference Proceeding Series* (pp. 34–42). <https://doi.org/10.1145/2330601.2330617>
- Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, 2(1), 3–10.
- Skrypnik, O., Joksimović, S., Kovanović, V., Gašević, D., & Dawson, S. (2015). Roles of course facilitators, learners, and technology in the flow of information of a CMOOC. *International Review of Research in Open and Distance Learning*, 16(3), 188–217.

Lifespan of an Idea

Sai Santosh Sasank Peri

LINK Research Lab
University of Texas at Arlington
saisantoshsasank.peri@mavs.uta.edu

Angela Liegey Dougall

Department of Psychology
University of Texas at Arlington
adougall@uta.edu

George Siemens

LINK Research Lab
University of Texas at Arlington
gsiemens@uta.edu

ABSTRACT: This paper presents an ongoing study on how ideas evolve, develop, and survive in learning networks and knowledge spaces. Ideas can be transmitted from one individual to another in a model similar to the dynamics of disease spreading over the population, reflective of underlying network structure, which can be studied using epidemic models. This work provides a unique framework for evaluating learning and knowledge creation among learners through a combination of Connectivism, Knowledge Building, and Epidemiological theories. We present our analysis methods, preliminary results, and planned future research building on our framework. We anticipate that these results will be useful to study idea development, creativity, and knowledge development in digital learning environments.

Keywords: Ideas, Epidemiology, Networked Learning, Knowledge Creation, Connectivism

1 INTRODUCTION

Ideas are entities containing information that is transmitted from one individual to another through personal interaction, similar to a contagion (Goffman & Newill, 1964). This transmission can be considered as an epidemic process. A study of this transmission dynamic, from its origin to spread until survival, is called a “Lifespan of an Idea”. Idea generation in groups through activities like brainstorming (Paulus & Yang, 2000) has been extensively studied in cognitive psychology. Attributes like creativity, prior knowledge, and unconscious thoughts (Ritter, van Baaren, & Dijksterhuis, 2012) also play a role in idea generation. While these studies focus on the psychological processes of idea generation, which lead to a better understanding of individual human brains, an epidemiological approach offers an understanding of the behavioral characteristics of a population when they encounter an idea. This has been supported by research on social contagion processes, such as emotion and behavior through populations utilizing the principles of epidemiology, social reinforcement and association of strong ties in network threshold theory (Christakis & Fowler, 2010; Centola, 2010).

Other network theories of idea spread (e.g., Woo & Chen, 2016) are specific to a particular domain such as diffusion of innovations for adoption of technology, word-of-mouth effects for business and marketing, opinion dynamics for the spread of news, and virality prediction to estimate influence and popularity of people and topics. Epidemiological models have proved to be a valuable approach in studying the spread of scientific ideas using published literature as infectious material (Bettencourt et al., 2006; Kiss, Broom, Craze, & Rafols, 2010).

The flow of knowledge through a population and its underlying network structure can be evaluated to predict the likelihood of an existing or novel scientific idea spreading. In the process of developing an idea, an essential learning and knowledge building activity will generally originate with an individual and is then transmitted to others in a network. As the idea moves through a network, it undergoes a process of assessment, revision due to the contributions of others, and evolution. Therefore, the epidemiological modeling of ideas in education resonates with learning theories of Connectivism (Siemens, 2014) and Knowledge Building (Bereiter & Scardamalia, 2014) where learning is considered as a network forming process and ideas are seen as the basis for knowledge creation. This novel combination of principles of epidemiological theories with learning theory forms the basis of our study regarding how ideas flow through digital networked learning space (like MOOCs, StackExchange) and identify the “contagious” ideas that are created, re-created, improved, spread and rejected by learners.

We assess multiple questions as part of our broader research work, such as the attributes of idea and structural design of the learning space. This workshop paper specifically addresses how does an idea originates, spreads, and is sustained and saturated. We describe our methodology of applying epidemic models to study ideas propagated by learners in an online question & answer (Q&A) community. We provide some illustrations to show the evolution of an idea over time.

2 PRELIMINARY RESULTS AND PLANNED FUTURE RESEARCH

Epidemic models can be represented as a homogenous mixture of individuals experiencing random interactions and divided into compartments reflecting their status based on a set of differential equations. These compartments are susceptible (S), infective (I) or removed (R) (Kermack & Mckendrick, 1927). They can also be represented as a network structure reflecting the properties of individuals in the network involved in the spread of disease. Previous studies have adopted these compartmental divisions and redefined them appropriately for transmission of ideas (Goffman, 1966). Briefly, individuals who are likely to get influenced by, or adopt, the idea are *susceptible*. Individuals who are in contact with the idea and can transmit it others are *infective*. Individuals who are no longer in contact with the idea and attain immunity are *removed*. Studies also used modified versions of the epidemic model and a weighted network model to estimate the spread through population (Bettencourt et al., 2006; Kiss, Broom, Craze, & Rafols, 2010).

We will use a publicly available dataset of an online learning environment called Data Science Exchange, which is a part of the Stack Exchange Q&A communities. The dataset consists of details of every question posted on the community since May 2014 to September 2018. Each question is usually identified with one or more keywords called a “tag” to classify the question category according to a relevant topic in the field of data science. Tags are used to identify the idea of the

posted question and attract learners specifically interested in this idea to provide a response and ask more questions. This process can be infectious when the number of questions related to a tag is in higher volume compared to other tags in the community. In a Q&A learning environment, the number of possible questions that might mention a tag at a time t are *susceptible*, S . Number of questions posted with a tag during the same period are *infective*, I . Number of questions that lose infectivity to other questions on the tag are *removed*, R . The rate at which a tag gets mentioned in questions is the rate of infection, α . The rate at which the tag occurrence in questions reduces is the rate of recovery, β . Since it is an online forum, there is a possibility of variation in total population over time called as rate of population growth, μ . As per logistic growth, the population grows based on upper limit to the population size called carrying capacity, K . As the incoming questions are initially susceptible, logistic growth term is added to the *susceptible* compartment. The total population, N at any time is the sum of S , I and R compartments. The interaction between these compartments and parameters can be expressed using a set of differential equations corresponding to a general epidemic SIR model, as given in equation (1).

$$\begin{aligned} dS/dt &= \mu(1-N/K)N - \alpha SI \\ dI/dt &= \alpha SI - \beta I \\ dR/dt &= \beta I \\ N &= S + I + R \end{aligned} \quad (1)$$

We will evaluate the proposed SIR model by fitting the data. Using Non-linear Least Squares method we solve the differential equations to refine the model parameters in successive iterations such that error between actual and estimated value is minimized. Simulated Annealing and Genetic Algorithm are employed for optimization of parameter estimation. The SIR model is evaluated for goodness of fit with mean square error and R square values. These optimized parameters allow for forecasting the growth of an idea in the future. Figure 1 illustrates occurrence of tags “machine-learning”, “Python” and “R” from 2014 to 2018 based on their occurrence where, population (y-axis) represents the number of times a particular tag was mentioned at a given time point (x-axis).

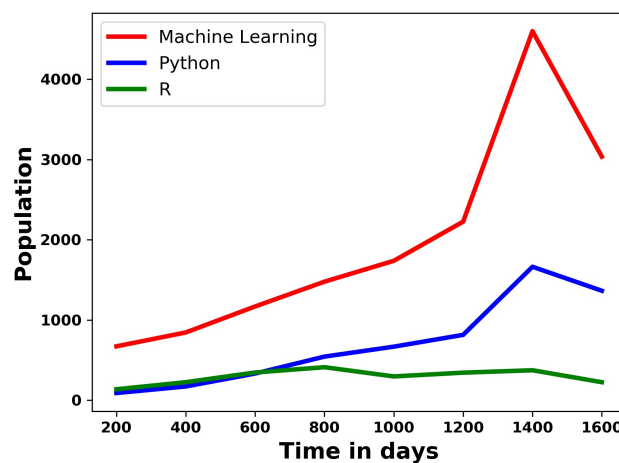


Figure 1: Comparison between tags “machine-learning”, “Python” and “R” from 2014 to 2018 based on their occurrence.

According to our assumptions in SIR model, Figure 1 represents the infective class for each tag, i.e. at day 600, there were about 1000 questions infected with tag “machine-learning”. The removed class (not pictured) consists of all the questions that lose effectivity after a certain time (here we assume 200 days), so about 900 questions. The susceptible class strength is deduced by using the relation of ‘N’, the total population from eq 1. The total population increases with time, as the online forum is an open platform allowing new users to post questions with tags leading to a different N value at each time point.

At present, our work is focused on fitting SIR epidemic model to the tag data from Data Science Exchange forum. We will utilize the planned research methods to estimate parameters for forecasting the rise and fall of trends in tag data to showcase which ideas are contagious among learners in the environment. Identification of these contagious ideas is the first building block of our framework to study idea development in digital learning environments. More specifically, the use of SIR model will be extended to include more sophisticated knowledge generation processes. In this paper, we presented a simple analysis of how programming languages, through the use of hashtags, trended over a period of time. Analyzing how complex ideas develop and evolve, for example, when a small group is attempting to solve a problem, is the next planned stage of our work.

REFERENCES

- Goffman, W., & Newill, V. A. (1964). Generalization of epidemic theory. *Nature*, 204(4955), 225-228.
- Paulus, P. B., & Yang, H.-C. (2000). Idea Generation in Groups: A Basis for Creativity in Organizations. *Organizational Behavior and Human Decision Processes*, 82(1), 76–87. <https://doi.org/10.1006/obhd.2000.2888>
- Christakis, N. A., & Fowler, J. H. (2010). Social Network Sensors for Early Detection of Contagious Outbreaks. *PLOS ONE*, 5(9), e12948. <https://doi.org/10.1371/journal.pone.0012948>
- Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996), 1194–1197. <https://doi.org/10.1126/science.1185231>
- Woo, J., Ha, S. H., & Chen, H. (2016). Tracing Topic Discussions with The Event-Driven Sir Model for Online Forums. *Journal of Electronic Commerce Research*, 17(2), 169.
- Goffman, W. (1966, October). Mathematical Approach to the Spread of Scientific Ideas.pdf. Retrieved from <https://www.nature.com/articles/212449a0>
- Kermack and McKendrick. 1927. *Proc. R. Soc. LISA*, 700.
- Bettencourt LM, Cintrón-Arias A, Kaiser DI, Castillo-Chávez C. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*. 2006 May 15;364:513-36.
- Kiss, I. Z., Broom, M., Craze, P. G., & Rafols, I. (2010). Can epidemic models describe the diffusion of topics across disciplines? *Journal of Informetrics*, 4(1), 74–82.
- Siemens, G. (2014). *Connectivism: A learning theory for the digital age*.
- Bereiter, C., & Scardamalia, M. (2014). Knowledge building and knowledge creation: One concept, two hills to climb. In *Knowledge creation in education* (pp. 35-52). Springer, Singapore.

Deconstructing the Evolution of Collaborative Learning Networks

Renzhe Yu

University of California, Irvine

renzhey@uci.edu

ABSTRACT: As social interaction becomes an integral component in online learning environments, analyzing the dynamic evolution of peer learning networks is necessary to better understand and support learners in these contexts. This study investigates a unique network of collaborative artifact composition within a college-level online course, focused on the co-evolution of this network and student engagement at the individual level. Using stochastic actor-oriented models (SAOM), I find that students tend to form cohesive subgroups but not to produce “super stars” in collaboration activities. Moreover, collaboration exerts peer influence on individual course engagement, but there is no evidence of engagement-based selection of collaborators. These identified trends can help the instructor(s) refine their course design and implement appropriate intervention to foster more effective learning communities.

Keywords: Social Learning Analytics; Connectivism; Learning Networks; Collaborative Composition; SuiteC; SAOM

3 BACKGROUND

Learning theories from earlier social constructivism to more recent connectivism have highlighted the role of social interaction in human learning (Siemens, 2005; Vygotsky, 1978). In these theories, learning occurs when people as nodes of knowledge make connections and knowledge flows within the interpersonal network. Empirically, research that employs social network analysis to examine online peer interaction partially justifies the theory of connectivism (e.g. Cho, Gay, Davidson, & Ingraffea, 2007; Dawson, 2008; Joksimović et al., 2016; Wang & Noe, 2010). However, most of these studies analyze the final network generated throughout the course period without attending to the dynamics of information flow and network changes, which is a central theme of connectivism. As such, analyzing the evolution of learning networks will add new insights to the understanding of peer interaction.

Towards this end, a handful of recent studies have leveraged statistical models of network dynamics to understand the temporal dependencies of learning network structures (Joksimović et al., 2016; Poquet, Dowell, Brooks, & Dawson, 2018; Stepanyan, Borau, & Ullrich, 2010; Zhang, Skryabin, & Song, 2016). Across these studies, reciprocity, individual performance and performance-based homophily consistently contribute to the formation of learning ties, while hierarchical structures including triad closure, preferential attachment and Simmelian ties are not always present. These studies are largely concentrated on discussion forums in MOOCs and may not generalize to other learning networks. To fill this void, the current study delves into the dynamics of artifact composition networks in formal higher education settings. It also traces the co-evolution of network structures and individual learning behavior, thus differentiating the underlying processes of influence and selection (Lewis, Gonzalez, & Kaufman, 2012) in peer learning environments.

4 FRAMING OF THE STUDY

This study takes advantage of SuiteC, a specially designed set of student-centered learning tools embedded within the Canvas learning management system (LMS). Partially informed by connectivism, this toolkit facilitates sharing, discussing and remixing student-contributed artifacts via three interconnected apps: Asset Library is a repository of such artifacts (a.k.a. assets) with rich social networking functions; Whiteboards is a platform for real-time collaboration on remixing assets; Engagement Index introduces a leaderboard to create a gamified vibe. (Jayaprakash, Scott, & Kerschen, 2017).

SuiteC enables more closely connected learning experience than traditional online learning environments. It is then meaningful to investigate how learning networks develop within this augmented system. As an exploratory step, this study delves into the learning network formed through collaborative composition in the Whiteboards (referred to as “whiteboard network”). This network differs substantially from a discussion network because the former engages learners in a process of working together towards a certain target while the latter involves direct and short communication between learners (Liu, Chen, & Tai, 2017).

In this context, I propose the following research questions:

1. What are the network structural properties (e.g. reciprocity, homophily) that characterize students’ collaboration in the Whiteboards over time?
2. Do collaborators exhibit similar levels of course engagement over time, or do students tend to collaborate with peers who have similar levels of engagement?

5 DATA AND METHODS

5.1 Dataset

The dataset comes from a fully online course offered to residential students of a four-year university in the US. The course was offered in Spring 2016 and lasted for 14 weeks. Each week students were required to share assets and interact with peer assets around the topic of that week. They were also required to collaborate on composing one or more whiteboards that feature the same topic.

All the actions within SuiteC apps were recorded, with a total count of 658,967. These actions were taken by 114 users and involved 1,366 whiteboards and 6,672 assets.

5.2 Modeling Strategy

Stochastic actor-based models (SAOM) were used to study the co-evolution of the whiteboard network and course engagement. This model family basically assumes that changes of network ties result from micro-level decisions of individual actors (nodes) decisions that maximize their current network function. When time-variant individual behaviors come into play, individual actors decide their behaviors by maximizing their behavior function. In the context of SuiteC, these assumptions seem reasonable and not very restrictive.

The whiteboard network was defined as a non-directed network among individual students, which resembles a co-authorship network. Engagement was originally defined for each learner as her total number of actions. For modeling purposes, the data were further transformed in two manners. First, the 14 weeks were divided into 4 periods based on the topic structure and a network was constructed for each period. Second, engagement values were first calculated within each period and then converted to a categorical variable with five levels.

To model the dynamic interplay between collaborative composition and engagement, network and behavior functions were used. The network function modeled local structures and attributes that contributed to the presence of a collaboration tie over time, including density (base effect), triangle, nodal degree, individual engagement and dyadic engagement similarity (Ripley, Snijders, Boda, Vörös, & Preciado, 2018). The behavior function, by contrast, modeled factors that influence observed behavior (engagement), including linear and quadratic terms of engagement and the average engagement similarity between a focal student and her collaborators.

6 RESULTS

Table 1 reports summary statistics of the whiteboard network across the four periods. Network density ranges from 0.02 to 0.04; it slightly moves up from period 1 to 2 before dropping heavily and then recovering through periods 3 and 4. On average, each student collaborates with two to three other students on composing whiteboards during each period. The Jaccard coefficients of the three transitions (not reported) are all above 0.3, a recommended threshold for applying SAOM.

Table 1: Summary statistics of the whiteboard network across four periods.

Period	1	2	3	4
Density	0.035	0.042	0.021	0.028
Average degree	2.843	3.422	1.735	2.313
Number of ties	118	142	72	96

Table 2 reports the estimated effects of function terms. Model 1 solely takes into account the evolution of whiteboard network (RQ 1), while Model 2 adds its interplay with course engagement (RQ 1 and RQ 2). In terms of network structures, the triangle effect is significantly positive whether engagement is incorporated or not, meaning that, if two students have both collaborated with the same third student, they are more likely to work together. By contrast, the significantly negative degree effect suggests that a student who already has multiple collaborators is less likely to collaborate with more peers. These effects combined suggest a tendency to form cohesive subgroups and to participate equally.

In Model 2, the engagement and engagement similarity effects on the whiteboard network are not significant. In other words, refusing any difference in the likelihood of pairwise collaboration for different combinations of engagement levels. By contrast, the average similarity effect on engagement is strongly positive. In other words, students tend to engage as much as their peers with whom they have collaborated. These results provide evidence for peer influence but against

peer selection, i.e. students being assimilated to their collaborators, instead of similar students being attracted to work together.

Table 2: Estimated effects of the network function and the behavior function.

Effect	Model	
	(1)	(2)
Whiteboard network		
<i>Rate</i>		
Period 1	3.539*** (0.602)	3.323*** (0.628)
Period 2	2.836*** (0.530)	2.982*** (0.566)
Period 3	1.532*** (0.289)	1.584*** (0.298)
<i>Structural</i>		
Density	-1.577*** (0.260)	-1.556 (0.268)
Triangle	1.782*** (0.244)	1.801*** (0.237)
Degree	-0.209*** (0.074)	-0.227*** (0.077)
<i>Covariate</i>		
Engagement		0.222 (0.149)
Similarity of engagement		1.190 (1.370)
Engagement		
<i>Rate</i>		
Period 1		3.807*** (0.834)
Period 2		34.168*** (11.202)
Period 3		10.427** (4.490)
<i>Behavior</i>		
Engagement linear		-0.100*** (0.032)
Engagement quadratic		0.109*** (0.029)
<i>Network</i>		
Average similarity of engagement		1.945*** (0.705)

Note: Standard errors reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

7 DISCUSSIONS

This paper reveals that cohesive subgrouping and equal participation are characterizing structures of students' collaborative composition network. It also finds that while students' general course engagement is influenced by their whiteboard collaborators, students who engage in the course environment to a similar extent are no more likely to collaborate on whiteboards than if they are different. These findings have implications both for social learning analytics researchers and for online learning practitioners. For one thing, research efforts should delve into the dynamic interplay between structures of learning networks and low-level learner behaviors in networked learning environments. Also, the artifact composition network exhibits more desirable structures than discussion networks, so online instructors may consider collaborative tasks more often when they intend to leverage the benefits of social interactions to foster student learning.

REFERENCES

Cho, H., Gay, G., Davidson, B., & Ingraffea, A. (2007). Social networks, communication styles, and learning performance in a CSCL community. *Computers & Education*, 49(2), 309–329.

- Dawson, S. (2008). A study of the relationship between student social networks and sense of community. *Educational Technology & Society*, 11(3), 224–238.
- Jayaprakash, S. M., Scott, J. M., & Kerschen, P. (2017). Connectivist Learning Using SuiteC - Create, Connect, Collaborate, Compete! In *Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK17)* (pp. 69–76). Vancouver, BC, Canada.
- Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., & de Kereki, I. F. (2016). Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)* (pp. 314–323). Edinburgh, United Kingdom: ACM.
- Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America*, 109(1), 68–72.
- Liu, C.-C., Chen, Y.-C., & Tai, S.-J. D. (2017). A social network analysis on elementary student engagement in the networked creation community. *Computers & Education*, 115, 114–125.
- Poquet, O., Dowell, N., Brooks, C., & Dawson, S. (2018). Are MOOC forums changing? In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18* (pp. 340–349). New York, New York, USA: ACM Press.
- Ripley, R. M., Snijders, T. A. B., Boda, Z., Vörös, A., & Preciado, P. (2018). Manual for SIENA version 4.0. Oxford, United Kingdom: University of Oxford, Department of Statistics; Nuffield College.
- Siemens, G. (2005). Connectivism : A Learning Theory for the Digital Age. *International Journal of Instructional Technology and Distance Learning*, 2(1), 1–7.
- Stepanyan, K., Borau, K., & Ullrich, C. (2010). A Social Network Analysis Perspective on Student Interaction within the Twitter Microblogging Environment. In *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 70–72). IEEE.
- Vygotsky, L. S. (1978). Interaction between Learning and Development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in Society: Development of Higher Psychological Processes* (pp. 71–91). Cambridge, MA, USA: Harvard University Press.
- Wang, S., & Noe, R. A. (2010). Knowledge sharing: A review and directions for future research. *Human Resource Management Review*, 20(2), 115–131.
- Zhang, J., Skryabin, M., & Song, X. (2016). Understanding the dynamics of MOOC discussion forums with simulation investigation for empirical network analysis (SIENA). *Distance Education*, 37(3), 270–286.

Multilevel Learning Behavior Analysis Method in Connectivist Learning Contexts

Wang Zhijun

Research Center of Educational Informatization, Jiangnan University 214122 Jiangsu Wuxi

Abstract: Connectivist learning is a learning centered on interaction and knowledge creation in the open complex information network environment. Network building and continuous knowledge creation are the two main goals of connectivist learning. Traditional learning behaviors analysis method can hardly to match the above two goals. Based on Connectivism, cMOOC and learning analytics, this study proposes that learning behavior in connectivist learning contexts should be analyzed from the perspectives of complex collective and independent subject learning behavior in the complex system. The “trinity of methods” includes analytics of 1) network structures including cognitive network, conceptual network, social network, and technical network; 2) content-oriented analysis including operational interaction, wayfinding interaction, sensemaking interaction, and innovative interaction, 3) process-oriented analysis including the evolve and interaction process analysis of the above four levels of networks and four types of interactions.

Keywords: Connectivist learning; learning analytics; multidimensional analysis; network analysis; content analysis; process analysis

1 INTRODUCTION

The development of Web2.0, social media and AI promoted our learning into a complex information networked era which the teaching and learning ecosystem is being reconstructed (Chen, 2016), the half-life of knowledge is shortened, the speed of knowledge change is increasing, and the way our learning is also being changed. Connectivism (Siemens, 2005a), proposed two completely different learning objectives from traditional learning. It centers on interaction and knowledge creation (Siemens, 2011), emphasizing learning as network creation (Siemens, 2005b). The analysis of learning behavior in a connectivist learning context is quite different from traditional learning. Existing research methods are difficult to explore this kind of learning comprehensively.

2 LITERATURE REVIEW

Connectionism was proposed based on chaos theory, self-organization theory, complex theory, and network theory, as a new learning theory (Siemens, 2005a), and it is the first theory to face directly with the complexity of learning. The connectivist learning encourages learners to create learning artifacts together by interacting with social media and knowledge, and learning resources sharing and created by more learners inside and outside the curriculum. There are two main parallel goals of connectivist learning: networks building, and knowledge growth and creation. cMOOCs are Massive open online courses that have been created and run in 2007 with the ideas of Connectivism. In cMOOCs, learning take place on a range of dedicated online learning applications as well as social media and networking applications for sharing information and resources among learners (Siemens, 2005). Fournier et al. (2014) pointed out that there are a large number of incomplete and distributed

data sets in cMOOCs which are more critical and maybe more in-depth than the data in the cMOOCs forum. It also brings a lot of challenges to researchers.

For the development of learning analytics, researches have been analyzed a series of qualitative and quantitative resources in cMOOCs (Fournier & Kop, 2015; Fournier, Kop, & Durand, 2014)). But these analyses also highlight the most prominent problems in this type of research (Kop, Fournier & Durand, 2017), that is, these MOOCs provide a larger data set than previous studies with educational data mining and learning analytics. Data tools are visualized as powerful tools like digital social networks, but the results of visualizations bring more questions than answers. Educational data mining and learning analytics cannot explain and answer the complexity of learning and its process, which push us to analyze these data qualitatively. We need to build a systematical research method to conduct in-depth research on this kind of complex learning.

3 THE THEORETICAL FOUNDATION FOR METHOD CONSTRUCTION

3.1 Actor-network theory (ANT)

The "symmetry", "translation", "network" and "network effect" emphasized in ANT (Fenwick & Edwards, 2010) are similarities with the Connectivism, and have important implications the Connectivism (Wang, 2017), such as (1) The principle of symmetry can guide the analysis of interaction; (2) Tracking the network formed by the actors is the basic step of learning behavior analysis; (3) It is important to find the obligatory points of passage in the formation of network and the knowledge flow; (4) the spatiality and temporality of the network should be analysed; (5) Both relationship thinking and process thinking the researcher should behave in doing research on connectivism; (6) Network action research can be take in research.

3.2 Connectivist Interaction and Engagement framework (CIE)

Connectivist Interaction and Engagement (CIE) Framework was constructed (Wang et al., 2014, Wang, Anderson, Chen, Barbera, 2017) to reveal the interaction process from cognition engagement of participants. It divides the interaction in connectivist learning into four levels: operational interaction, wayfinding interaction, sensemaking interaction, and innovation interaction. The interaction pattern in each level of interaction is also revealed. (Wang, Anderson, Chen & Barbera, 2017), it can evaluate the interaction quality of this kind of learning by content analysis.

3.3 Complex Systems Conceptual Framework of Learning (CSCFL)

Education is a complex system, the classical research methods treat the education as a linear, simple system and ignore the complex, diversity, and non-linear characteristic of education. The Complex Systems Conceptual Framework of Learning (CSCFL) (Jacobson, 2016) pointed out that learning occurs in complex systems with elements or agents at different levels-including neuronal, cognitive, intrapersonal, interpersonal, cultural-in which there are feedback interactions within and across levels of the systems so that collective properties arise (i.e., emerge) from the behaviors of the parts, often with properties that are not individually exhibited by those parts. The eight core concepts and characteristics of CSCFL at two levels are important dimensions and perspectives for guiding the analysis of connectivist learning behaviors.

4 TRINITY MULTILEVEL ANALYSIS METHOD

The founder of Connectivism proposed that the research methodology of Connectivism is mixed research method focused on the evaluation of (social, informative) connections occurs neuronally, conceptually, and socially (Siemens, 2011, p. 57). The author proposed that the process of knowledge creation and network evolution is also important. The evolution of learning over time and the switching and expansion of space also cannot be ignored in this kind of learning.

4.1 The conceptual framework of connectivist learning behavior analysis

In the development of learning analytics, Hoppe (2017) proposed the trinity of methodological approaches for the analysis of learning and knowledge building communities. The trinity of methods includes analytics of 1) network structures including actor-actor (social) networks but also an actor-artifact networks, 2) processes using methods of sequence analysis, and 3) content using text mining or other techniques of artifact analysis. According to the “trinity of methods”, combined with the ANT, CIE and CSCFL theories, this study proposes a conceptual framework of connectivist learning behavior analysis, as figure 1. This study states that the analysis of the learning behavior in the connectivist learning context should be conducted from the network, content and process aspects with the guidance of ANT, CIE, and CSCFL. Only by this, can we get the unique feature of learning behaviors in a connectivist learning context and a multi-dimensional understanding of Connectivism.

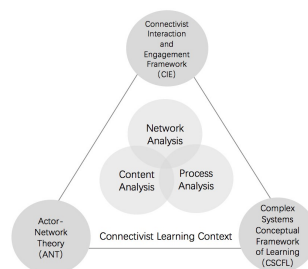


Figure 1 The conceptual framework of connectivist learning behavior analysis

4.2 The “trinity” of method for connectivist learning behavior analysis

The CIE framework combines the quality of interaction, the process of knowledge creation and the evolution of network in connectivist learning together and can be used in network, content and process analysis. The ANT emphasizes the symmetry of the subject and the object. It is necessary to track the actors, to find the obligatory points of passage in the network, to consider the temporal and spatial characteristics, and the relationship and process thinking in research. The two levels of the CSCFL highlight the collective and the individual perspectives of the analysis. The eight characteristics provide important guidance for trinity of method. Connectivist learning is supported by various technologies. The connection between technologies is actually an important part for social connections. Therefore the technical network also needs to be analyzed. So the network analysis includes four levels: technology network, social network, concept network, and neural network. The content analysis mainly refers to the four levels of operational interaction, wayfinding interaction, sensemaking interaction, and innovation interaction. In terms of process analysis refers to the evolution process of four types of networks, and four levels of interaction, and the interaction

among these 8 elements. Based on the above considerations, the connectivist learning behavior method is proposed as figure 2.

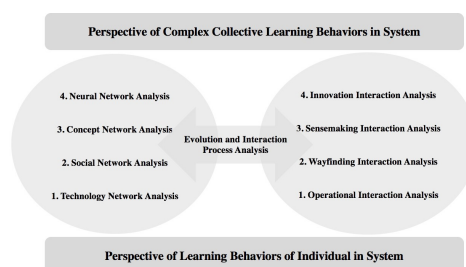


Figure 2 The multi-level learning behavior analysis in connectivist learning contexts

5 SUMMARY AND DISCUSSION

The analysis of the connectivist learning behavior should be conducted from the complex systems with macroscopic analysis methods to form a stereoscopic and comprehensive understanding of this learning. The Trinity Multilevel Analysis Method of Connectivist Learning Behavior is proposed by this research. It is a macroscopic and comprehensive analysis method. However, the analysis is dependent on the development of many other disciplines, such as brain science, cognitive science, and learning analytical tools which can collect and link the distributed dataset together (Zouaq, Jovanovic, Joksimovic, & Gasevic, 2017) and visualize them. The detail of analysis for each specific dimension will be discussed in detail further in a future article.

REFERENCE

- Fenwick, T. & Edwards, R. (2010). Actor-Network Theory in Education. *Oxen: Routledge*. Preface 2.
- Fournier, H., & Kop, R. (2015). MOOC learning experience design: Issues and challenges. *International Journal on E-Learning*, 14(3), 289–304.
- Fournier, H., Kop, R., & Durand, G. (2014). Challenges to research in MOOCs. *MERLOT Journal of Online Learning and Teaching*, 10(1), 1-15.
- Hoppe, U. (2017). Computational Methods for the Analysis of Learning and Knowledge Building Communities. *In Handbook of Learning Analytics*. 23-33
- Jacobson, M. J., Kapur, M., & Reimann, P. (2016). Conceptualizing debates in learning and educational research: Toward a complex systems conceptual framework of learning. *Educational Psychologist*, 51(2), 210-218.
- Kop, R., Fournier, H., & Durand, G. (2017). A Critical Perspective on Learning Analytics and Educational Data Mining. *In Handbook of Learning Analytics*, 319.
- Siemens, G. (2005a). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, 2, 1, 3-10.
- Siemens, G. (2005b). Connectivism: Learning as network-creation. Retrieved Dec, 10, 2018, from <http://www.elearnspace.org/Articles/networks.htm>.
- Siemens, G. (2011). Orientation: Sensemaking and wayfinding in complex distributed online information environments. Unpublished doctoral dissertation, *University of Aberdeen*.
- Wang, Z., Anderson, T., Chen, L. and Barbera, E. (2017). Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework. *British Journal of Educational Technology*. 48(2):683–699
- Wang, Z., Chen, L., Anderson, T. (2014). A Framework for Interaction and Cognitive Engagement in Connectivist Learning Contexts. *International Review of Research in Open & Distance Learning*, 15(2), 121-141.
- Wang, Z. (2017). A New Research Perspective on Instructional Interaction in Connectivist Learning: Actor Network Theory. *Modern Distance Education Research*, (6):28-36.
- Zouaq, A., Jovanovic, J., Joksimovic, S., & Gasevic, D. (2017). Linked Data for Learning Analytics: Potentials and Challenges. *In Handbook of Learning Analytics*, 347-355.

VISLA: Visual Approaches to Learning Analytics

Katrien Verbert, Robin De Croon, Tinne De Laet, Tom Broos, Martijn Millecamp

KU Leuven, Belgium
{firstname.lastname}@kuleuven.be

Xavier Ochoa

New York University, USA
xavier.ochoa@nyu.edu

Robert Bodily

Brigham Young University, USA
bodilyrobert@gmail.com

Judy Kay

The University of Sydney, Australia
judy.kay@sydney.edu.au

Hendrik Drachsler

University of Frankfurt, Germany
drachsler@dipf.de

Cristina Conati

University of British Columbia, Canada
conati@cs.ubc.ca

ABSTRACT: One of the most visible tools used in Learning Analytics is the dashboard. These dashboards use a wide range of visualization techniques to explore and understand relevant user traces that are collected in various (online) environments and to improve (human) learning. The design and evaluation of learning analytics dashboards within the educational practice does not receive enough research attention. The goal of our workshop is to build a strong research capacity around visual approaches to learning analytics. The longer-term goal is to improve the quality of learning analytics research that relies on information visualization techniques. This proposal describes the goal and activities of the VISLA 2019 workshop on Visual Approaches to Learning Analytics.

Keywords: learning analytics dashboards, visualization, HCI

1 THEME AND WORKSHOP BACKGROUND

In recent years, many learning analytics dashboards have been deployed to support insight into learning data. The objectives of these dashboards include providing feedback on learning activities, supporting motivation, and reducing dropout. Several visualization techniques have been used in learning analytics dashboards to help teachers, learners, and other stakeholders explore and

understand relevant user traces collected in various (online) environments. The overall objective is to improve (human) learning.

As important as they are for learning analytics, the design and evaluation of dashboards for the educational practice have, in the most part, being ad-hoc processes that limit their impact and preclude the collection and sharing of best practices. Moreover, the lack of rigorous research focus on the selection and use of different visualization techniques for different types of data lead to sub-optimal, and sometimes harmful, designs. Research has also shown that many learning dashboards are often deployed without conducting usability tests (Bodily and Verbert, 2017): this is detrimental to the research field as a lack of usability could be the reason why students do not like or use dashboards. The goal of our workshop is to build a strong research capacity around visual approaches to learning analytics. The longer-term goal is to improve the quality of learning analytics research that relies on information visualization techniques.

Authors were invited to submit original unpublished work. To facilitate comparison and generalization, all submissions were organized according to the following questions (Klerkx et al., 2015): 1) What kind of data is being visualized? What tools were used to clean up the data (if any)? 2) For whom (learner, teacher, manager, researcher, other) is the visualization intended? 3) Why: what is the goal of the visualization? 4) How is data visualized and why? Which interaction techniques are applied? What tools, libraries, data formats, etc. can be used for the technical implementation? What workflows and recipes can be used to develop the visualization? 5) How has the approach been evaluated or how could it be evaluated? 6) What were the encountered problems and pitfalls during the visualization process?

2 ORGANIZATIONAL DETAILS

During our 1-day workshop, we aim to facilitate a very interactive and engaging event where we want to avoid death by powerpoint at all causes and promote discussion activities over presentational ones. In the first half of the workshop, we will, therefore, ask participants to shortly present the work of *another* submission and to relate it back to their *own* work. The facilitators can potentially allocate challengers per presentation to move the discussion around common themes and differences in approaches.

During the second half of the workshop, we invite the participants to share their tools, workflows, and recipes in a hands-on discussion session so that they can benefit from each others' knowledge, apply their visual approaches on either their own dataset or on the dataset that we provide.

3 CONTRIBUTIONS

We accepted six submissions and invited Prof. Sharon Hsiao (Arizona State University, USA) to give an invited talk, entitled "Visual Learning Analytics in Computing Education."

The first group of papers focus on visualizing learner data to assist educators to identify trends and anomalies or to gain insights about students and their learning progress. Different visualization techniques are used to achieve these objectives. Rohloff et al. (2019) used a Sankey diagram which shows the students' transitions between course sections by grouping them into different buckets.

Askinadze et al. (2019) used a combination of Venn, Sankey, and UpSet diagrams to perform an in-depth analysis by investigating the effects of individual courses and their combinations. Finally, Kickmeier-Rust (2019) uses jitter to display complex learning data with scatter plots.

The other three papers focus on dashboards for students that aim to provide actionable insights. They report on dashboards that could be used to regulate students learning (Molenaar et al. 2019), help them reflect on their choice of study (Hoppenbrouwers et al. 2019), or to plan interventions in the next run of the course (Hlosta et al. 2019). Molenaar et al. (2019) designed two self-regulated learning interventions based on adaptive learning technology trace data that either let students draw their own dashboards or provide students with advanced personalized visualization. Hlosta et al. (2019) help educators to identify key milestones in the educational process, where the paths of successful and unsuccessful students start to split. Finally, Hoppenbrouwers et al. (2019) showed that *“a more visual representation, confined to only the most essential information, provides a better overview, leads to more and deeper insights while displaying less information and context, and has better usability and attractiveness scores than a more textual version.”*

4 ABOUT THE ORGANISERS

Katrien Verbert is an Associate Professor at the HCI research group of KU Leuven, Belgium. Her research interests include recommender systems, visualization techniques, visual analytics, and applications in healthcare, learning analytics, precision agriculture and digital humanities.

Robin De Croon is a postdoctoral researcher at the HCI research group of KU Leuven, Belgium. His research interests include healthcare informatics, visualization techniques, and gamification.

Tinne De Laet is Associate professor at the Faculty of Engineering Science, KU Leuven. She is the Head of the Tutorial Services of Engineering Science. Her research focuses on using learning analytics, conceptual learning in mechanics, multiple-choice tests, and study success.

Tom Broos is a PhD student at the HCI research group of KU Leuven, Belgium. He researches scalable learning analytics interventions to support first-year students in their transition to higher education. He emphasizes the active receiver, analytical transparency and privacy.

Martijn Millecamp is a PhD student at the HCI research group of KU Leuven, Belgium. His research interests include user interfaces for music recommender systems and dashboards for learning analytics.

Xavier Ochoa is an Assistant Professor of Learning Analytics at the New York University - Steinhardt School of Culture, Education and Human Development. His main research interests include Multimodal Learning Analytics and the application of Artificial Intelligence to solve educational problems.

Bob Bodily received his PhD in the Instructional Psychology and Technology program at Brigham Young University. His research interests include learning analytics, educational data mining, learner dashboards, open educational resources, research trends, and student-generated assessments.

Judy Kay is Professor of Computer Science at the University of Sydney where she leads the multi-disciplinary Human Centred Technology Research Cluster. Her research aims to create new technologies for lifelong, life-wide learning. One key strand creates Open Learner Models (OLMs), interfaces designed to support self-monitoring, reflection, and planning.

Hendrik Drachsler is Professor of Educational Technologies and Learning Analytics and affiliated with the German Leibniz Institute for International Educational Research (dipf.de), the Goethe-University Frankfurt am Main, and the Open University of the Netherlands. His research interests include Learning Analytics, Personalisation technologies, Recommender Systems, Educational data, mobile devices, and their applications in the fields of Technology-Enhanced Learning and Health 2.0.

Cristina Conati is a Professor of Computer Science at the University of British Columbia, Vancouver, Canada. Her research aims to create intelligent interactive systems that can capture relevant user's properties (states, skills, needs) and personalize the interaction accordingly, spanning areas such as Affective Computing, User Modeling, Intelligent Tutoring Systems, User-Adaptive Visualizations, Explainable AI.

5 REFERENCES

- Askinadze, A., Liebeck, M., Conrad, S., (2019). Using Venn, Sankey, and UpSet Diagrams to Visualize Students' Study Progress Based on Exam Combinations. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-5 (accepted)
- Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, 10(4), 405-418.
- Hlosta, M., Kocvara, J., Beran, D., Zdrahal, Z., Visualisation of key splitting milestones to support interventions. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-9 (accepted)
- Hoppenbrouwers, N., Broos, T., De Laet, T., (2019) Less (context) is more? Evaluation of a positioning test feedback dashboard for aspiring students. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-12 (accepted)
- Kickmeier-Rust, M. D., (2019). Using Jitter and Sampling Techniques to Improve the Comprehensibility of Scatter Plots: A Practical Example. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-5 (accepted)
- Klerkx, J., Verbert, K., Duval, E. (2017). Learning analytics dashboards. In: Lang C., Siemens G., Wise A., Gasevic D. (Eds.), *Handbook of Learning Analytics*, Chapt. 12 Society for Learning Analytics Research.
- Molenaar, I., Horvers, A., Dijkstra, R., Baker, R., (2019). Designing Dashboards to support student' Self-Regulated Learning. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-5 (accepted)
- Rohloff, T., Bothe, M., Meinel, C. (2019). Visualizing Content Exploration Traces of MOOC Students. Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), 1-5 (accepted)

Visualizing Content Exploration Traces of MOOC Students

Tobias Rohloff, Max Bothe, Christoph Meinel

Hasso Plattner Institute, Potsdam, Germany

{tobias.rohloff,max.bothe,christoph.meinel}@hpi.de

ABSTRACT: This workshop paper introduces a novel approach to visualize content exploration traces of students who navigate through the learning material of Massive Open Online Courses (MOOCs). This can help teachers to identify trends and anomalies in their provided learning material in order to improve the learning experience. The difficulty lies in the complexity of data: MOOCs are structured into multiple sections consisting of different learning items and students can navigate freely between them. Therefore, it is challenging to find a meaningful and comprehensible visualization that provides a complete overview for teachers. We utilized a Sankey diagram which shows the students' transitions between course sections by grouping them into different buckets, based on the percentage of visited items in the corresponding section. Three preceding data processing steps are explained as well as the data visualization with an example course. This is followed by pedagogical considerations how MOOC teachers can utilize and interpret the visualization, to gain meaningful insights and execute informed actions. At last, an evaluation concept is outlined.

Keywords: MOOCs, Learning Analytics, Content Exploration, Sankey Diagram

1 INTRODUCTION

Massive Open Online Courses (MOOCs) are attended by thousands of learners (Shah, 2018), which makes it hard for teachers to keep the overview of their students' progress. Dashboards have been proven to be a helpful tool by providing different data visualizations and statistics (Klerkx, 2017), as implemented by many MOOC platforms. Some of these visualizations are easy and intuitive to understand and some require a slightly longer learning curve. Especially when the data becomes more complex, it gets harder to understand the visualizations. One difficult case in MOOCs is to comprehend how learners navigate through the course material. Courses are usually structured into sections, which represent course weeks or topic chunks. Each section consists of different learning items, mostly video lectures, texts, exercises and quizzes. Even if an order is given through the structured material, learners can explore the course in any sequence or skip content at all. Thus, it is complicated to visualize these content exploration traces for all students of a course. However, teachers could benefit for example by identifying anomalies within their provided material.

With this workshop paper, we introduce one possible visualization technique that we implemented for the HPI MOOC platform¹. We looked at approaches from other disciplines, like conversion rates in web analytics (Zheng, 2015) and funnel charts. In the end, we decided to implement a Sankey diagram which displays transitions between course sections based on their amount of visited learning items of each student. This paper explains how the data is being processed and visualized. Additionally, it

¹ <https://open.hpi.de/>

discusses how MOOC teachers and instructors can utilize and interpret the visualized data, to obtain meaningful insights from their students' learning behavior and execute informed actions and interventions. At last, an evaluation concept is presented to investigate the helpfulness and comprehensibility of the visualization, amongst other things.

2 DATA PROCESSING

In order to visualize the student's content exploration traces with a Sankey diagram, the captured interaction data needs to be processed first. The platform stores all user interaction events in redundant analytics storages within a central learning analytics service (Rohloff, 2018). These storages are realized with different database technologies to enable different query techniques (like SQL or NoSQL) for an optimized performance of each implemented metric. The events are structured in an xAPI-alike format (Renz, 2016). For the intended visualization, the users' learning item visits need to be processed into aggregated nodes and links for the Sankey diagram. Therefore, a new metric was introduced within the learning analytics service, which takes care of three processing steps (one database query and two post-processing steps) to compute the final data structure. These three processing steps are explained briefly in the following sections. The data processing performance depends on the size of the data basis. With our current infrastructure and more than 360,000,000 user interaction events today on a single platform, the load is too high to generate the data on-demand with every request. Therefore, we decided to process the data once per day for each course and store the results. The persisted data is then displayed to teachers.

2.1 Process Raw Events into Visit Counts per Section for each User

The first step processes the raw events stored in the database and calculates the unique item count per section for every user. Therefore, the SQL-based storage was used (PostgreSQL). The data is stored in an event table and queried. It returns a list of dictionaries with each element containing a unique combination of a `user_id` and `section_id`, as well as the distinct visited `item_count` (visit count). Additionally, only events captured during the regular course runtime have been taken, since self-paced course activity should be examined separately. The maximum length of this list is $u * s$, where u is the number of users and s is the number of sections.

2.2 Process Visited Percentage for all Sections for each User

Now, for each user, the visited percentage for all sections of a course is calculated based on the visit count from the previous step. This ensures that values for all sections are generated, even for sections without any visits. Additionally, the visited percentages are sorted according to the section positions. This results in a dictionary where every `user_id` points to a sorted list of visited percentages, representing the different course sections.

2.3 Process Bucketized Nodes and Node Links

In the third step, the different percentage values of every user are aggregated. Each Sankey node layer will represent a different course section. The nodes of one layer will display different visited percentage buckets. Therefore, the visited percentages ranging from 0.0 to 1.0 were split into specific intervals. We decided on the following configuration:

$$[0.0] \cup]0.0, 0.2[\cup [0.2, 0.4[\cup [0.4, 0.6[\cup [0.6, 0.8[\cup [0.8, 1.0[\cup [1.0]$$

We treated no visits (0.0) and all items visited (1.0) as special cases to identify learners who never showed up in a section (no-shows) and very engaged learners (completers). Now, we had to determine for each possible link between the nodes of adjacent layers (source node of layer a to target node of layer b) the number of users. This resulted in $i * i * (s - 1)$ links, where i is the number of defined intervals and s is the number of sections. For our configuration of 7 intervals and a course with 7 sections, this would produce 294 links for all nodes. The size of each node can be derived from the links, by summing up the corresponding user counts. The number of nodes is $i * s$, resulting in 49 nodes for our given example. With all nodes and links in place, the Sankey diagram can be drawn.

3 DATA VISUALIZATION

To render the diagram as part of the platform's web-based teacher dashboard a D3² Sankey plugin³ was used. The output for our example configuration is shown in Figure 1. The 7 vertical node layers represent the course sections (the section labels are omitted in the figure). Each layer consists of 7 nodes, which are annotated with the corresponding visited percentage intervals. The nodes and links are color-coded, ranging from red for no visited items to green for all items visited. The colors of the inner nodes are interpolated with a ratio based on the nodes' interval threshold. This enables a dynamic colorization based on the interval configuration. The links have the same color as their target node, with a slight transparency to display overlaps.

The specific user count value of a node or link is shown when hovered in the web browser. Additionally, all links connected to a node are highlighted when hovering the node. This can help teachers to comprehend cohorts of students with unusual behavior. For example, if a larger group of students, who visited a lot of learning items in a section, only visits a few items of the following section.

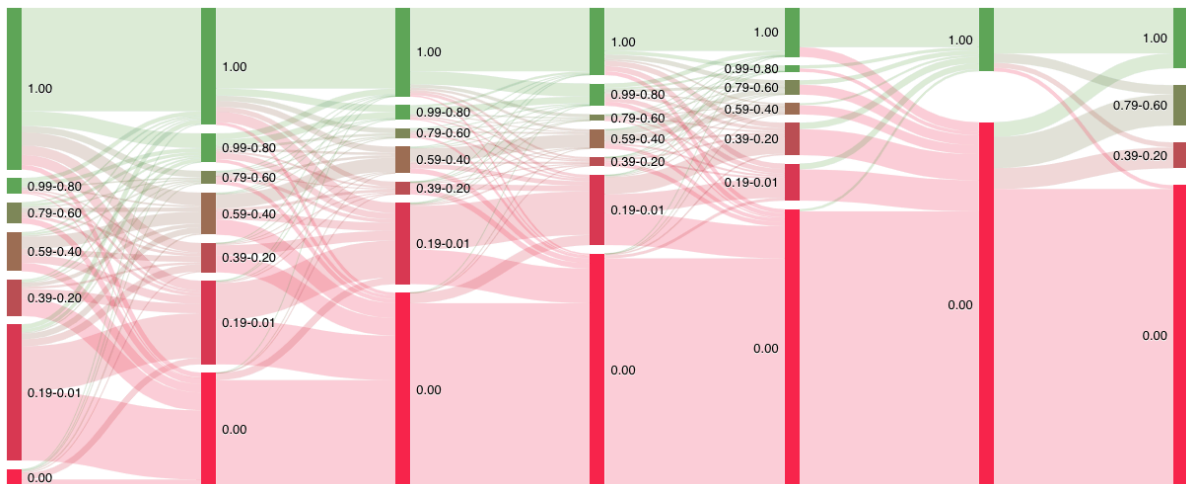


Figure 1: Content Exploration Traces of a MOOC with 7 Sections visualized as a Sankey Diagram.

² <https://d3js.org/>

³ <https://github.com/q-m/d3.chart.sankey>

The course chosen to visualize Figure 1 included 5,284 students. The first five sections were successive course weeks, each consisting of video lectures, self-tests and a weekly graded quiz. In the sixth section a final exam was conducted, followed by an “I like, I wish” final section to gather the students’ feedback. It can be seen that in this example, that the majority of completers also complete the following section and only a few of them visit fewer items. Also, only a minority of no-shows come back to visit content in the next sections. The midfield shows a more diverse picture and there is also a general trend visible, that the number of no-shows increases from section to section, who will end as drop-outs most probably. However, the reasoning behind requires the knowledge of a human expert. Visualizations like this can only support evaluations and decisions resulting from it.

4 PEDAGOGICAL VALUE

After the presentation of technical implementation details, this section will introduce how teachers and instructors in MOOCs can utilize this diagram and benefit from it. Related and similar visualizations either used stacked bar charts for a weekly student participation overview, or state transition diagrams for learning items (Coffrin, 2014). However, our proposed diagram provides the advantages of both approaches: it shows a complete course overview with different stacked user subgroups and also displays the transitions of these subgroups between course sections. This should serve as a starting point for teachers to get a first overview of student activity and their content exploration traces in a course. Thereby, the diagram is meant to complement other visualizations which focus on more detailed aspects of a course, like assignment grade distributions, video navigation charts or forum activity graphs (Stephens-Martinez, 2014).

Therefore, it is placed in the teacher’s central course dashboard as one of the first visualizations. Based on the displayed data, the teacher is able to quickly spot unusual behavior and anomalies across the whole course. Then, the corresponding content can be delimited and identified to further examine the issue and execute informed interventions. Thereby, the two main elements of the Sankey diagram can be used to interpret the data. First, the stacked nodes show how many active and less active students participate in a certain course section. These numbers can be either compared with other course sections, other iterations of the same course, or different courses to see how well perceived a certain section is. Second, the links show the transitions of different engaged student subgroups, which can be helpful in various ways. No-shows in one section are highly likely to be no-shows in the following section as well since they never appear in the course again and can be ignored. However, a transition of a large portion of highly engaged users in one week to a low activity group in the next week is unusual and may indicate an issue in the preceding week, like too tiring video lectures or a too difficult weekly exam. But here the interpretation possibilities are reaching the limits of this diagram and depend on the respective case. Nevertheless, further investigations can be done with other visualizations which focus on certain aspects instead of a complete overview, as discussed before.

A future evaluation is necessary to test our assumptions if the Sankey diagram is interpretable enough for real-world MOOC instructors and if they consider it as helpful to monitor their students’ activity to make informed and meaningful actions. This evaluation will be done on different deployments of the HPI MOOC platform. It is planned to do this separately, but also as part of a larger teacher dashboard evaluation. Interviews can be used for qualitative analysis, and for quantitative analysis surveys and usage data. The usability and comprehensibility will be investigated, but also the specific value as a learning analytics tool, e.g. by measuring its EFLA score (Scheffel, 2017). Even if the diagram is able to

realize the goal of an overview of a whole course, teachers need to explore the details of an identified trend or anomaly to better examine the cause. Therefore, the diagram must be complemented with more detailed visualizations, which show what happens inside sections between different learning items. Here, also the difficulty of quizzes or comprehensibility of videos can be utilized for example, next to the item visits. This needs to be implemented as well before the evaluation is conducted.

5 CONCLUSION

This paper introduced a novel approach on how a Sankey diagram can be used to visualize students' content exploration traces between sections of a MOOC. Based on captured user interaction events, three processing steps were explained to generate the data for the nodes and links of the Sankey diagram, by using vertical node layers as a representation of different MOOC sections. Each node displays the share of a certain interval of the total visited learning items percentage of a section. By interacting with the Sankey diagram, the teacher can highlight connected notes to comprehend cohorts of students with unusual behavior. An example is shown for a real-world MOOC with possible conclusions. Additionally, the pedagogical value of the visualization was discussed, how instructors can use and interpret the visualization and gain meaningful insights to take informed actions. Also, an evaluation concept was outlined to test our assumptions and examine the helpfulness and comprehensibility. All in all, this work showed a possibility to display a complete overview for MOOC teachers, how their thousands of students navigate through the course material.

REFERENCES

- Coffrin, C., Corrin, L., de Barba, P., Kennedy, G. (2014, March). *Visualizing patterns of student engagement and performance in MOOCs*. Paper presented at the 4th International Conference on Learning Analytics and Knowledge. <https://doi.org/10.1145/2567574.2567586>
- Klerkx, J., Verbert, K., Duval, E. (2017). Learning Analytics Dashboards. In Lang, C., Siemens, G., Wise, A., Gasevic, D. (Eds.), *Handbook of Learning Analytics* (Chapt. 12). Society for Learning Analytics Research.
- Renz, J., Navarro-Suarez, G., Sathi, R., Staubitz, T., & Meinel, C. (2016, April). *Enabling Schema Agnostic Learning Analytics in a Service-Oriented MOOC Platform*. Paper presented at the 3rd ACM Conference on Learning @ Scale. <https://doi.org/10.1145/2876034.2893389>
- Rohloff, T., Bothe, M., Renz, J., & Meinel, C. (2018, September). *Towards a Better Understanding of Mobile Learning in MOOCs*. Paper presented at the 5th IEEE Conference on Learning with MOOCs. <https://doi.org/10.1109/LWMOOCs.2018.8534685>
- Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017, September). *The Proof of the Pudding: Examining Validity and Reliability of the Evaluation Framework for Learning Analytics*. Paper presented at the 12th European Conference on Technology Enhanced Learning. https://doi.org/10.1007/978-3-319-66610-5_15
- Shah, D. (2018, January). By The Numbers: MOOCs in 2017 [Blog post]. Retrieved from <https://www.class-central.com/report/mooc-stats-2017/>
- Stephens-Martinez, K., Hearst, M. A., Fox, A. (2014, March). *Monitoring MOOCs: Which Information Sources Do Instructors Value?* Paper presented at the 1st ACM Conference on Learning @ Scale. <https://doi.org/10.1145/2556325.2566246>
- Zheng, J., Peltsverger, S. (2015, January). Web Analytics Overview. In Khosrow-Pour, M. (Ed.), *Encyclopedia of Information Science and Technology, Third Edition* (Chapt. 756). IGI Global.

Using Venn, Sankey, and UpSet Diagrams to Visualize Students' Study Progress Based on Exam Combinations

Alexander Askinadze, Matthias Liebeck, Stefan Conrad
Heinrich Heine University Düsseldorf, Germany
{askinadze, liebeck, conrad}@cs.uni-duesseldorf.de

ABSTRACT: Educational dashboards allow educators to gain insights about their students and their learning progress. It is essential to understand why students may drop out of the university. In our educational dashboard, we used a combination of Venn, Sankey, and UpSet diagrams to perform an in-depth analysis by investigating the effects of individual courses and their combinations. We present our visualizations based on student data from a computer science course at a German university.

Keywords: student visualization, Venn diagram, Sankey diagram, UpSet diagram

1 INTRODUCTION

Educational institutions are collecting more and more data that can be analyzed. Learning Analytics (LA) is a research field that deals with the analysis of such data and is defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs.” (Siemens, 2010)

Arnold and Pistilli (2012) presented the “Course Signals” (CS) system, a successful example of the use of learning analytics at Purdue University. Based on available data in the CS system, an algorithm predicts students' risk levels. The predicted risk levels can be used by instructors, for example, to intervene by posting a traffic signal indicator (green, yellow or red indicating the likelihood of being successful or unsuccessful) on a student's dashboard or by e-mail messages. Their evaluation showed that the usage of CS delivered significantly higher retention rates than without.

While predictive models can be used to identify at-risk students and intervene as mentioned above, Charleer, Klerkx, and Duval (2014) mention that such predictive models may be black boxes and give users no insight into the reasons for the decisions made. This motivates us to use learning analytics dashboards that apply information visualization techniques to display students' data so that educators may gain deeper insight and become empowered to make own decisions, rather than relying on automated decisions (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013).

With our work on a dashboard for educators in an academic context, our research belongs to the research area of LA. We want to provide a better insight into the overall study progress of different student cohorts by proposing new visualizations that show which exams students passed jointly. Students can drop out of studies for many reasons (Sagenmüller, 2018). Among other things, it may be due to certain mandatory exams. The proposed visualizations, especially for the analysis of drop-out students, may help to find potential causes of drop-out associated with the curriculum. The core questions that we want to answer are: 1) How can we visualize which exams or exam combinations

our graduates or drop-out students passed until a selected semester?; 2) How can we use the visualizations to identify certain exams that are difficult for students?; 3) How can we compare the study progress of different cohorts of students in one visualization?

2 RELATED WORK

In the past, many visualization techniques have been applied to present in an educational dashboard. Charleer, Klerkx, and Duval (2014) report that most dashboards consist of basic visualizations, e.g., bar and line charts or scatter plots. Gray, Teahan, and Perkins (2017) list further common visualization techniques, such as pie and donut charts or tables, as well as some advanced visualization techniques related to the information visualization community, such as tag clouds, stream diagrams, heat maps, sunburst diagrams, aster plots, bubble diagrams, and radar plots. Sankey diagrams are another visualization technique that has been used in the learning analytics context. They were originally intended to visualize the energy efficiency of a steam engine. Sankey diagrams can be used to display study progress over several semesters. Morse (2014) investigated cohorts of students and utilized Sankey diagrams to visualize how the students changed their major, graduated or dropped out throughout multiple semesters. Similarly, Heileman, Babbitt, and Abdallah (2015) used Sankey diagrams to debunk myths about student progress by visualizing student cohorts from different majors as being enrolled, having graduated or stopped studying.

In contrast to previous research, we need visualizations that show which exams students have passed until a given semester. This would be possible, for example, with simple bar charts, but then the information about which exams were passed jointly would be lost. In addition, the time component that shows in which semester the respective exams were passed, would be lost. Therefore, we rely on and combine the strengths of Venn, Sankey, and UpSet diagrams as an advanced visualization technique to perform an in-depth analysis by investigating the effects of individual courses and their combinations on the overall study progress in the next chapters.

3 METHOD

In this chapter, we discuss the three proposed visualization techniques in general and show their application on real data in the following chapter. Usually, visualizations depict courses or modules individually. We now focus on our core questions and investigate which exams are passed jointly. For our educational dashboard, we implemented a drop-out analysis method which allows us to analyze students who have completed their studies without a degree until a certain semester by selecting multiple courses. For the analysis of the drop-outs in the first semesters, we recommend selecting the courses that are scheduled for the first semester in the curriculum.

Table 1: Pros & cons of Venn, UpSet, and Sankey diagrams in our dashboard.

	Venn	UpSet	Sankey
shows jointly passed exams until the selected semester	x	x	x
temporal information (shows information per semester until selected semester)			x
can compare different cohorts in one diagram			x
visualization scales well for more than three exams		x	
exam combinations are displayed clearly and intuitively	x		

One approach we have not seen before in the context of LA are Venn diagrams which can display, for example, intersections, unions, differences, and symmetric differences. With Venn diagrams, we want to investigate which exams are passed before students drop out. For each exam combination, the corresponding intersection set shows the number of students who passed all exams in this intersection. Venn diagrams are easy to understand for a small number of exams. However, they are impractical for a large number of analyzed exams (> 3) since the number of combinations grows exponentially with the number of exams. We use the implementations of Venn diagrams from the JavaScript library D3¹ (Bostock, Ogievetsky, & Heer, 2011) and venn.js¹. D3 is extendable and uses modern web technologies, such as SVG and canvas, for the interactive visualization of complex data.

Therefore, we propose to additionally use the visualization technique UpSet (Lex, Gehlenborg, Strobel, Vuilleumot, & Pfister, 2014). For every module combination, UpSet also visualizes the number of jointly passed exams with a bar chart for which the legend below indicates which combination corresponds to each bar, as visualized in Figure 1 (b). The UpSet diagram is sorted in descending order, allows to hide specific combinations (which is not possible in the Venn diagram) (Khan & Mathelier, 2017), and can display a high number of exam combinations since the diagram grows horizontally and can still be scrollable. A disadvantage of UpSet over Venn diagrams is that all intersections for a specific module do not need to be directly side by side due to the sort order and are, therefore, more difficult to interpret. Therefore, we show both visualizations side by side in our dashboard to combine their respective strength. For the implementation, we relied on upset.js¹.

Although Venn and Upset diagrams are suitable visualization techniques for displaying combinations of exams cumulatively up to a particular semester, they cannot be used to show study progress for different semesters. Additionally, they are not able to visualize different cohorts, e.g., graduates and dropouts, in one figure at the same time. We suggest using Sankey diagrams for both purposes. Each node in a Sankey diagram is a combination of exams in a specific semester, as visualized in Figure 1 (c). When analyzing many exams, the visualization of the names for each exam combination is challenging. Therefore, we decided to use a binary encoding. For n different exams, there are 2^n possible combinations. Below the visualization is a legend, which assigns the binary numbers to their corresponding combination of passing (1) and failing (0) exams. For example, the notation 1_100 represents those students that only passed the first of three selected exams at the end of the first semester. For the implementation, we used the D3 extension d3-sankey¹. Since Sankey diagrams display detailed temporal data, they are powerful, yet difficult to understand, especially for a high number of exams and a time span of multiple semesters. We summarized the advantages and disadvantages of the three visualization methods in Table 1.

4 ANALYSIS / APPLICATION ON REAL DATA

We present our visualizations based on students' progress data of different cohorts of a German computer science degree program. The courses of the first-semester curriculum are Calculus I, Linear Algebra I, and Programming. All of these courses are mandatory for the bachelor's degree. Since already 25% of all dropouts occur until the end of the second semester and we are interested

¹ github.com/d3/d3; github.com/benfred/venn.js; github.com/chuntul/d3-upset; github.com/vasturiano/d3-sankey

in helping students as early as possible in their studies, we start by investigating students that drop out before their third semester. For the analysis, we filtered out all students who did not have a single exam attempt and for whom no study history data is therefore available. Figure 1 (a-c) shows passed exams of students that dropped out of their studies until the end of their second semester in Venn, UpSet and Sankey diagrams, respectively. However, by examining the passed exams, it is not clear at all whether the students even tried to pass the exams. Therefore, we propose to additionally display exam attempts, as depicted in Figure 1 (d).

Our visualizations show that students are less likely to pass the mathematical exams (and especially both of them) than the computer science course Programming. Dropouts that are able to pass a math exam are, to a large extent, able to also pass the other two exams. From the Venn diagram in (a), we can see that most of the dropouts do not pass a single course. Compared with the UpSet diagram of exam attempts (d), we notice that the majority of them at least tries to pass the Programming exam. Also, we see that more than half of the students that try to pass the Programming exam do not even try to pass one of the math exams. Different reasons might account for this behavior, for example, that it may be easier to meet the examination requirements for Programming or that math subjects appear too difficult for the students and they, therefore, do not even register for the exam. In any case, the math exams seem to be the most significant obstacles.

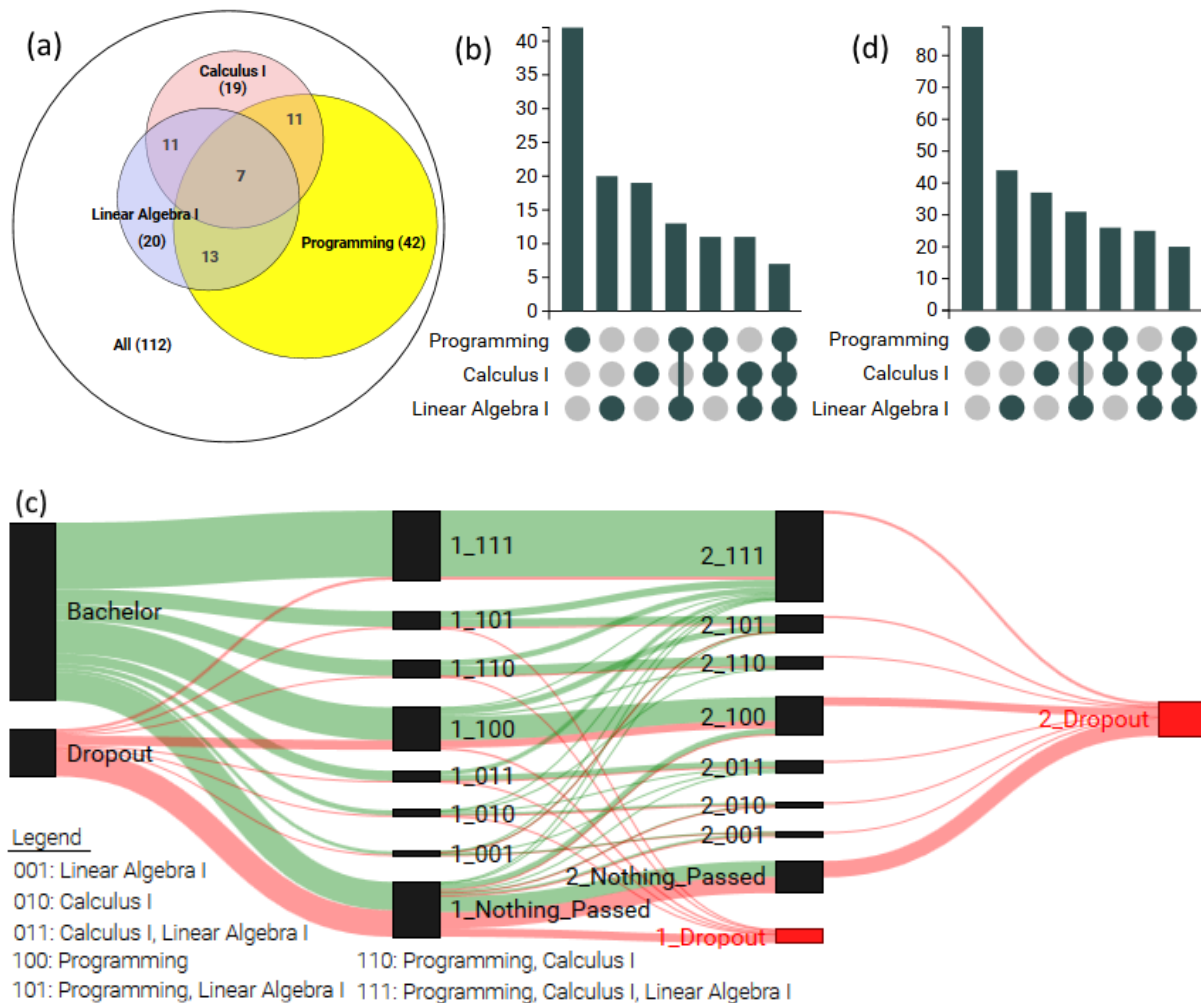


Figure 1: Venn (a), UpSet (b) & Sankey diagrams (c) showing jointly passed exams of students who dropped out of studies till the end of the second semester. UpSet (d) diag. shows exam attempts.

The Sankey diagram (c) visualizes the study progress of students that dropped out until the end of the second semester as well as graduates. We can see that many students who have passed all three freshman courses after the first semester will most probably not drop out until the end of the second semester. Also, most of the dropouts before the end of the second semester did not pass any of the freshman courses or only the Programming course. Since teachers have little chance of helping students who drop out due to personal reasons, it now makes sense to focus in detail on the courses and compare the behavior of dropouts with that of graduates. We can see that most of the students (including graduates) who have only completed the Programming exam at the end of the first semester will not have passed the math exams until the end of the second semester. This suggests problems with the math exams for which solutions should be found.

5 CONCLUSION

In this work, we presented three visualizations that are suitable to illustrate combinations of exams. The first two visualizations, the Venn diagram and the UpSet diagram, can be displayed side by side as they complement each other. Together, they can deliver new informative insights into the study behavior of dropouts. However, they cannot visualize temporal study progress over multiple semesters. Therefore, we proposed the usage of Sankey diagrams that are able to visualize the study progress by combining passed and failed exams. The presented visualization methods allowed us to gain new insights into the data that we would not have seen without them. In the future, we want to evaluate our visualization on different data, make our dashboard more interactive, and use Sankey diagrams to visualize the temporal progress of course exercises.

REFERENCES

- Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on LAK*. ACM, 267-270.
- Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ Data-Driven Documents. *IEEE Transactions on Visualization & Computer Graphics*, (12), 2301-2309.
- Charleer, S., Klerkx, J., & Duval, E. (2014). Learning Dashboards. *Journal of Learning Analytics*, 1(3), 199-202.
- Gray, C. C., Teahan, W. J., & Perkins, D. (2017). Understanding our Analytics: A Visualization Survey.
- Heileman, G. L., Babbitt, T. H., & Abdallah, C. T. (2015). Visualizing Student Flows: Busting Myths About Student Movement and Success. *Change: The Magazine of Higher Learning*, 47(3), 30-39.
- Khan, A., & Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC bioinformatics*, 18(1), 287.
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE transactions on Visualization and Computer Graphics*, 20(12), 1983-92.
- Morse, C. (2014). *Visualization of Student Cohort Data With Sankey Diagrams via Web-Centric Technologies* (master's thesis). University of New Mexico, Albuquerque, New Mexico.
- Sagenmüller, I. (2018). Student retention: 8 reasons people drop out of higher education. Retrieved January 13, 2019, from u-planner.com/blog/student-retention-8-reasons-people-drop-out-of-higher-education
- Siemens, G. (2010). 1st International Conference on Learning Analytics and Knowledge 2011. Retrieved January 16, 2019, from <https://tekri.athabascau.ca/analytics/>
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10), 1500-1509.

ACKNOWLEDGEMENTS: This work was partially funded by the IST-Hochschule University of Applied Sciences.

Designing Dashboards to support learners' Self-Regulated Learning

Inge Molenaar

Behavioural Science Institute, Radboud University Nijmegen

i.molenaar@pwo.ru.nl

Anne Horvers

Behavioural Science Institute, Radboud University Nijmegen

a.horvers@pwo.ru.nl

Rick Dijkstra

Behavioural Science Institute, Radboud University Nijmegen

rick.dijkstra@student.ru.nl

Ryan Baker

University of Pennsylvania

ryanshaunbaker@gmail.com

ABSTRACT: This contribution reports on the development of two learner-faced dashboards that support learners' self-regulated learning during practice activities in an adaptive learning technology (ALT). While learners learn using adaptive learning technologies on tablets, they leave rich traces of data that capture many details of their learning processes. The data can be used to create dashboards that support learners to make valid inference about how they regulate control and monitor their learning. Such personalized visualizations are a new tool to support learners regulation. In this paper we describe two designs of personalized dashboards supporting SRL. The first dashboard is drawn by learners themselves based on ALT achievement data. Learners are asked to set goals at the start of each lesson and add their achievements after each lesson. This is used as input to monitor progress and determine whether adaptation is needed to reach their goals. Learners draw elements of the dashboard themselves and hence make their own personalized visualizations. The second dashboard follows the same logic, but the visualization process is automated in an app. Again learners set their goal at the start of each lesson and view their achievement and progress in the dashboard after each lesson. Additionally, learners also are presented with their learning path based on Moment-by-Moment Learning Curves and cues to translate data into actionable feedback to efficiently reach learning goals. The contribution of this paper is to discuss the design and rationale for the two dashboards that support young learners SRL based on ALTs trace data.

Keywords: Adaptive Learning Technologies, Self-Regulated Learning, Personalized Visualisations

1 BACKGROUND

This contribution describes two approaches to translate learners' trace data from Adaptive Learning Technologies (ALTs) into personalized visualizations that function as dashboards to support learners' self-regulated learning (SRL). In the Netherlands alone, over 250,000 students in primary education

learn Mathematics, Dutch and English using adaptive learning technologies (ALTs) such as Snappet, Muiswerk, Taalzee/Rekentuin, Got it, and PulseOn on a daily basis (Kennisset, 2014). These systems provide learners with instructional materials and practice opportunities that are aligned with the current level of learners' knowledge (Aleven, McLaughlin, Glenn, & Koedinger, 2016a; Klinkenberg, Straatemeier, & Van Der Maas, 2011a). When learners learn with adaptive learning technologies on tablets, they leave rich traces of data that capture many details of their learning process (Gašević, Dawson, & Siemens, 2015). Although ALTs successfully use learner data to adjust instructional materials to learners performance, supporting learners' self-regulated learning is not a focus of most ALTs being used at scale (Winne & Baker, 2013). Even though the important role of self-regulated learning (SRL) has been emphasised in the field of learning analytics and quite a few learner-faced dashboards have been developed aimed to support SRL (Winne & Baker, 2013), these dashboards do not use trace-data nor support learners to translate data into appropriate actions (Bannert, Molenaar, Azevedo, Järvelä, & Gašević, 2017).

Dashboards are loosely defined as: *"Single displays that aggregated different indicators about learners, learning processes and or learning contexts into one or multiple visualizations"* (Schwendimann et al., 2017). Research around dashboards traditionally has a strong focus on the learning analytics and educational data and less attention is paid to the pedagogical value and connection to learning sciences (Jivet, Scheffel, Specht, & Drachsler, 2018). Although SRL theory is the most common foundation for learner-faced dashboards, most of these dashboards only visualize indicators of learner achievement to support students awareness or reflection (Bodily & Verbert, 2017). Dashboards often fail to support learners in translating awareness into actions to improve regulation. Moreover, none of the dashboards reviewed in a recent review by Jivet et al (2018) used trace data to support SRL. This is especially surprising considering the well-established measurement problems with self-report measurements of SRL (Azevedo, 2009). The relative rarity of trace data used as support for SRL can be explained by the challenges to understand what learner trace data reveal about SRL (Bannert et al., 2017; Molenaar & Järvelä, 2014). Hence the purpose of this contribution is to explore how trace data from ALTs can be used to develop dashboards that supports learners' SRL and provide learners with actionable feedback. Especially for young learners in primary education learner-faced dashboard have been under represented in research and we are unaware of any learner-faced dashboard supporting SRL with trace data (Jivet et al., 2018). This contribution starts with the pedagogical basis for this dashboards discussing SRL theory and explicitly grounding the dashboard design in SRL theory. Next, we discuss the dashboard design including the data used, explanation of the visualizations selected and the interaction techniques and implementation in the educational setting and workflow.

1.1 SRL theory as basis for the design of the dashboards

SRL theory defines learning as a goal-oriented process in which learners make conscious choices working toward learning goals (P. H. Winne & Hadwin, 2017; Zimmerman, 2000). Self-regulated learners use cognitive activities (read, practice, elaborate) to study a topic, use metacognitive activities (orientation, planning, monitoring, and evaluation) to control and monitor their learning, and motivate themselves to engage in an appropriate level of learner effort (Azevedo, Moos, Greene, Winters, & Cromley, 2008). Following the COPES model (Winne, 2018; Winne & Hadwin, 1998) regulation unfolds in 4 loosely coupled phases: i) the task definition phase in which learners

generate an understanding of the task, ii) the goal setting phase in which learners set their goals and plan their actions, iii) the enactment phase in which learners execute their plans working towards their goals and finally iv) the adaption phase which is activated when progress towards the goals is not proceeding as planned and adjustments in strategies, actions or tactics are required. These phases occur in the context of task conditions, standards that learners set to represent their goals and operations performed by learners that lead to new products in the form of knowledge or skills. The control and monitoring loop are at the heart of COPES model. In cognitive evaluations learners relate their achieved products to their standards in order to assess progress towards their goals. Although the COPES model explains how learners' internal feedback functions, it is well established that learners often face a utilization deficiency (Winne & Hadwin, 2013). This is the failure to adequately activate control and monitor loop during learning. Dashboards are potentially a powerful tool to overcome this utilization deficiency as they can help learners with objective data about the current products obtained (achievement), how they relate to learning goals (progress) and how that relates to standards (Molenaar, Horvers, & Baker, 2019). This form of external feedback can consequently drive the adaptation phase, helping learners' adjust learning behaviour leading to optimized strategies, adjustments to plans or different actions in the enactment phase.

Hence when internal feedback fails, dashboards can support learners with external feedback to adjust the regulation during learning (Butler & Winne, 1995). Learners often receive external feedback from the teacher or the ALT indicating the correctness of an answer to a problem (Alevin, McLaughlin, Glenn, & Koedinger, 2016b). Although this supports local corrections, this type of feedback does not provide sufficient information to adjust control and monitoring. Specifically, this feedback does not trigger cognitive evaluation which is important for learners that do not regulate their learning sufficiently (Azevedo et al., 2008). Different techniques (e.g., prompts (Bannert, Hildebrand, & Mengelkamp, 2009), scaffolding (Azevedo et al., 2008), intelligent tutor systems (Azevedo et al., 2016)) have been used to assist learners' regulation in ALTs. Although these techniques are initially effective, they are less successful in sustaining regulation during learning in absence of the tools. A drawback of these techniques is that they do not help learners to make explicit inferences about how their actions are related to progress towards learning goals (Winne & Hadwin, 2013). The fit between achievement (products) and internal representations of the learning goals (standards) remains underspecified and the contribution of actions to progress is unclear. In order to engage in cognitive evaluations learners need reliable, revealing, and relevant data in order to be able to draw valid inferences about their own learning process (Winne, 2010). Data from ALTs can be used to provide learners with continuous feedback about their achievement, progress and above all to understand how progress towards their learning goal is related to their actions. This entails that the role of dashboards needs to be extended from discussing *what* learners learned to also incorporate *how* learners learned. Hence dashboard can be the basis for developing a promising way to overcome learners' utilization deficiencies of regulatory strategies, and consequently increase learners' SRL skills for future learning.

Learner-faced dashboards have just recently become a more prominent way of providing SRL support e.g. Bodily et al., (2018), although visualizations on learners' achievements have been used in some learning systems for some time (Arroyo, et al. 2007; Koedinger et al., 2007). However, a recent review by Jivet (2018) and colleagues indicates that most of these dashboards do not provide actionable information for learners to improve their regulation. Following the learning analytics

process model learners need to translate awareness into action (Bodily & Verbert, 2017). They need a ‘representative reference frame’ to interpret the data (Wise, 2014). Both achievement and progress can be valuable ways to create such a reference frame, but as described above only when learners have internal standards, against which they are evaluated (Winne & Hadwin, 2013). These standards help learners to set criteria that indicate *how* to know that a learning goal is reached. Frequently, learners are in need of additional external help to create standards. This is also referred to as *feed-up*, which represents an external trigger to support learners to articulate *when* learning goals are reached (Hattie & Timberley, 2007). Feed-up interventions can be used to support learners to explicitly set standards. Consequently, this can support learners’ cognitive evaluations in the enactment phase. Only when learners establish that there is a difference between their achievement and standards set, they realize that progress is not as anticipated and adaptation is needed. This may cue re-evaluation of plans and adjustment of strategies, but only when learners are able to determine next steps to reach the learning goal. External feedback to articulate this is named *feed-forward* (Hattie & Timperley, 2007), when a learner’s verbalizes how to adapt learning strategies and actions to ensure future learning. Thus, next to assessment feedback that indicates how a learner is doing on one task (feedback), *feed-up* and *feed-forward* are external feedback that can help learners to effectively monitor and control their learning. A comprehensive approach towards learner-faced dashboards includes both the assessment of learners achievement on a cognitive level (feed-back on achievement) as well as information on progress to stimulate cognitive evaluation by supporting the monitoring loop (feed-up) and recommendations to drive adaptations in the control loop to proceed towards the learning goal (feed-forward).

The learners’ data traces in ALTs provide indications of learners’ achievement and progress towards their learning goal (Molenaar et al., 2019) and specifically the relation between learning actions and progress i.e. the learning path. Therefore the data can be used to help learners explicitly reflect on achievement and progress towards their learning goals (Winne, 2010). To indicate the relation between actions and progress explicit we use Moment-by-Moment Learning Curves (Baker, HersHKovitz, Rossi, Goldstein, & Gowda, 2013; Baker, Goldstein, & Heffernan, 2011). These curves show how much the learner is likely to have learned at each problem-solving opportunity, which is a representation of progress over time. This may function as a tool to show learners how they regulate their learning over time. Research has shown that Moment-by-Moment Learning Curves show specific patterns that are not only associated with learning but also regulation of accuracy (Molenaar, Horvers, & Baker, submitted). Hence, these patterns could potentially help learners understand the development of progress during a lesson and subsequently triggering adaptation. Consequently, dashboards visualizing achievement, progress towards learning goals and the learning path may play a central role in guiding learners to optimize their regulation.

2 THE DASHBOARD DESIGN: DATA, VISUALIZATION AND INTERACTION TECHNIQUES

In this contribution we explore two possible types of dashboards to support SRL and serve as an form of external feedback for learners. The dashboards are developed in the context of ALT which also generates the data used.

2.1 Data from the adaptive learning technology

The *adaptive learning technology (ALT)* used in this study is widely used for spelling and arithmetic education throughout the Netherlands. This technology is applied in blended classrooms in which the teacher gives instruction after which learners practice on their tablets. First, learners solve non-adaptive problems, which are the same for each student in the class. After this, the learners work on adaptive problems. Adaptive problems are selected after each problem solved based on an estimate of the learner's knowledge called the ability score (Klinkenberg, Straatemeier, & Van Der Maas, 2011b). This score is calculated by a derivative of the ELO algorithm (ELO, 1978). Based on the learner's ability score, the ALT selects problems with a probability of 75% that the learner will answer the problem correctly. After a learner has answered approximately 25 problems, the system has a reliable indicator the ability score. This ability score is used as indicator of *achievement*. The difference between the previous ability score and the new score is the indicator of *progress*.

Next to adaptive problems, Learners are given direct feedback (correct or incorrect) after entering an answer to a problem and teachers can follow learners in teacher dashboards (Molenaar & Knoop-van Campen, 2018).

The log data from the ALT consist of: A date and time stamp, learner identifier, problem identifier, learning objective identifier, ability score after the mentioned problem and the correctness of the answer the learner gave.

2.2 Techniques to transform data: Moment-by-moment learning curves

The ALT data are used to create *Moment-by-Moment Learning Curves* (MbMLC) using an algorithm developed by Baker, HersHKovitz, Rossi, Goldstein, & Gowda (2013). These curves are used to visualize a learner's learning over time. The probability a learner has just learned a skill is plotted across the learner's problem solving attempts over time while practicing on a specific skill. A newly developed Python script is used to label the MbMLC based on Baker et al. (2013) following the rules in Table 1. A peak is defined as a point more than 0,015 higher than the point before or after. A new common pattern was found, with two peaks, so this pattern is added as 'double spike'.

Table 1: Rules for coding moment-by-moment learning curves.

Curve	Rules
Immediate drop	The curve starts high, drops quickly after solving problems and remains low afterwards.
Immediate peak	The curve starts low, peaks within the first 10 problems and remains low afterwards.
Double spikes	The curve starts low and shows 2 peaks over the course of problem solving.
Close multiple spikes	The curve starts low and shows more than 2 peaks within the first 25 problems and remains low afterwards.
Separated multiple spikes	This curve starts low and continues to show multiple peaks, even after 25 problems

2.2.1 Dashboard A: drawing your own dashboard

In study A learners are asked to draw their own dashboard. At the start of first three lessons, learners are asked to answer four questions regarding their learning goals: 1. How skilled do you want to become at that particular subskill? 2. How many lessons do you need to reach that goal? 3. How skilled do you want to become in this particular lesson? These questions are answered on a scale from 1 (not very good) to 6 (excellent). Also, learners are asked which percentage of problems they wanted to solve in one attempt (0% to 100%). Learners answered by drawing the bars below the questions, see the left side of Figure 1. The chosen colour represent different levels of achievement also used in the ALT to indicate achievements. This stage was designed to act as a *feed-up* intervention in which learners clearly articulated their learning goal and set their standards to evaluate progress.

After the first three lessons, learners are asked to reflect on their learning by answering three questions: 1. What is your current knowledge on the subskill studied today?; 2. How much effort did you put in today's lesson?; 3. What is percentage of problems you solved in one attempt? Like above, learners answered by drawing the bars below the questions, see the left side of Figure 1. Learners based their answers with regard to *achievement* on the ability score indicated by the ALT. Next, students were asked to compare part 1 with part 2 to determine their *progress* and to see how far they are from reaching their goal. This stage was designed to act as a *feed-forward* intervention in which learners clearly articulated progress towards their learning goal and engage in cognitive evaluation.

Before the rehearsal lesson, the learners were asked to review all their dashboards and determine which subskills they need to work on in the rehearsal lesson. Again students set goals for the rehearsal lesson and evaluate on those before working on the post-test. Thus the feed-up, feed-forward cycle is repeated 4 times during the experiment.

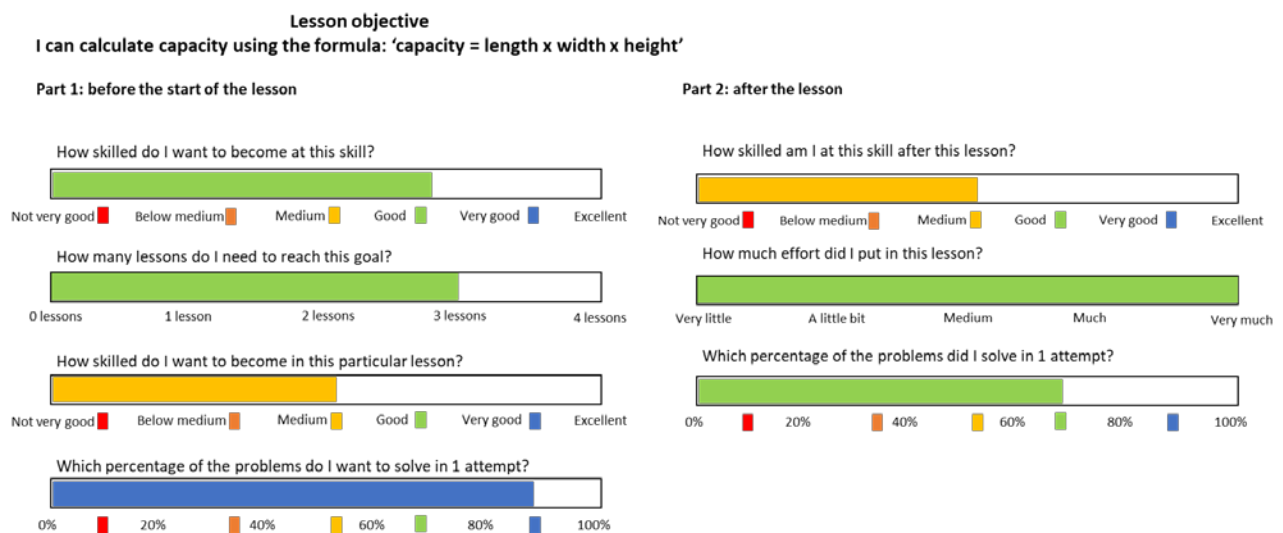


Figure 1. Dashboard drawing by a learner.

2.2.2 Dashboard B: The learning path app

In study B, learners were asked to set a goal at the start of each lesson in the *learning path app*¹. In the overview screen, learners clicked on the dolphin of a particular arithmetic subskill. Then they were shown the goal setting screen, see Figure 2. In this screen, learners were asked to indicate how skilled they wanted to become at that particular subskill and what their goal was for this lesson. The learners filled in their goals by moving the flag on a scale from 0 to 100%. This stage was designed to act as a feed-up intervention in which learners clearly articulated their learning goal and set their standards to evaluate their progress.

After the lesson, learners were asked to look at their progress in the overview screen and in the goal setting screen. On the overview screen learners can see their combined progress on all the three subskills which was communicated by the position of the dolphin. The position of the dolphin on the horizontal level indicates the ability score of the learner as calculated by the ALT. Hence the more to the right the better you know this subskill. Additionally, the size of the dolphin increases with the number of problems solved so this gives an indication of the number of problems a student made for the progress made. Moreover, the dolphins colour provides information about the progress in relation to the overall learning goal set. A grey dolphin indicates no learning goal is set, an orange dolphin indicates the learners has not yet reached their personal learning goal and a green dolphin shows that the learning goal is reached. The hoop around the dolphin indicates that the lesson goal is reached, but the end goal for this skill is not yet reached. This stage was designed to act as a feed-forward intervention in which learners clearly articulated progress towards their learning goal and engage in cognitive evaluation.

When learners click on a dolphin, they go to the goal-setting screen with more detailed information on the learner' progress. The blue bars indicate progress based on the ability score as calculated by the ALT. When the ALT did not yet provide an ability score, learners were shown a grey bar. The colour of the flag shows how this progress is related to the goals set. An orange flag indicates that the learner has not reached their goal yet and a green flag indicates that particular goal is reached.

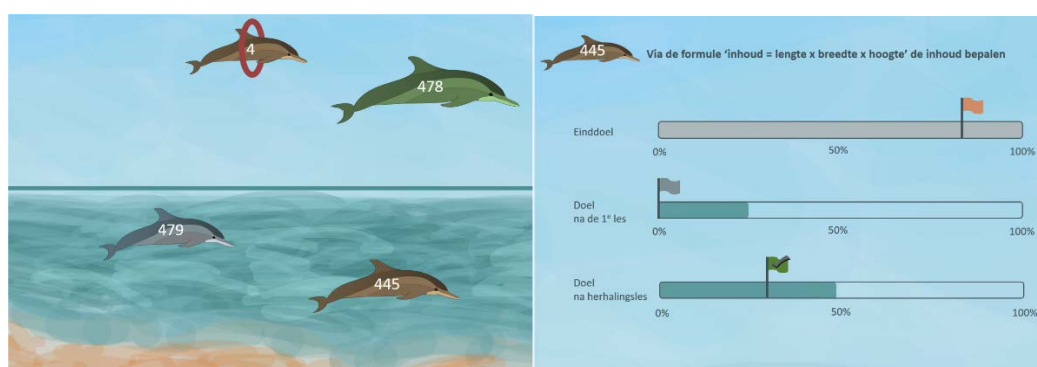


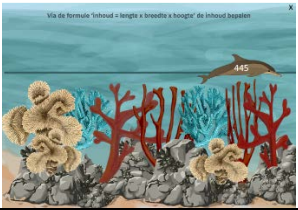
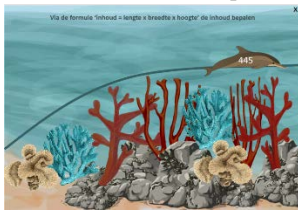
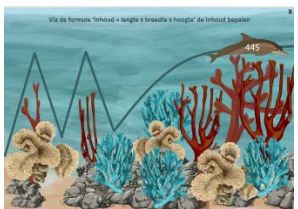
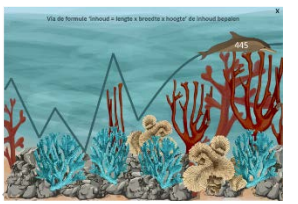
Figure 1. Goal setting screen

When learners click on the progress bars, they go to the *personalized visualizations screens*. Here learners see the learning paths they followed for a particular subskill. The learning paths show how

¹ Leerpaden app in the google appstore

a learner's learning evolved during the practice activities. The personalized visualizations are based on the Moment-by-Moment Learning Curves calculated based from the ALT data. Learners were shown 5 types of learning paths called high swimmer (immediate drop), quick swimmer (immediate peak), climber in two steps (double spikes), slow climber (close multiple spikes) and climber and descender (separated multiple spikes), see Figure 3. The learning path visualize how learners actions contribute to their achievement and show their progress over time. To make these visualizations actionable, learners are explained the meaning of the learning paths. On the poster students are also given actionable feedback to adapt their learning. For example, when a learner showed a close multiple spikes this means that he/she learned the skill slowly and that more practice is still needed. Students are advised to actively monitor their accuracy and increase their effort to ensure they are practicing at their level. Hence, these patterns may help learners understand the development of their effort and accuracy during a lesson and subsequently triggering adaptation.

The feedback is printed on posters that are positioned central in the classroom for all learners to see. Additionally, teachers are given instructions to support learners to understand the learning paths and their implications. A protocol was provided to the teachers that explicitly discusses the function of each step in the intervention. Moreover, teachers are instructed to help learners formulate which actions they could take depending on their learning paths

Personalized dashboards	Planning	Monitoring
High swimmer: Immediate drop 	You already know this skill. → Please practice a different skill.	Your accuracy is high, well done!
Quick riser: Immediate peak 	You have learned this skill quickly after the teacher explained it. → You can practice until you have reached proficiency (green dolphin) and then continue on the next skill.	Your accuracy is high, well done!
Riser in two stages: Double Spikes 	You have learned this skill in two stages during guided instruction and class wide practice. → Please practice until you have reached proficiency.	→ Please monitor your accuracy during practice. → Do you feel that you can put in a little more effort? Try to become a quick riser!
Slow riser: Close multiple spikes 	You are learning this skill somewhat slowly. → Please continue to practice in adaptive mode until you have reached proficiency.	→ Please monitor your accuracy during practicing. → Do you feel that you can put in a little more effort? Try to become a riser in two stages!

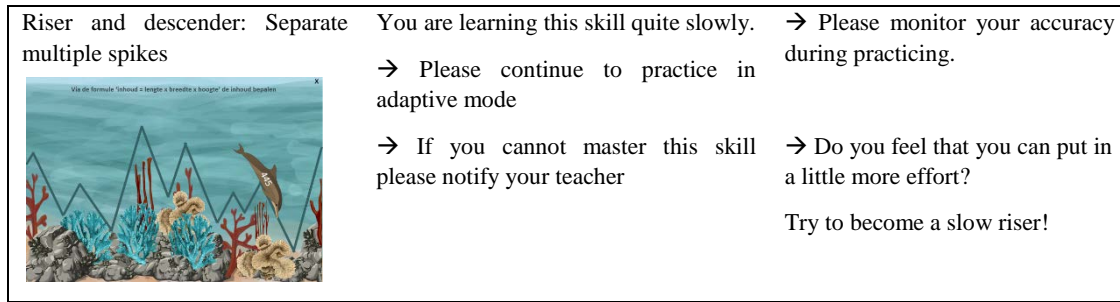


Figure 3. Personalized dashboards

3 EVALUATION FRAMEWORK

We have evaluated the dashboards in two experimental studies. The experiments examine the effects of the dashboard intervention on learning outcomes and transfer of knowledge. Effort and accuracy are included as indicators of self-regulated learning.

Study A evaluates dashboard A and consisted of 71 learners in grade 4 who were divided over the experimental goal setting condition ($n=37$) and the control condition ($n=34$). Study B investigates the learning path app with 93 learners divided over the experimental personalized visualizations condition ($n=63$) and the control condition ($n=30$). Both studies followed a similar design in which learners worked on 3 arithmetic skills in 4 lessons of 50 minutes, see Figure 4. The lessons consisted of a mix of teacher instruction and practice activities. The three skills were easy, medium and hard in terms of difficulty. Learners' learning was measured with a pre and post-test and a transfer-test.

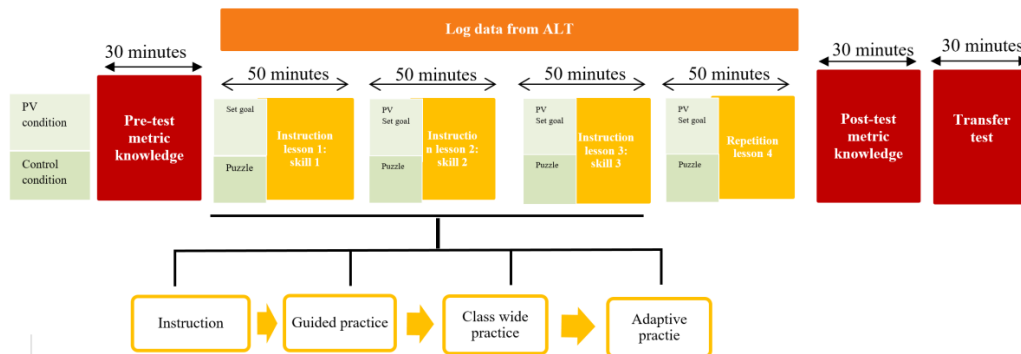


Figure 4. Study design

3 PRELIMINARY RESULTS

Study A. A repeated measurement ANOVA was used to investigate the effect of the dashboard on learning with pre and post-test as within subject variables and condition as between subject variable. The results showed a significant main effect of Time $F(1, 69) = 89.13, p < .001$. All learners post-test scores ($M = 19.01, SD = 3.56$) were higher compared to the pre-test scores ($M = 14.03, SD = 5.31$). We also found a significant interaction effect between Time *Condition $F(1, 69) = 4.09, p =$

0.05. Learners in the experimental condition made more progress ($M = 6.00$, $SD = .25$) than learners in the control condition ($M = 3.88$, $SD = .26$). An ANOVA showed a significant difference on the transfer test $F(1,69) = 5.15$, $p = .026$. Learners in the experimental condition scored lower on the transfer test ($M = 10.19$, $SD = 3.97$) than learners in the control condition ($M = 11.97$, $SD = 2.36$).

Study B. Data are currently analysed and will be ready for presentation at the workshop. We expect that learners in the personalized visualization condition will outperform learners in the control condition both on learning outcomes as well as their effort and accuracy regulation.

4 SCIENTIFIC SIGNIFICANCE

In this paper we outlined the design of two dashboards to support learners' regulation. These dashboards are grounded in the COPES theory of self-regulated learning. We propose a comprehensive approach towards learner-faced dashboards that includes learners' achievement, information on progress and the learning path which connects learners' actions to their progress. This transforms the role of dashboards from discussing *what* learners learned to also incorporating *how* learner learned. In this way dashboards could be a promising way to overcome learners' utilization deficiencies to effectively apply self-regulated learning. Unique to these dashboards is that trace data is used to help students understand their regulation in learning paths. MbMLC are used to help learners understand how their actions relate to progress.

These dashboards are designed to function as a reference for learners and to support learners to engage in cognitive evaluation. Prior to learning, the feed-up intervention ensures students set standards and formulate learning goals. After learning, the feed-forward intervention helps learners to translate the dashboard data into adaptations that help them to proceed towards their goals. The explicit instructions show how learners can be supported to follow up on the provide data on achievement, progress and learning paths. This provides a very transparent interface into how data are transformed into actionable feedback for learners.

The preliminary results indicate that these dashboard indeed improved learners learning, but did not enhance transfer of learners' knowledge. When differences are found in learner effort and accuracy, this may imply that the intervention also affects how learners regulate their learning. Additional effects of personalized visualizations will be presented at the workshop.

REFERENCES

- Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016a). Instruction Based on Adaptive Learning Technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of Research on Learning and Instruction* (2nd ed., pp. 522–560). New York: Routledge. <https://doi.org/10.4324/9781315736419.ch24>
- Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Mehranian, H., Fisher, D., Barto, A., Mahadevan, S. and Woolf, B. (2007). Repairing disengagement with non invasive interventions. *International Conference on Artificial Intelligence in Education*, (August 2016), 195–202.
- Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning*, 4(1), 87–95. <https://doi.org/10.1007/s11409-009-9035-7>
- Azevedo, R., Martin, S. A., Taub, M., Mudrick, N. V., Millar, G. C., & Grafsgaard, J. F. (2016). Are

- Pedagogical Agents' External Regulation Effective in Fostering Learning with Intelligent Tutoring Systems? In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 197–207). Cham: Springer International Publishing.
- Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development*, 56(1), 45–72. <https://doi.org/10.1007/s11423-007-9067-0>
- Baker, R. S., Hershkovitz, A., Rossi, L. M., Goldstein, A. B., & Gowda, S. M. (2013). Predicting Robust Learning With the Visual Form of the Moment-by-Moment Learning Curve. *Journal of the Learning Sciences*, 22(4), 639–666. <https://doi.org/10.1080/10508406.2013.836653>
- Baker, R. S. J. D., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, 21(1–2), 5–25. <https://doi.org/10.3233/JAI-2011-015>
- Bannert, M., Hildebrand, M., & Mengelkamp, C. (2009). Effects of a metacognitive support device in learning environments. *Computers in Human Behavior*, 25(4), 829–835. <https://doi.org/10.1016/j.chb.2008.07.002>
- Bannert, M., Molenaar, I., Azevedo, R., Järvelä, S., & Gašević, D. (2017). Relevance of learning analytics to measure and support students' learning in adaptive educational technologies. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*. <https://doi.org/10.1145/3027385.3029463>
- Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open Learner Models and Learning Analytics Dashboards: A Systematic Review. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. <https://doi.org/10.1145/3170358.3170409>
- Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*. <https://doi.org/10.1109/TLT.2017.2740172>
- Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, 65(3), 245–281. <https://doi.org/10.3102/00346543065003245>
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let ' s not forget : Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
- Hattie, J., & Timberley, H. (2007). The power of feedback. *Medical Education*, 44(1), 16–17. <https://doi.org/10.1111/j.1365-2923.2009.03542.x>
- Hattie, J., & Timperley, H. (2007). Review of Educational The Power of Feedback. *Review of Educational Research*. <https://doi.org/10.3102/003465430298487>
- Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate: preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18*. <https://doi.org/10.1145/3170358.3170421>
- Kennisnet. (2014). Onderwijs met een eigen device. Retrieved from https://www.kennisnet.nl/fileadmin/kennisnet/publicatie/Onderwijs_met_eigen_device.pdf
- Klinkenberg, S., Straatemeier, M., & Van Der Maas, H. L. J. (2011a). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, 57(2), 1813–1824. <https://doi.org/10.1016/j.compedu.2011.02.003>
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., Mark, M. A., Koedinger, K. R., Anderson, J. R., ... Mark, M. A. (2007). Intelligent Tutoring Goes To School in the Big City To cite this version : HAL Id : hal-00197383 Intelligent Tutoring Goes To School in the Big City.
- Molenaar, I., Horvers, A., & Baker, R. S. J. D. (submitted). What can Moment-by-Moment learning curves tell about students' self-regulated learning?
- Molenaar, I., Horvers, A., & Baker, R. S. J. D. (2019). Towards Hybrid Human-System Regulation: Understanding Children' SRL Support Needs in Blended Classrooms. In *In proceedings of the 9th*

International learning analytics & knowledge conference.

- Molenaar, I., & Järvelä, S. (2014). Sequential and temporal characteristics of self and socially regulated learning. *Metacognition and Learning*, 9(2). <https://doi.org/10.1007/s11409-014-9114-2>
- Molenaar, I., & Knoop-van Campen, C. (2018). How Teachers Make Dashboard Information Actionable. *IEEE Transactions on Learning Technologies*. <https://doi.org/10.1109/TLT.2018.2851585>
- Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Shirvani Boroujeni, M., Holzer, A., ... Dillenbourg, P. (2017). Understanding learning at a glance: A systematic literature review of learning dashboards. *IEEE Transactions on Learning Technologies*.
- Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, 45(4), 267–276. <https://doi.org/10.1080/00461520.2010.517150>
- Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology*. <https://doi.org/10.1111/bjep.12173>
- Winne, P. H., & Baker, R. S. J. d. (2013). The Potentials of Educational Data Mining for Researching Metacognition, Motivation and Self-Regulated Learning. *JEDM - Journal of Educational Data Mining*, 5(1), 1–8. <https://doi.org/10.1037/1082-989X.2.2.131>
- Winne, P. H., & Hadwin, A. F. (2013). nStudy: Tracing and supporting self-regulated learning in the Internet. In *International handbook of metacognition and learning technologies* (pp. 293–308). Springer.
- Winne, P. H., & Hadwin, A. F. (2017). Studying as self-regulated learning, (January 1998).
- Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In *Metacognition in educational theory and practice*. <https://doi.org/10.1016/j.chb.2007.09.009>
- Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*. <https://doi.org/10.1145/2567574.2567588>
- Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. *Handbook of Self-Regulation*, 13–39. <https://doi.org/10.1016/B978-012109890-2/50031-7>

Visualisation of key splitting milestones to support interventions

Martin Hlosta, Jakub Kocvara, David Beran, Zdenek Zdrahal

Knowledge Media institute, The Open University

martin.hlosta@open.ac.uk

ABSTRACT: The paper presents an approach to help staff responsible for running courses by identifying key milestones in the educational process, where the paths of successful and unsuccessful students started to split. By identifying these milestones in the already finished courses, this information can be used to plan the interventions in the next runs. This is achieved by finding the earliest time when the differences in behaviour or key performance metrics of unsuccessful students start to become significant. We demonstrate this approach in two case studies, one focused on a course level analysis and the latter on a whole academic year. This suggests its generic nature and possible applicability in various Learning Analytics scenarios.

Keywords: Learning Analytics, Visualisation, At-Risk Students, Intervention support

1 INTRODUCTION - SETTING THE SCENES

Identifying students at-risk of failing either the course or the whole qualification is a very topical issue of Learning Analytics. Further analysis of reasons why the student is lagging behind may suggest interventions that guide him/her to the successful completion of the course (Jayaprakash et al., 2014). Usually, two sources of data are available: data about the student and data about the course. Student data include their demographics, their study history, and activities within the course. The data related to the course are the study plan i.e. study materials, dependencies between different study resources, time allocated to each task, assignment to be completed by the student to prove that he/she has mastered the expected content and progression rules, which define criteria of student's success or failure in the course. Often student data from previous presentations of the course are available and machine learning techniques can be used for developing predictive models (Wolff et al., 2014). This problem specification applies both to classroom-based and to distance education. One of the typical issues is selecting a moment in the course to use the predictive model for interventions so that the predictions are accurate yet early enough for at-risk students to get back on track. Howard et al., (2018) selected this point based on manual inspection of decrease of the error between week 4 and 5.

In this paper, we offer a different view on the learning analytic tasks. As mentioned above, by assessing whether the student satisfied the course progression rules, we distinguish two groups of students: those who pass and those who fail. In fact, we may extend this dichotomy by an additional group of students who have not met all progression rules, but there is a reasonable chance that they can complete the missing requirements in the future and finish the course. For example, the student has not acquired all credits required to successfully pass, but then he/she has earned enough key credits and therefore may be allowed to continue and complete the missing credits in the next years. Consequently, we may distinguish three groups of students denoted as fail, continue and pass.

By analysing already completed course presentations, we have noticed that there are "points" in the study plan where the "homogeneous" cohort divides into two or three of these groups. This split can

be verified by a suitable statistical test and without early intervention, it is usually persistent to the end of the presentation. Once the student starts losing pace with the study plan, the gap is likely to grow and eventually, the student may resign and fail. The same situation can apply to the continue/pass split. Such points are usually identified by the manual analysis. For example, Simpson (2004) identified different withdrawal routes of students by showing the proportion of students not submitting their assignments using the 'river' diagram and only very few of them returning to submit the next ones. In the same paper, he suggested that different withdrawal types might indicate different interventions with students. Coffrin et al. (2014) presented state transition diagrams for students who completed the course and those who did not. Using these diagrams, users can observe transitions of students between the assessments and the differences between the completed and non-completed groups, although these differences are not stated explicitly. Teasley (2018) mentions this identification of important points in courses when discussing what it means to do learning analytics, referring to finding a "point of no return" when poorly performing students are likely not to succeed in the course.

The recent survey analysing 52 papers in Visual learning analytics found that most of the work focuses on Understanding Collaboration and Instructional Design, with analytics on students for instructors being most prevalent (Vieira et al., 2018). Some of the work focuses on time changes, especially students progressing in the course, e.g. a simple approach in (Breslow et al., 2013) using line plots to show different activity types used in different weeks. Chen Y. et al., (2016 October) helps to explain the behaviour of students in different clusters based on their predictions and actual results. Moreover, some papers support the identification of interesting points in time. Chen Q. et al. (2016) visualises the peaks in the videos from the clickstream to better design the videos in the future.

The aforementioned approach in (Corfin et al, 2014) can be used to identify points when students start to drop out and also the one in (Hlost et al., 2014) to spot the typical patterns of students before the first assignment leading to failure.

We have demonstrated, that if the pattern of characterising that the students are approaching split point is identified before the split became persistent and the instructors intervene, the student retention or successful completion can be dramatically improved. Identifying the split points will be demonstrated and visualised in the following sections. This builds on our previous work (Zdrahal et al., 2016) and also (Wolff et al., 2014) and its aims to provide a generalisable and visual approach for early phases of Learning Analytics process.

To conclude, there is work that highlights the identification of the milestones to support the intervention. Moreover, some of existing research in visual analytics can help with this identification but to the best of our knowledge, there is a gap in automatic identification and visualisation of these milestones during the learning process. Also, the existing papers focus on a limited context, such as MOOCs, closed classroom. Providing that relevant data are available, our work aims at generalizing across different learning contexts.

First, we provide the description according to 5 questions from the workshop proposal call for paper. Then, we present two case studies from different learning scenarios showing the visualisations and concluding with the further work.

2 THE APPROACH

2.1 What kind of data is being visualized? What tools were used to clean up the data?

The visualisation expects data from the university system with the final result of students in either a course or a whole academic year. The approach expects partial measurements of students' progress towards achieving the learning goal recorded as events in time. These usually include assessment scores, optionally weighted by their importance. In addition, data of any recorded student activities can be used.

The pre-processing has been performed using SQL and Python with its common libraries for manipulating data, i.e. Pandas and Numpy and SciPy for statistical evaluation¹.

2.2 For whom is the visualization intended?

The visualisation has been designed for staff responsible for running the courses, potentially for researchers in Learning Analytics. Realising the key milestones, the course directors receive hint when to plan the interventions or where the design of the course might be updated. The users are not expected to be experts in visualisation, they should be familiar with the structure of the course.

2.3 Why: what is the goal of the visualization? What questions about the data should it answer?

The goal is to support the identification of **important milestones** in a course or academic year using visualisation. It should answer questions such as: When does the difference in measured value between successful and unsuccessful students start to be statistically significant? When is a convenient time to make interventions for poor performing group provided that a similar pattern of student behaviour will prevail the next run? What is the best splitting value of the measured value between the groups of students in time?

We expect the approach to be used for initial course analysis before building a machine learning algorithm that might be more complex and resource expensive. On a higher level of abstraction, the usage workflow consists of the following steps:

1. Identify the indicator of students' progress, e.g. assessment score, number of credits. Visualising this should provide the first insight of where the students start to split.
2. Select a behavioural characteristic of students, e.g. number of clicks, time spent in the VLE and use the visualisation on a more granular level.

2.4 How is data visualized and why? Tools, libraries, data formats used for technical implementation? What workflows and recipes can be used to develop the visualization?

The data is visualised using the line plot representing the median for each performance group, with the variance of the captured metric between the 25th and 75th percentile. The variance is shown using the same colour with added transparency level. This was a preferred variant over boxplots as they would make the graph more challenging to read, especially when shown for more performance groups. The first identified milestone is visualised using the vertical line through the whole graph,

¹ Pandas - <https://pandas.pydata.org/>, Numpy - www.numpy.org, SciPy - <https://www.scipy.org/>

with bold style in the region between the two medians, where the difference was measured. The black horizontal line denotes the best split between the performance groups.

The measurements are taken in various regular time intervals during the whole duration of either a course or the academic year, typically days or weeks. The approach provides retrospective analysis, so the results of students are required to assign students in the performance groups.

Python with its data manipulation libraries and matplotlib² for visualisation have been utilised. Similar results might have been also achieved with R or with some javascript library.

The approach consists of four steps:

1. Preparing the common data input format - This includes extracting the source data of student events and converting them in a time-sliced data table, where all students have records of all available measurements, i.e. not only when they change.
2. Identifying the important milestones - starting at the beginning of the measured period, the algorithm continuously examines the difference between the successful and unsuccessful students. In each time slice, a statistical test is performed to detect if the difference in the observed metric is statistically significant. If the conditions for unpaired t-test are met (normality of both group distributions and homoscedasticity), it is used as a preferred variant. Otherwise, the Wilcoxon rank sums test is used.
3. Best Splitting values - starting in the identified milestone, for each following time the best splitting values in the measurement is computed by minimising the error of that split, i.e. proportion of wrong predictions to all predictions. It represents the quality of the predictions that would be achieved if this splitting point was used to classify students into good and at-risk student groups.
4. Visualising the lines, variance bands, early milestone and splitting points. The graph can be enhanced by adding manually annotated events, e.g. the start of Christmas break, dates of the assessments, etc.

2.5 How has the approach been evaluated or how could it be evaluated?

The quality of each milestone split can be evaluated in terms of statistical significance. The approach counts with taking the data distribution into consideration. In each point, we can also compute the error of the split that is made based on this factor.

The goal of the visualisation is to convey a clear message to either researchers or course designers to help them understand when the intervention should happen. Understanding this and acceptance of this information can be viewed as one of the key evaluation strategies. As the next step we want to run a user study with 10-15 participants and various types of roles - i.e. tutors, course designers and researchers. We plan to use a combination of a questionnaire designed by the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh, 2003) combined with open ended questions. The UTAUT uses four constructs (performance expectancy, effort expectancy, social influence and facilitating conditions) to explain the users' technology acceptance and use. The open ended questions will focus on providing information about current actions around the identified points in time, about perceived importance and potential interventions that might be possible to plan. We have conducted a similar procedure during the case study 1 in the past but in more informal way.

² Matplotlib - <https://matplotlib.org/>

2.6 Encountered problems and pitfalls during the visualization process?

One of the problems was examining zero values, i.e. if zero measurement should be included/excluded from the statistical tests. The other challenge was making the approach generic enough to cope with various x-axis unit, in our case days relatively counted from the start of the course or calendar dates.

3 TWO CASE STUDIES

3.1 Classroom-based university – Progress through the academic year

A faculty from a classroom-based university with the face to face teaching had poor progression rates of their first-year students. The students acquire credits by completing one-semester long courses prescribed in the study plan. Acquiring credits is stored as events. Typically, there are 6-8 courses per semester, the number of credits earned in the course depends on its difficulty and importance for the study program. Based on the number of credits earned at the end of the two-semester academic year, a student falls into one of the four groups (fail, fail-winter, continue, pass). Groups “pass” and “continue” progress to next study year, students in the “fail” group are deregistered, “fail-winter” fail even before the end of the winter semester. The trajectory of students is shown in Figure 1. We are interested in the difference between the “fail” and “continue” groups. There are 943 students in total, i.e. 245 pass, 198 continue, 54 fail, 446 fail-winter. The number of credits within the groups is not normally distributed, neither the homoscedasticity has been satisfied, hence Wilcoxon sum rank test was used. The groups start splitting before the Christmas break, meaning that students who have not collected enough credits at that time are already at risk. By the end of the winter exam period, the inter-group differences are very noticeable. The flat part that follows, corresponds to the period of lectures in the summer term usually without credit-earning exams. Next opportunity for earning further credits is in the summer exam period. Though the winter and summer exam periods are well-defined, the examiners may offer a few “early exam terms” up to 4 weeks before the start of the exam periods. It is visible in Figure 1, that the “pass” students take this opportunity more often than the students of the “continue” group. Moreover, Figure 1. shows, that the students in the “fail” group do not earn significant (if any) credits before the start of exam period.

This visualisation triggered a conversation with the faculty management and led to designing a precaution intervention strategy, reminding this to all the students and then repeating this to the ones that haven’t collected enough credits. To the great surprise of university academics and ourselves, this has resulted to the increase of students progressing to the second year by 49%. Specifically, comparing with the best year so far, 49% of students expected to fail progressed to another year. The letter of recognition of the faculty dean is available on 3.

³ The letter of recognition of the faculty dean available on our website:
https://analyse.kmi.open.ac.uk/resources/documents/letter_of_recognition.pdf

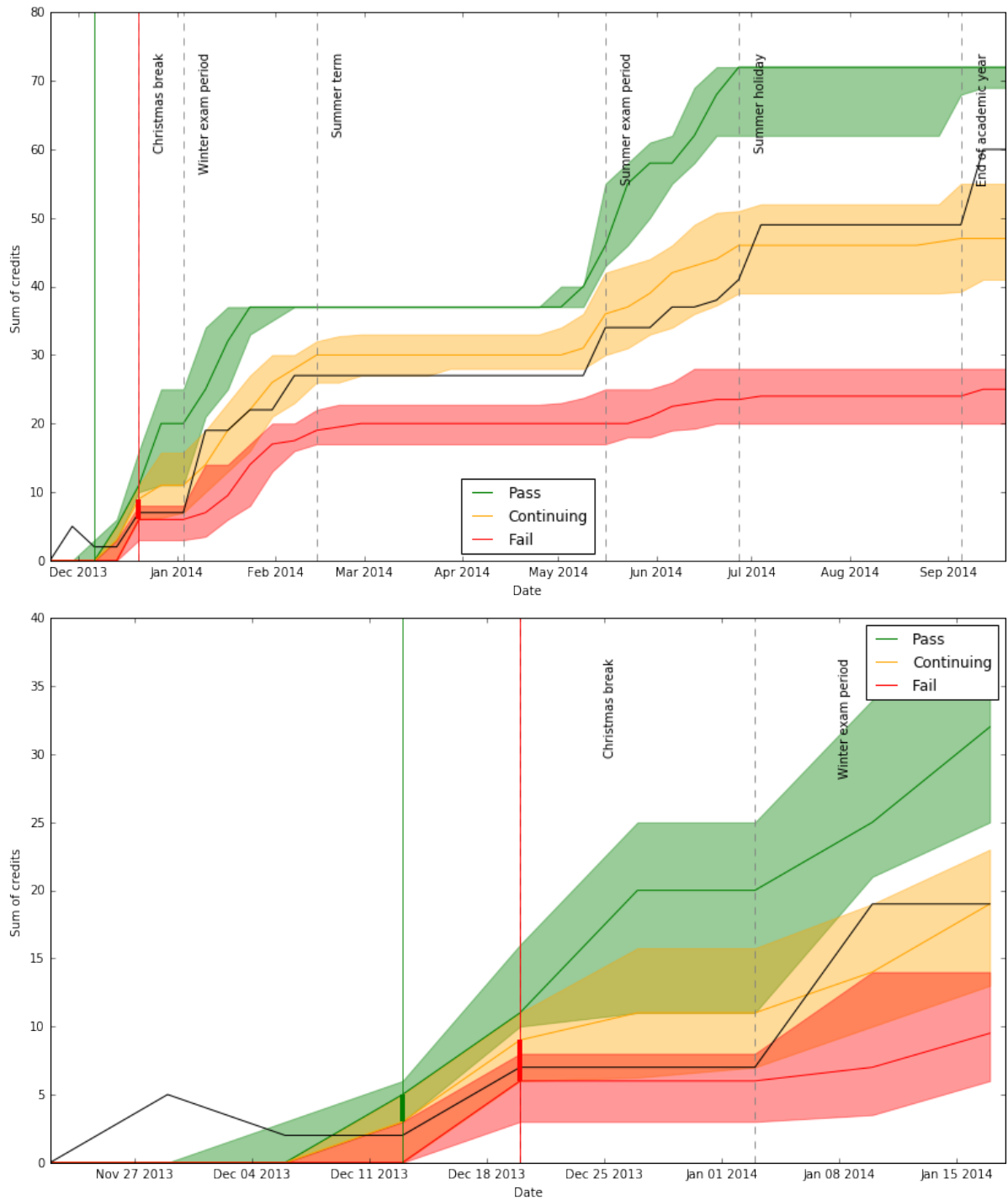


Figure 1: a) case study - students achieving credits in face to face university b) zoomed view highlighting the first statistically significant difference between the groups.

3.2 Distance education course

The second example comes from a publicly available OULAD dataset from the Open University (Kuzilek, 2014). Using this dataset allows better reproducibility of this approach. We selected a level-one course that is fully online - EEE/2014J. The rest of the courses in 2014J can be found in the GitHub repository⁴. Students gain a score after submitting their assessments, which enable them to pass the course. Their final result is either Distinction, Pass, Fail or Withdrawn. Moreover, student

⁴ Github repository with figures – <https://github.com/hlostam/milestone-vis>

sum of clicks per days is captured. We used the weighted assessment score to account for the importance of the assessment. Figure 2 shows that the first important difference is just after the submission of the first assessment with the best splitting point for score 13.

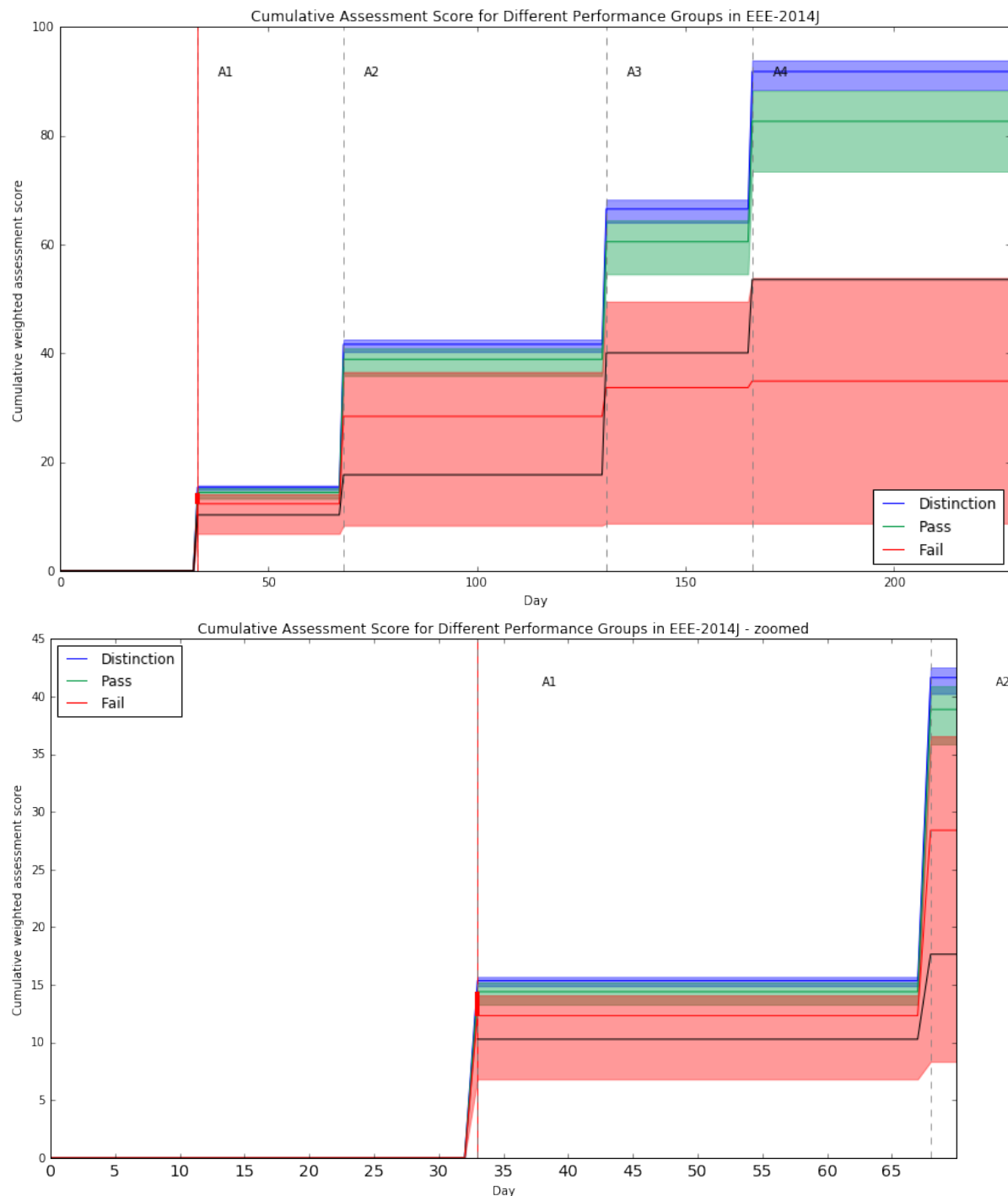


Figure 2: a) Distance education course and acquiring the weighted assessment score b) the detail highlighting the first statistically significant difference between the groups.

This might justify focusing on intervening even before the first assessment. Focusing on more detailed student online behaviour, Figure 3 shows that for the sum of the clicks in VLE, the first observed difference between both Failed and Passed and between Passed and Distinction is in the first day of the course.

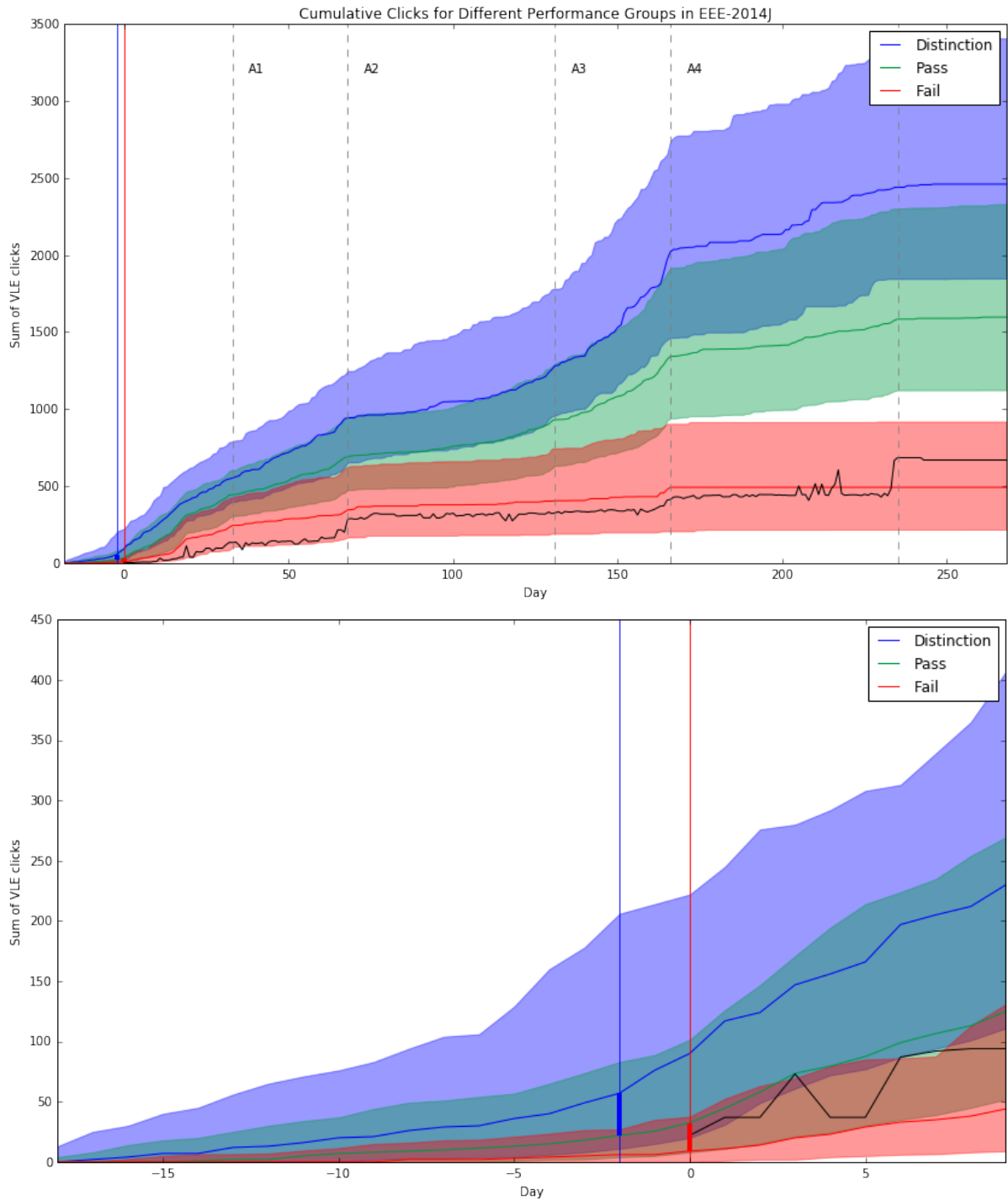


Figure 3: a) Number of clicks in distance education course for different performance groups b) the detail highlighting the first statistically significant difference between the groups.

It should be mentioned that the identified key milestones do not mean that potential predictive models would be accurate enough to split between the successful and unsuccessful students. It gives only a signal that starting this point, the differences between the behaviour of these two groups in terms of the measured variable started to be statistically significant.

4 CONCLUSIONS AND FURTHER WORK

Until now we have deployed this framework in three case studies, two of which we share here. In the case of the conventional university, the usefulness and impact of the approach have been

demonstrated by successfully improving the retention by about 49% in two consecutive years. In both cases we compare results with the lowest retention achieved in 2013/14 i.e. before the described predictions and interventions have been deployed. Our current focus is to include the study history of the students, which might help to identify groups of students where interventions might have higher impact.

REFERENCES

- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment*, 8, 13-25.
- Chen, Q., Chen, Y., Liu, D., Shi, C., Wu, Y., & Qu, H. (2016). Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses. *IEEE Transactions on Visualization & Computer Graphics*, (10), 2315-2330.
- Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., & Qu, H. (2016, October). DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. In *Visual Analytics Science and Technology (VAST), 2016 IEEE Conference on* (pp. 111-120).
- Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 83-92). ACM.
- Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., & Wolff, A. (2018). Modelling student online behaviour in a virtual learning environment. In: *Machine Learning and Learning Analytics Workshop at The 4th International Conference on Learning Analytics and Knowledge (LAK14)*, 24-28 Mar 2014, Indianapolis, Indiana, USA.
- Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education*, 37, 66-75.
- Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6-47.
- Kuzilek, J., Hlosta M., Zdrahal Z. (2017). Open University Learning Analytics dataset In: *Sci. Data* 4:170171 doi: 10.1038/sdata.2017.171.
- Klerkx, J., Verbert, K., Duval, E. (2017). Learning analytics dashboards. In: *Handbook of Learning Analytics*, Chapt. 12 Society for Learning Analytics Research.
- Simpson, O. (2004). The impact on retention of interventions to support distance learning students. *Open Learning: The Journal of Open, Distance and e-Learning*, 19(1), 79-95.
- Teasley, S. D. (2018). Learning analytics: where information science and the learning sciences meet. *Information and Learning Science*.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.
- Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122, 119-135.
- Wolff A., Zdrahal Z., Herrmannova D., Kuzilek J., Hlosta M. (2015). Developing predictive models for early detection of at-risk students on distance learning modules In: *Machine Learning and Learning Analytics Workshop at The 4th International Conference on Learning Analytics and Knowledge (LAK14)*, 24-28 Mar 2014, Indianapolis, Indiana, USA.
- Zdrahal, Z., Hlosta, M., & Kuzilek, J. (2016). Analysing performance of first year engineering students. In: *Data Literacy for Learning Analytics Workshop at Learning Analytics and Knowledge (LAK 16)*, Edinburgh, UK.

Using Jitter and Sampling Techniques to Improve the Comprehensibility of Scatter Plots: A Practical Example

Michael D. Kickmeier-Rust

University of Teacher Education, St. Gallen, Switzerland

michael.kickmeier@phsg.ch

ABSTRACT: Displaying complex data including the interrelationships of several variables is one of the key challenges for information visualization. This is particularly true when the target audience has little data literacy, which is oftentimes the case in the context of learning analytics (where stakeholders are students, parents, teachers, or administrators). In this paper, I introduce a practical scenario in the context of the Swiss educational system and present an innovative solution to display complex learning data with scatter plots. By techniques such as jitter and data sampling, the scatter plot can be advanced and presented in a more comprehensible way, even when large data sets are displayed.

Keywords: Learning Performance, Performance Comparison, Visualization, Scatter Plot, Jitter

1 INTRODUCTION: SWISS TEST AND TRAINING PLATFORMS

The new Swiss national curriculum *Lehrplan 21*, released in 2015, describes the educational policy for compulsory schools. It sets the educational goals at all school levels and informs all stakeholders about the competencies to be achieved in compulsory education. *Lehrplan 21* divides the eleven years of compulsory schooling into three cycles: (i) kindergarten, 1st and 2nd grade, (ii) 3rd through 6th grade, and (iii) 7th through 9th grade. The curriculum, furthermore, breaks the subjects down into competence areas, which focus on skills/abilities (e.g., listening, reading, speaking, writing in the languages) as well as thematic areas (e.g., “numbers and variables” in mathematics). Within these areas, competencies are defined in the form of typical “I can” statements, pointing to the abilities, which students are intended achieve at the end of each of the three cycles. The set of competencies in each of the subjects are ordered by seven competence levels, which are summarizing descriptions of the abilities and competencies the students hold. There is a strong relation to developmental theories (cf. Siegler et al., 2014): the levels are characterized by an increase in factual, conceptual, and procedural knowledge, by an increase in perceptive demands (e.g., speech comprehension), by increasingly complex application scenarios as well as the degree of self-regulation and independence that need to be applied. Related to *Bloom’s Taxonomy* of cognitive development (Anderson, 2013), a higher level encompasses the abilities and competencies of a lower level.

In a number of Swiss cantons (e.g., St. Gallen or Zürich), online-based test and training platforms are deployed; *Lernlupe* (www.lernlupe.ch) for 3rd – 6th grade and *Lernpass plus* (www.lernpassplus.ch) for 7th – 9th grade. These platforms provide individual training facilities and standardized online tests along the competencies and levels defined by *Lehrplan 21*. The feedback for students is formative and competence-oriented in nature. For example, students receive a verbal description of their abilities and their current competence level. The results of the standardized tests provide clear indications of strengths and existing competence gaps and they are utilized by the teachers to plan

an individual support of students. The tests, moreover, provide practical indications for career planning. For example, the *Jobskills* platform (www.jobskills.ch) enables a comparison of various job specifications with students' individual competence profiles.

Lernlupe and Lernpass plus feature IRT-based computer adaptive testing functions (cf. van der Linden, 2016). Adaptive testing allows optimizing the assessment quality within minimal testing times. The test items are selected on an individual basis so that the item difficulty matches the estimated ability of the student as exact as possible. The prerequisite for adaptive testing is extensive standardization studies to identify the item characteristics (e.g., item difficulties) based on a representative sample. In the cantons St. Gallen and Zürich, such large-scale studies (5000 students per age group) have been carried out in the past years. Based on these results, a metric between 200, which corresponds to the lowest level of difficulty or ability, and 800, which is the highest value, is established. The mean of this scale is 500. The scale is a "historic" IRT scale and used in a variety of standardized academic achievements tests, for example the GMAT (www.mba.com/exams/gmat).

The test and training platforms Lernlupe and Lernpass plus are used on a frequent basis by cantonal schools and accordingly rich is the basis of available data. The main user groups of the data are students, on the one hand, and teachers and parents on the other hand. The feedback formats generated by the systems are used, for instance, to inform parents about the learning progress of their children. The feedback is designed cautiously and restricted to the individual score (on the 200-800 scale) in relation to the achieved competence levels, accompanied by verbal categorical descriptions of strengths and achievements. An increasingly important aspect of data visualization refers to the identification of performance indicators and performance comparisons for administration and management on a local school level but also on a regional, political level. In comparison to the "traditional" methods of gathering data about the performance of schooling (e.g., the OECD PISA studies), the data from the aforementioned (and other) platforms are more up-to-date, rather longitudinal in nature, and more detailed. This increases the utility of the data significantly - and also the interest of local, regional, and national authorities.

2 VISUALIZING GENERAL PERFORMANCE DATA

To inform stakeholders about general learning performance, for example, of entire cohorts, a visualization type is necessary that includes all relevant information and that particularly offers a comparison of local data (the data of a specific school) with regional data and the data of the standardization studies. Relevant variables are cohort/age group, subject and competence areas, gender, as well as the temporal progression over years. The dependent variables are student performance on the 200-800 scale and the achieved competence levels. The challenge is to develop a form of visualization that conveys the meaning of these complex data in a very simple way.

For Lernlupe and Lernpass plus we designed a set of visualization formats including density diagrams and pie charts. A group of experts and users chose a scatter plot approach as the most intuitive form of visualization (cf. Figure 1). The main advantage of the scatter plot was seen in the fact that each student can be represented as an individual point. This was considered being highly intuitive to understand for a broad variety of users with a broad range of data literacy levels (i.e., children, parents, teachers, school leaders, administrative leaders, politicians).

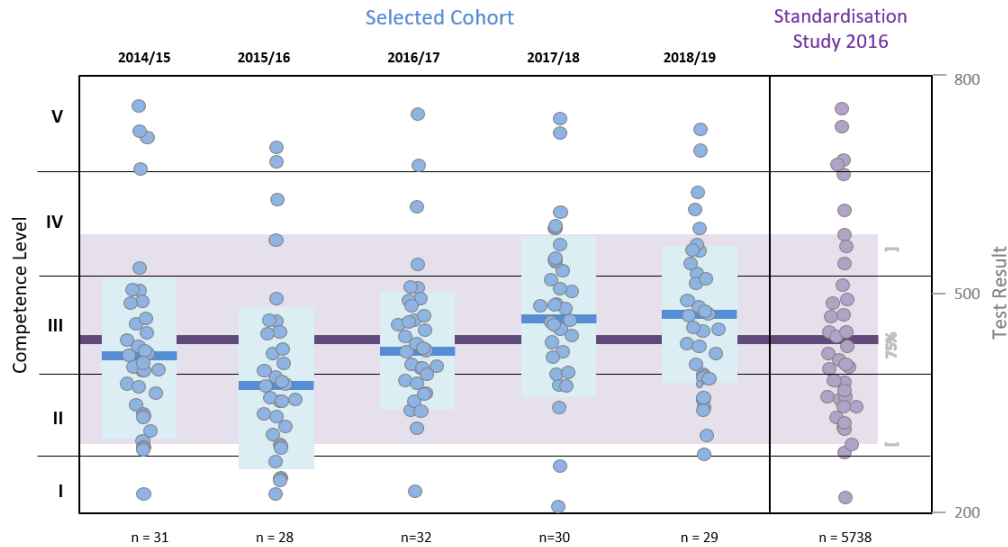


Figure 1: Visualizing learning data with a scatter plot. In this example, the performance of boys and girls is displayed on a yearly basis in comparison to the results of the standardization sample; the performance is shown as competence levels (scale on the left) and test scores (scale on the right).

3 JITTER AND SAMPLING TECHNIQUES

Scatterplots are amongst the most effective forms of understanding bivariate relationships, since they nicely display the relationship between a variable x and a variable y . However, typical scatter plots are only effective for two continuous variables. If one variable is discrete, other techniques for visualizing the data may be more appropriate. In our case, we have the continuous performance variables, however, only few classes (e.g., gender or years; cf. Figure 1). Also, we face the challenge that comparably small groups of students (e.g., 20 children of a class) are supposed to be compared to huge groups (e.g., 5000 students of the standardization studies). When displaying such a large amount of data points, the scatter plot gets illegibly crowded and confusing. Therefore, the visualization would lose its major strength. To overcome these issues, we developed a visualization approach combining sampling and jittering techniques to improve comprehensibility.

Jittering refers to adding random noise to data in order to prevent data points being over plotted by others. This over plotting specifically occurs when a continuous measurement (such as the standardized competence value) is rounded or aggregated. In large data sets, such as the standardization sample, over plotting is very likely and reduces the comprehensibility of the visualization substantially. Jittering can be done by adding small random changes to the actual values along the x or y -axes.

In our case, we have a small number of discrete classes (e.g., gender or the year), so we applied jittering along these classes (on the x -axis). A purely random jittering, however, may result in a low comprehensibility of the plot. We experienced that users tend to misinterpret the deviations from the center of a class. As a result, we developed an algorithmic jittering, which plots the data points with a minimum overlap to surrounding points. Given two points with exactly the same value, the

points overlap by a certain percentage. The more points exist with the same value, the higher is the percentage of overlapping (and therefore the smaller the visible area of the point) until a maximum overlap is reached. In a second step of the plotting approach, the same principle is applied to points with higher and lower values. By this means, the data points are not randomly jittered but iteratively placed within the outlines of the functional shape of the data (usually a bell curve). Technically, the basis for the algorithmic jitter function is *SinaPlot*, which is an enhanced jitter strip chart package in *R* where the width of the jitter is controlled by the density distribution of the data within each class (Sidiropoulos, et al. 2017, 2018). Figure 2 illustrates the approach.

Sampling refers to selecting only a (more or less representative yet small) sample from a pool of available data and to display this sample in the scatter plot. The technical challenge is to find the right sampling method for a given visualization purpose and a given target audience. In our example, sampling is done when the selected group of students exceeds 99 students and sampling is applied for comparisons with the standardization study.

One sampling method is to select random data points (i.e., students). Assuming that the original data follow a normal distribution, this results in a suitable representation of the original data. This approach, however, cannot guarantee that the most extreme students are represented, which is an important information. Also, this approach works well for very large data pools, such as the standardization group. For smaller pools, for example when a user selects the students of multiple classes, this method likely results in inadequate, most often a too uniform, sampling of observations. Our solution is a threshold-based sampling algorithm; for a specific number of observations (e.g., for 5 students with exactly the same value) only a single point is plotted. The algorithm also assures that for all individual data values, at least one point is maintained. When the number of points exceeds a maximum number (e.g., the aforementioned 99 points), the threshold is recursively raised. Moreover, the threshold follows a normal distribution, meaning that the threshold value is higher in middle areas than at the tail ends. This method allows displaying a distribution as close to the shape of the original one, with losing as little individual values as possible and without losing the most extreme values.

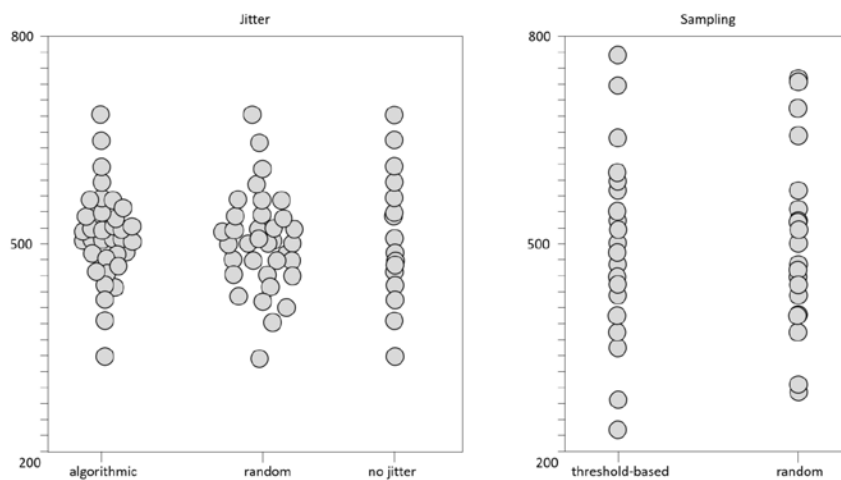


Figure 2: The left panel shows a comparison of the algorithmic jitter function and random jitter, opposed to the plot without jitter; the right panel shows a comparison of sampling methods

4 CONCLUSIONS

The approach of using jitter and data sampling turned out to be a suitable solution to make scatter plots more comprehensible and perhaps more applicable to wider scenarios. However, the algorithmic approach is arbitrary, in a way, because it bears a large degree of freedom. Therefore, finding the optimal settings for a specific set of data and use cases still is difficult. A critical factor, for example, is the screen resolution. The larger the screen, the more data points can and should be displayed. This, in turn, strongly influences the setup of the optimal plotting algorithm.

In user interviews with teachers and school leaders, the scatter plot was chosen as the most appropriate chart type to visualize the achievements of groups of students without losing the information about individuals. Even for large students groups (e.g., the standardization sample), the scatter plot was preferred over more conventional methods such as density functions, typical bell-shaped curves, or pie charts. The individual data points could easily be associated with “real” students, which was not the case with the rather abstract area below a curve, as an example. The downside of the scatter plot, particularly when larger amounts of data (i.e., > 50 points) are displayed on a typical computer screen, is a rapid decrease of comprehensibility and legibility, mainly due to an overlap of data points. I presented two techniques to maintain the strength of the scatter plot and reducing the issue of over plotting. This approach to display student data was the favorite among a group of potential users.

To explore different characteristics of the plotting algorithm for different scenarios, in further steps, we will conduct simulation studies to compare different configurations of the algorithm in terms of legibility and ease of comprehension.

REFERENCES

- Anderson, L. (2013). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Cambridge, UK: Pearson Publishing.
- Sidiropoulos, N., Sohi S.H., Rapin N., & Bagger F.O. (2017). *An Enhanced Chart for Simple and Truthful Representation of Single Observations over Multiple Classes*. CRAN R package. Available online at <https://cran.r-project.org/web/packages/sinaplot/vignettes/SinaPlot.html>
- Sidiropoulos, N., Hadi Sohi, S., Pedersen, T.L., Porse, P.T., Winther, O., & Rapin, N. (2018). SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations Over Multiple Classes. *Journal of Computational and Graphical Statistics*, 27(3), 673-676.
- Siegler, R. S., DeLoache, J. S., Eisenberg, N., & Saffran, J. (2014). *How Children Develop*, 4th edition. New York: Worth.
- van der Linden, W. (Ed.) (2016). *Handbook of Item Response Theory (Volume 1)*, 1st Edition. London, UK: Chapman and Hall/CRC;

Less (context) is more? Evaluation of a positioning test feedback dashboard for aspiring students.

Nicolas Hoppenbrouwers

Faculty of Engineering Science, KU Leuven

Tom Broos

Dept. of Computer Science, KU Leuven

tom.broos@kuleuven.be

Tinne De Laet

Tutorial Services, Faculty of Engineering Science, KU Leuven

tinne.delaet@kuleuven.be

ABSTRACT: Aspiring engineering students profit from feedback regarding how their mathematical skills compare to the requirements and expectations of an engineering bachelor program. The positioning test is a non-binding test used in Flanders, Belgium assessing the mathematical skills of aspiring students. This paper elaborates on the research on and development of a learning analytics dashboard (LAD) that provides feedback on a participants' obtained results. Its objective is to provide actionable insights and to raise awareness and reflection about the participants' strengths and weaknesses, and subsequently their choice of study. To reach the final dashboard, the design went through six iterations, 662 students were surveyed, and 60 persons were thoroughly interviewed, including study advisors, students, and visualization experts. The final dashboard was evaluated using the EFLA, SUS, and a custom-made questionnaire, and a framework of factual, interpretative, and reflective insights. The results show that the developed dashboard is a considerable improvement over a comparable state-of-the-art dashboard. Furthermore, results show that a more visual representation, confined to only the most essential information, provides a better overview, leads to more and deeper insights while displaying less information and context, and has better usability and attractiveness scores than a more textual version.

Keywords: learning analytics, information visualization, student dashboard, positioning test, learning technologies

1 INTRODUCTION

The first bachelor year is often cited as the most essential to future academic success [1, 2, 11, 28]. A wide range of research focuses on identifying predictors of academic success in the first bachelor year, before students enroll in university programs, as this would shed light on the skills and knowledge students need to be successful. Apart from the obtained grade-point average in secondary education [3, 29], literature often describes mathematical ability as the most significant predictor of persistence and attainment in STEM fields [18, 20, 22, 23]. Starting mathematical competences is identified as one of the primary factors determining whether a student will continue studying in a STEM field, and certainly for engineering [21, 27]. Once the relevant skills are identified, learning analytics dashboards (LAD) can be developed to provide aspiring students with feedback, hereby supporting them in the

transition from secondary to higher education (HE). LADs are an effective and commonly used tool in learning analytics (LA) to visualize information [5, 7, 14, 15, 26]. Just like the general objective of information visualization, they can be used to represent large and complex quantities of data in a simple way [15, 19]. Few [16] defines a dashboard as ‘a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance’. Unlike most other countries, students in Flanders do not have to complete any formal application procedure or test in order to enroll in a university program. Furthermore, the tuition fee of EUR 922 per year is relatively low compared to other nations. Consequently, students are free in their choice of program, resulting in a large degree of heterogeneity in the first bachelor year regarding knowledge, skills, and educational background. This results in a drop-out of 40% in STEM fields. Since 2011, the Flemish universities offering engineering bachelor programs have joined efforts for organizing the ‘positioning test’, a non-obligatory and non-binding diagnostic test for the candidate students’ ability to solve math problems [31]. The focus on mathematics is not surprising considering the importance of mathematical ability as a predictor for student success in STEM [18, 20, 22, 23]. The positioning test typically contains 30 multiple choice questions and is organized in the summer between the end of secondary education and the start of higher education.

This paper presents the research that aimed at developing a LAD that provides aspiring engineering students with feedback on their mathematical problem-solving skills, based on their results on the positioning test. The developed LAD aims at visually triggering insights in the obtained results. More specifically, the LAD should provide actionable insights, making students more aware of their strengths and weaknesses, and allowing students to reflect on their study choice. The objective of the LAD is similar to that of the positioning test itself, in that it tries to encourage and motivate students that do well on the positioning test (score > 14/20) to consider engineering as a viable and interesting study option, participants who obtain a low score (score < 8/20) to reflect on their study choice, and support the middle group to take remedial actions (e.g. a summer course) to improve their mathematical abilities in order to successfully attain an engineering degree. To achieve these objectives, the research ran through all phases of a user-centered design process, including a preliminary data-analysis, a large survey of 622 end users, pilot testing, and 55 in-depth interviews. Different evaluation metrics were used to assess the developed dashboard: EFLA [24, 25], SUS [4], and a custom-made questionnaire, and the framework of factual, interpretative, and reflective insights [10]. Finally, this paper compares the developed dashboard with an existing feedback dashboard [6] for the positioning test.

2 RELATED WORK

The literature describes several guidelines for developing effective LADs. For example, Few [16] describes thirteen commonly made mistakes when developing dashboards. Together with the general graphical integrity and design aesthetic principles defined by Tufte and Graves-Morris [30], they serve as the basis for the development of the dashboard. The most commonly used visualization types in LADs are bar charts, line graphs, tables, pie chart, scatterplot, simple text, world clouds and traffic lights. De Laet [12] however warns not to use traffic lights, and mentions how wording is essential in LA applications. Predictive LA applications have uncertainty and it is important this uncertainty is also displayed [12]. LADs should avoid speaking too much in terms of “chances of failure” and “success”

[12]. Two additional relevant guidelines are defined by Charleer et al. [8]. They recommend that LADs should be sufficiently aggregated or abstract as an uncluttered representation incites more detailed explorations of the LA data. Secondly, they recommend that LADs should provide functions that increase the level of detail in the data [8].

The LAD of this paper focuses on the transition from one education system to the other (secondary to HE), while most examples in the literature are more concerned with monitoring study progress during an educational program, either for a student or a tutor. Several LADs were used as an inspiration for the LAD of this paper, such as the OLI dashboard [13], the Course Signals dashboard [1], the Student Activity Meter (SAM) [17], and the LISSA-dashboard [9]. The most related dashboard is that state-of-the-art dashboard by Broos et al. [6], which also aims at providing feedback after the positioning test. This LAD referred further on to as the “reference dashboard” provides, beside feedback on the mathematical problem-solving skills of students, feedback on learning and studying skills, and the prior education of students [6]. The reference dashboard by Broos et al. contains elaborate textual explanations and feedback to contextualize the participants’ obtained result.

LADs can incorporate insights of other research while visualizing data. Vanderroost et al. [31] analyzed the predictive power of the positioning test for engineering studies in Flanders. More specifically, the research examines whether it is possible to “predict” first-year academic achievement using the results of the positioning test. More specifically, the goal is to identify three distinct groups of students: group A are students who perform well in their first bachelor year, achieving a study efficiency of over 80% after the first exam period in January; group C are with a study efficiency below 30%; group B are students with a SE between 30 and 80 %. Earlier research [31] showed that participants obtaining a high score on the positioning test ($>13/20$) more often obtain good study results (study efficiency (SE) $>80\%$) in the first semester (51%), while students with a low score on the positioning test ($<8/20$) more often do not enroll (35%), drop-out (6%), or have low academic achievement (SE $<30\%$) in the first semester (39%). Vanderroost et al. also showed how the study efficiency in the first semester of the first bachelor year strongly predicts if a student will complete the engineering bachelor and in which time frame (in 3, 4 or 5 (or more) years).

3 CONTEXT

The positioning test consists of approximately thirty multiple-choice questions assessing participants’ problem-solving skills. Formula scoring is used to calculate the overall result (on 20) based on each participant’s responses. Each question is assigned to one of five mathematical categories: (1) reasoning, (2) knowledge of concepts, (3) spatial visualization ability, (4) skills (calculating derivatives, solving systems of linear equations, combinatorics, geometry, etc.) (5) and modeling questions (problem solving questions in a physical context that need combination and modeling of different inputs). Additionally, each question is assigned to one of four difficulty levels. The difficulty level of a question is determined by the percentage of participants that correctly answered the question: the 25% best answered questions of the 30 questions have a difficulty level of 1, while the 25% worst answered questions have a difficulty level of 4.

End-users of the existing reference LAD are participants of the positioning test, consisting mainly of students that just completed secondary education. They receive access to their personalized LAD through a feedback email, typically three days after completing the test. Apart from these aspiring

engineering students, other stakeholders are also involved. The Tutorial Services of the Faculty of Engineering Science heavily participates in the development of the LAD.

They are represented by the head of the unit and two study advisors (SAs), who from their pedagogical experience and educational expertise give feedback on content and design. SAs are concerned with guiding and helping students with any questions they might have. They can also be considered end-users of the dashboard, as they use the LAD to start the conversations with participants that need more feedback and advice during a private meeting. LA researchers and visualization specialists, represented by three experts of an HCI research group, evaluate the quality of the design.

4 DESIGN

Design process. A user-centered design process was followed to develop the dashboard. The design passed six iterations before reaching its final state. Throughout the iterations, the design principles by Tufte and Graves-Morris [30], the commonly defined dashboard mistakes by Few [16] and a set of self-defined design requirements served as guidelines for the development of the dashboard. The self-defined design requirements are formal specifications of the general objective described in Section 1 identified based on interviews with the involved stakeholders. They consist of eight functional requirements and six non-functional requirements. An example of a functional requirement is: ‘the ability to compare your own result with the results of other participants’. An example of a non-functional requirement is: ‘a good balance between textual and visual elements’.

In total the dashboard was developed and improved in six iterations. Each iteration is characterized by a different objective, format, and evaluation method. The first iterations focused more on functional requirements, finding out expectations, and determining the right content. Later iterations focused more on non-functional requirements and correctly choosing and improving the visualizations. The final design was programmed using D3.js. Different methodologies were used for creation and evaluation of the dashboard, such as co-designing, rapid prototyping, guidelines and general principles, the EFLA and SUS questionnaire, formal presentations with feedback, and semi-structured as well as informal interviews, based on distinct protocols, for instance scenario-based with concurrent

The content of the dashboard has changed throughout the six iterations. We conducted semi-structured interviews with two study advisors of the Bachelor of Engineering Science at the Catholic University of Leuven (KU Leuven), informal interviews with the head of Tutorial Services of the faculty, and a questionnaire among 662 students. In the questionnaire, students scored 14 different content part suggestions on a 7-point Likert scale for relevance and usefulness to include in a feedback system after participation in the positioning test. Results show that students like to see their total score, a comparison to other participants, and the typical performance (in terms of SE) of first-year bachelor students that obtained a similar score on the positioning test the previous year. They also liked to see the aggregated score per mathematical category and the score and original assignment per question. Students were divided when it comes to displaying the typical performance (in terms of grades on the course) on each individual course of first-year bachelor students who obtained a similar score on the positioning test previous year. They also disagreed regarding the presence of a specific, personalized, pre-defined study choice advice, due to insufficient face validity. Confirmed by the results of a data-analysis, which showed a lack of predictive power for these features, we decided to remove them

from the dashboard. The conclusions of the interviews with the study advisors (SAs) are similar to those of the survey. Examples of features that were added throughout the iterative process are the aggregated score per degree of difficulty and a picture of the original question of the positioning test, as both study advisors and students reacted positively to these suggestions.

Final design. The final design of the dashboard exists in two variants, differing in one part. Fig. 1 displays the first variant and consists of five major parts. Each part has a tooltip presenting more detailed information, following the general guidelines proposed by Charleer et al. [8], described in Section 2. Fig. 3 shows the tooltips for part A and B of Fig. 1. Furthermore, a help icon on the top right corner of each graph contains more context and explanation, e.g. explaining the color of a graph. The five major parts of the LAD (Fig. 1) and its tooltips allow students to:

- (A) review their obtained total score and compare themselves to the other participants by showing the general score distribution of all participants;
- (B) review each question, its difficulty, its mathematical category, its original assignment, the answer they submitted and the correct answer;
- (C) review their aggregated obtained score per mathematical category or degree of difficulty, allowing them to find their strengths and weaknesses and see whether they score well/bad on certain categories or easier/harder questions, permitting them to discover whether they lack basic knowledge or only more advanced skills;
- (D) compare themselves to other participants for each mathematical category and degree of difficulty, by plotting the score distribution per topic ;
- (E) view the typical first-year academic achievement, in terms of SE, of students in earlier cohorts based on their total positioning test score, via a Sankey diagram.

The second variant, in part displayed in Fig. 2, differs only on the strengths & weaknesses section (part C and D in Fig. 1). It combines the information of these two parts in on large histogram, displaying the distribution of the total positioning test score of all participants, and five small histograms, displaying the score distribution per category or degree of difficulty. The objective of the two separate variants is to see which visualization leads to more insights and whether the click functionality of the first variant is intuitive.



Figure 1: Feedback dashboard for future students after positioning test: first variant of final design. Corresponding to the displayed letters: (A) Donut chart showing the individually obtained total score on the position test (on 20). (B) Matrix showing the score and difficulty per question. Red indicates the student provided a wrong answer, grey for no answer and green for a correct answer. The percentage of correct responses by all test participants is indicated by the horizontal bar for each item. (C) Bar chart illustrating the participant's strengths and weaknesses, by showing the score per mathematical category (reasoning, knowledge of concepts, spatial visualization ability, skills, and modeling) and per degree of difficulty. (D) Histogram showing performance of peers for each mathematical category and degree of difficulty. The student's score is positioned using a vertical line. (E) Sankey diagram showing performance of previous students in the first bachelor year with a comparable total score on the positioning test. The diagram shows the outcomes in study efficiency (e.g. green arrows for students achieving 70% of study points, black arrows for students dropping out of the program) for three groups of positioning test outcomes (less than 10/20, from 10 to 13 and above 13).

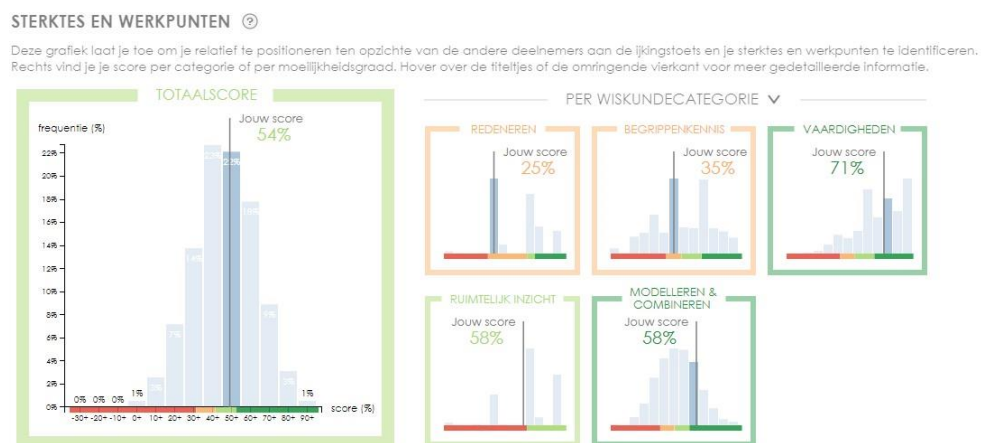


Figure 2: An alternative visualization for the score per category in the final dashboard, substituting part C and D of the dashboard.



Figure 3: Two examples of tooltips in the final dashboard. The tooltip on the left is shown when clicking or moving the mouse over part A in Figure 1. It shows more detailed information about the performance of the student on the positioning test and compares the result to that of other participants. The tooltip on the right clicking or moving the mouse over any of the items in part B in Figure 1. It shows detailed information for each question on the test, including the given and correct answer, the difficulty level, and question category. At the bottom of the tooltip, the question is represented in the same way as in the positioning test (on paper) to aid visual recall.

5 EVALUATION

Both variants, described in Section 4, are evaluated and compared to the reference dashboard [6], described in Section 2. The latter is currently in use in the Flemish universities organizing the positioning test. It is text-heavy in comparison to the evaluated alternatives, which allows to assess the added value of inclusion of such elaborate textual guidance.

Evaluation of the two final variants of the dashboard and reference dashboard [6] is based on 48 in-depth interviews (16 per dashboard), each lasting between 40 minutes and 1 hour. Each interview consists of four stages. The first phase of the interview is scenario-based, using the concurrent think-aloud protocol. End-users have to imagine having participated in the positioning test and now getting their result. Three scenarios are possible. Either they get a score in which they belong to group A (total score of 15/20), either group B (12/20) or group C (6/20). Anonymized data is used from the dataset described in Section 4. Each test user says out loud the insights they obtain upon visualization of the dashboard. The framework by Claes et al. [10] is used to measure these insights. The framework defines three levels of insights: 1) factual insights: simple, objective statements or questions that are triggered by the dashboard, e.g. “I obtained a total score of 14/20.”; 2) interpretative insights: interpretation of the displayed data, relying on the participant’s knowledge and experiences, e.g. “I mainly score well on the easier questions.”; 3) reflective insights: subjective, emotional and personal connotations triggered by the dashboard, leading to further awareness and reflection, e.g. “I feel like I did not do well enough at this test, making me doubt about whether I should go for another study program.”. Each insight is categorized into one of these levels. The test user can also mention when something in the dashboard is unclear, but the monitor of the test does not intervene and only writes down all statements made by the test person.

In the second phase, the interview switches to a task-based interview with active intervention. The monitor gives the test persons tasks based on the information or insights they missed during the first phase and finds out why these parts and insights have been missed. This phase tries to examine whether the dashboard is intuitive and has any shortcomings.

In the third phase, the test person fills in the SUS, the EFLA and a custom-made questionnaire, which verifies whether design requirements have been met. The EFLA questionnaire has been translated to Dutch and adapted to reflect the topic of the dashboard, identical to the evaluation of the dashboard of Broos et al. [6]. The design requirements questionnaire test consisted of 21 statements, to which the user could “Strongly disagree” or “Strongly agree”, using a 5-point Likert scale.

Finally, in the fourth phase the test persons get to see the two other dashboards and can express their preference. This last phase was optional.

6 RESULTS

Based on the recorded insights during the interviews 13 types of factual, 11 of reflective, and 8 types of interpretative insights were identified. All types of insights occurred more often with the participants for the LAD developed in this research compared to the reference dashboard (Table 1).

Table 1: Subset of the 13 types of factual (F), 11 types of reflective (R), and 8 types of interpretative (I) insights identified during the interviews and the percentages of interviewees in which these insights were found for the reference dashboard (B) of [6] and the two variants described in this paper.

Description	insight	%B	%V1	%V2
(F1)	My total score on the positioning test was ...	100	100	100
(F2)	My total score placed me in group A/B/C ...	100	94	94
(F3)	I answered X questions correct/wrong/blank	75	100	100
(F4)	My answer to this question was correct/wrong/blank	81	100	100
(F5)	On average this question was replied well/badly	56	88	94
(I1)	My total score compared wrt other participants	100	100	100
(I2)	This question was difficult/easy	56	88	81
(I5)	I score especially well in easy/difficult questions	56	56	63
(R1)	Reflection on total score	100	100	100
(R2)	Reflection on comparison wrt peers	69	100	94
(R3)	I guessed/left blank too many questions	44	56	63
(R4)	Reflection on particular question	56	88	81
(R10)	Reflection on future academic achievement	69	88	94
(R11)	Reflection on study choice	75	100	94

Fig. 4 shows the total SUS and EFLA score and the score per EFLA-dimension. The first variant has an overall average SUS-score of 81, the second variant 76, both statistically significant ($p < 0.01$) higher than the score of 47 of the reference dashboard. A score of more than 68 is considered above average [4], implying that the developed LAD has a better usability design than the reference dashboard. The differences between the averages of the two variants of the final dashboard are not statistically significant ($p > 0.2$). The total EFLA-score of the first variant is 74 and of the second variant is 70. Only

the EFLA score of the first variant is statistically significantly higher than the one of the reference dashboard score of 59.

4

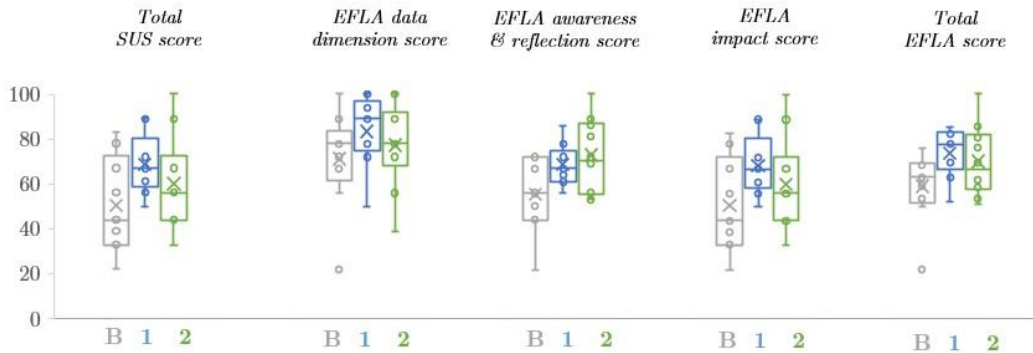


Figure 4: The total SUS and EFLA score and the score per EFLA-dimension: the data dimension (questions D1+D2), the awareness and reflection dimension (A1-A4) and the impact dimension (I1+I2). Gray boxplots ('B') denote the reference dashboard [6], blue box-plots ('1') denote the first variant of the final design of this paper and green ('2') the second variant.

The results of the design requirements questionnaire showed that each of the three dashboards successfully helps participants in understanding whether their current mathematical skills are matched with the expected mathematical skills and incites users of the LAD to awareness and reflection. Both variants, however, scored significantly better than the reference dashboard on the ability to use the dashboard independently, give a better overview of strengths and weaknesses, give a better detailed overview of the obtained result and allow participants to compare themselves more to the other participants. The users also indicated that these dashboards are better at displaying only factual, objective information, without giving interpretations or conclusions, but indicated that the dashboards can also be more confronting. Furthermore, they found that the two variants were more personalized, immediately gave an indication of the most important information, were better at showing only information that is relevant, were better at providing context, were more aesthetically pleasing, add less ambiguity and have a better balance between textual and visual elements, compared to the reference dashboard. For most design requirements, the differences between the two variants are not statistically significant.

7 DISCUSSION AND CONCLUSION

7.1 Implications for LAD design

This dashboard provides feedback to participants of the positioning test for the engineering program, inciting awareness and reflection about their strengths and weaknesses, and consequently their choice of study. The results of this LAD are interesting, as it focuses on the transition from secondary

school to higher education, while most LADs in the literature focus on monitoring students when they are already at university or college. Furthermore, a comparison has been made with the reference dashboard [6] that is currently used for feedback to the participants of the positioning test. The LADs developed in this research are more visual compared to the reference dashboard. Following thorough evaluation of the six iterations of the dashboard, the most important advantages of the more visual dashboards in this paper are that they have better usability scores, provide a better overview of the obtained results and a participant's strengths and weaknesses and visualize only relevant and objective information. A surprising result is that, while the visual dashboards contain less context and explanation, they still lead to more interpretative and reflective insights. Users declare that they think the layering of detail is better in the more visual dashboards. The main screen provides a good overview and immediately gives an indication of the essence, while the tooltips allow for more detailed information, consistent with the guidelines of Charleer et al. [8]. According to the tests, the reference dashboard of Broos et al.[6] has too much unnecessary information and text, which leads to users getting lost and not knowing what they should learn as take-away message. Some test persons also admit skipping parts of this dashboard because they "do not want to read so much text", causing them to miss out on important information.

The first most important general conclusion is that confining LADs to the most essential information, not displaying an overload of context and explanations, but using intuitive and simple visualizations, displaying less information, may lead to more awareness and reflections. An important part of LA applications is to make sure the end-users cannot get the incorrect interpretation, often leading to a lot of textual clarification. This research tries to convey to the designer that more text not necessarily means better insights, but well-designed and intuitive visualizations do.

Secondly, many test users mention how the dashboards of this paper are aesthetically pleasing and "fun to play with". Animations direct the user's attention to the most important information but are also specifically included to make the dashboard more aesthetically pleasing and show that the data is dynamic and interactive. While this result seems only of minor importance, it should not be underestimated. Several users mention how the aesthetics make them want to play more with the dashboard and spend more time with the dashboard. This eventually leads to more insights, which is essentially the final goal of this LAD. A lot of LADs do not spend enough time on the aesthetics of the dashboard, underestimating the effect this has on the effectiveness of the dashboard.

Finally, another objective was to see which of the two variants is more effective. The differences in the results are however not statistically different. Most users prefer the first variant, as it seems less cluttered at first sight, but end-users often miss some of the functionality in this variant. Further iterations should combine the best elements of both visualizations.

7.2 Future work and limitations of the study

The more visual dashboards however also have several disadvantages and pose new challenges. As all information is displayed on a single screen, some users observe the dashboard in an unstructured way, sometimes leading to less interpretative or reflective insights and confusion. Most participants observed the dashboard in a structured manner, but further research could examine whether a different arrangement of the various graphs could resolve this issue, keeping the visual character of the dashboard. Suggestions are a more sequential ordering of the graphs, similar to a grade report in

high school, or to use a guided tour to force the correct logical flow. Secondly, extra care is needed for the placement and highlighting of text. Because the visual dashboard looks more intuitive, users are less inclined to read any text at all, acknowledged by several test persons. While the graphs are mostly clear by themselves and lead to more interpretative and reflective insights, this is a real concern for the development of a dashboard. Further research should examine how to highlight text to force the user's attention to the surrounding text, even if they already understand the graph.

This study presents both qualitative and quantitative results of thorough four-stage evaluations with test users. It must be noted that the evaluation of the LADs happened with more experienced students asked to imagine being in the randomly assigned scenario of a student in transition from secondary to higher education. Test users completed the SUS, EFLA and custom questionnaires after an in-depth and a task-based interview (see Section 6). This may contribute to the explanation of inter-study differences between results reported previously [6] for the reference LAD (overall EFLA score of 72) and those reported in this paper (overall EFLA score of 59). In the former study, the actual target group of the reference LAD was surveyed using an on-screen questionnaire available within the dashboard itself. Further work is necessary to assess if, once accounted for methodological influence, outcome differences indicate that experienced students have different needs and preferences for LADs than newcomers.

REFERENCES

- [1] Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM.
- [2] Besterfield-Sacre, M., Atman, C. J., & Shuman, L. J. (1997). Characteristics of freshman engineering students: Models for determining student attrition in engineering. *Journal of Engineering Education*, 86(2), 139-149.
- [3] Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). Predictions of freshman grade-point average from the revised and recentered SAT® I: Reasoning Test. *ETS Research Report Series*, 2000(1), i-16.
- [4] Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
- [5] Broos, T., Peeters, L., Verbert, K., Van Soom, C., Langie, G., & De Laet, T. (2017, July). Dashboard for actionable feedback on learning skills: Scalability and usefulness. In *International Conference on Learning and Collaboration Technologies* (pp. 229-241). Springer, Cham.
- [6] Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2018, March). Multi-institutional positioning test feedback dashboard for aspiring students: lessons learnt from a case study in Flanders. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 51-55). ACM.
- [7] Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2017). Small data as a conversation starter for learning analytics: Exam results dashboard for first-year students in higher education. *Journal of Research in Innovative Teaching & Learning*, 10(2), 94-106.
- [8] Charleer, S., Klerkx, J., Duval, E., De Laet, T., & Verbert, K. (2016, September). Creating effective learning analytics dashboards: Lessons learnt. In *European Conference on Technology Enhanced Learning* (pp. 42-56). Springer, Cham.
- [9] Charleer, S., Moere, A. V., Klerkx, J., Verbert, K., & De Laet, T. (2018). Learning analytics dashboards to support adviser-student dialogue. *IEEE Transactions on Learning Technologies*, 11(3), 389-399.
- [10] Claes, S., Wouters, N., Slegers, K., & Vande Moere, A. (2015, April). Controlling in-the-wild evaluation studies of public displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 81-84). ACM.
- [11] Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution.

- [12] De Laet, T (2018). The (non)sense of “chances of success” and predictive models. <http://blog.associatie.kuleuven.be/tinnedelaet/the-nonsense-of-chances-of-success-and-predictive-models/>. Accessed 4 April 2018.
- [13] Dollár, A., & Steif, P. S. (2012). Web-based statics course with learning dashboard for instructors. *Proceedings of computers and advanced technology in education (CATE 2012), Napoli, Italy*.
- [14] Duval, E. (2011, February). Attention please!: learning analytics for visualization and recommendation. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 9-17). ACM.
- [15] Elias, T. (2011). Learning analytics. *Learning*, 1-22.
- [16] Few, S. (2006). Information dashboard design.
- [17] Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012, May). The student activity meter for awareness and self-reflection. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 869-884). ACM.
- [18] Green, A., & Sanderson, D. (2018). The roots of STEM achievement: An analysis of persistence and attainment in STEM majors. *The American Economist*, 63(1), 79-93.
- [19] Khalil, M., & Ebner, M. (2016). What is learning analytics about? A survey of different methods used in 2013-2015. *arXiv preprint arXiv:1606.02878*.
- [20] Kokkelenberg, E. C., & Sinha, E. (2010). Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students. *Economics of Education Review*, 29(6), 935-946.
- [21] Leuwerke, W. C., Robbins, S., Sawyer, R., & Hovland, M. (2004). Predicting engineering major status from mathematics achievement and interest congruence. *Journal of Career Assessment*, 12(2), 135-149.
- [22] Moses, L., Hall, C., Wuensch, K., De Urquidi, K., Kauffmann, P., Swart, W., ... & Dixon, G. (2011). Are math readiness and personality predictive of first-year retention in engineering?. *The Journal of psychology*, 145(3), 229-245.
- [23] Pinxten, M., Van Soom, C., Peeters, C., De Laet, T., & Langie, G. (2017). At-risk at the gate: prediction of study success of first-year science and engineering students in an open-admission university in Flanders—any incremental validity of study strategies?. *European Journal of Psychology of Education*, 1-22.
- [24] Scheffel, M (2018). Evaluation Framework for LA (EFLA). <http://www.laceproject.eu/evaluation-framework-for-la> , Accessed 1 March 2018.
- [25] Scheffel, M., Drachsler, H., & Specht, M. (2015, March). Developing an evaluation framework of quality indicators for learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 16-20). ACM.
- [26] Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., ... & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1), 30-41.
- [27] Tinto, V. (2005). *College student retention: Formula for student success*. Greenwood Publishing Group.
- [28] Solberg Nes, L., Evans, D. R., & Segerstrom, S. C. (2009). Optimism and College Retention: Mediation by Motivation, Performance, and Adjustment 1. *Journal of Applied Social Psychology*, 39(8), 1887-1912.
- [39] Stumpf, H., & Stanley, J. C. (2002). Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement*, 62(6), 1042-1052.
- [30] Edward Tufte and P Graves-Morris. 1983. The visual display of quantitative information. Vol. 2. Cheshire, CT Graphics press.
- [31] Vanderroost, J., Van Soom, C., Langie, G., Van den Bossche, J., Callens, R., Vandewalle, J., & De Laet, T. (2015, June). Engineering and science positioning tests in Flanders: powerful predictors for study success?. In *Proceedings of the 43rd Annual SEFI Conference* (pp. 1-8).

AutoTutor Tutorial: Conversational Intelligent Systems and Learning Analytics

Bor-Chen Kuo^{1*}, Chen-Huei Liao¹, Kai-Chih Pai¹, Chia-Hua Lin¹, Xiangen Hu^{2,3}, Zhiqiang Cai², Art Graesser².

¹National Taichung University of Education, Taiwan

²University of Memphis, USA

³Central China Normal University, China

* kbc@mail.ntcu.edu.tw

ABSTRACT: Conversational Intelligent tutoring system is a class of Adaptive Instructional Systems that are among the most studied and efficiently implemented in the last 20 years. This tutorial will introduce the most successful example C-ITS called AutoTutor and focuses on the authoring of AutoTutor lessons and Data analysis process of Tutoring data. Authoring of AutoTutor lessons include a) implementing discourse strategies in AutoTutor dialogues and trialogues, b) creating conversation elements (such as media elements); c) conversation rules, and d) using existing well-made authoring templates. Data analysis process of tutoring data include applying learning analytics methods, such as Bayesian Knowledge Tracing (BKT), Additive Factors Model (AFM), ...etc., to leverage the sequences of observations from student-ITS interaction log files to continually update the estimate of student latent knowledge.

Keywords: AutoTutor, Student Models, Learning Analytics

1 TUTORIAL BACKGROUND

Institute of Electrical and Electronics Engineers (IEEE) recently approved a standard committee ([P2247.1 - Standard for the Classification of Adaptive Instructional Systems](#)). This is a significant milestone for advanced personalized learning, which is identified by the National Academy of Engineering one of the grand challenges of the 21st century (<http://www.engineeringchallenges.org/9127.aspx>). Conversational Intelligent tutoring systems (C-ITS) is a class of AIS that are among the most studied and efficiently implemented in the last 20 years. This tutorial will bring you the most successful example C-ITS called AutoTutor (Graesser, Hu, & Person, 2001; Graesser et al., 2004; Nye, Graesser, & Hu, 2014; Nye, Graesser, Hu, & Cai, 2014; Person et al., 2000). AutoTutor holds conversations with the human in natural language. The authors of the proposed tutorial are among those who have development multiple versions of AutoTutor that teaches Critical Thinking (Wallace et al., 2009), Computer Literacy (Person, 2003), Physics (Graesser et al., 2003), Reading (Graesser et al., 2016), Electronics (Morgan et al., 2018), Chinese reading and mathematics learning (Liao, Kuo, & Pai, 2012).

AutoTutor applications are built with the guidance of human learning principles (A. C. Graesser, Halpern, & Hakel, 2008), such as Deep Questioning, to help students learn by holding deep reasoning conversations (Arthur C. Graesser & Person, 1994). AutoTutor converses with learners follow the Expectation-Misconception Tailored (EMT) dialog (Arthur C. Graesser et al., 2004). An AutoTutor conversation often starts with a main question about a certain topic. The goal of the conversation is to help students' construct an acceptable answer (expected answers) to the main question. Instead of telling the students the answers, AutoTutor asks a sequence of questions (hints, prompts) that target specific concepts involved in the ideal answer to the main question. AutoTutor systems respond to students' natural language input, as well as other interactions, such as making a choice, arranging some objects in the learning environment, etc.

This tutorial focuses on the authoring of AutoTutor lessons and Data analysis process of Tutoring data:

1. Authoring of AutoTutor lessons include a) implementing discourse strategies in AutoTutor dialogues and trialogues, b) creating conversation elements (such as media elements); c) conversation rules, and d) using existing well-made authoring templates.
2. Data analysis process of tutoring data include applying learning analytics methods, such as Bayesian Knowledge Tracing (BKT), Additive Factors Model (AFM), ...etc., to leverage the sequences of observations from student-ITS interaction log files to continually update the estimate of student latent knowledge.

2 ORGANIZATIONAL DETAILS

This event is a full-day tutorial. A Moodle website will be set up to continuously add more details and materials for participants. Tutorial will be announced on AutoTutor website (autotutor.org). Participants need to bring laptops. An example AutoTutor lesson will be provided to participants. Participants will create one's own AutoTutor lesson by modifying the example lesson. The proposed agenda is presented below in Table 1.

Table 1: Proposed agenda

Time	Session	Content
9:00-9:15	Introduction to AutoTutor	Introduction – Introduction of presenters and participants
9:15-10:30		Overview and Demo of AutoTutor Systems
10:30-11:00	Coffee Break	
11:00-12:30	AutoTutor Script Authoring Tool	A step by step guidance to creating an AutoTutor lesson
12:30-14:00	Lunch Break	
14:00-15:30	Student Models and Learning Analytics	Student Models in AutoTutor
16:00-17:30	Learning Analytics for AutoTutor	AutoTutor log data analysis by using Bayesian Knowledge Tracing (BKT) and Additive Factors Model (AFM)

3 TUTORIAL OBJECTIVES OR INTENDED OUTCOMES

The objectives of the tutorial include but not limited to 1) Understand the theoretical foundations, enabling technologies, and practical applications of C-ITS through hands-on and worked-out examples of AutoTutor. 2) Familiar with simple, common, and advanced data analysis methods that apply to the analysis of AutoTutor Data.

The intended outcomes of the tutorial include 1) All attendees will be able to create a complete C-ITS module. 2) All attendees will understand the data structure of the interaction between C-ITS and learners, 3) All attendees will be able to analyze data using the data analytical methods introduced.

REFERENCES

- Graesser, A. C., Cai, Z., Baer, W. O., Olney, A. M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the Center for the Study of Adult Literacy. *Adaptive Educational Technologies for Literacy Instruction*, 288–293.
- Graesser, A. C., Halpern, D. F., & Hakel, M. (2008). *25 principles of learning*. Task Force on Lifelong Learning at Work and at Home Washington, DC.
- Graesser, A. C., Hu, X., & Person, N. K. (2001). Teaching with the help of talking heads. *Proceedings of the 2001 IEEE International Conference on Advanced Learning Technologies*, 460-461.
- Graesser, A. C., Jackson, G. T., Matthews, E. C., Mitchell, H. H., Olney, A., Ventura, M., et al. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1-5). Boston: Cognitive Science Society.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36, 180-193.
- Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31, 104–137.
- Liao, C.-H., Kuo, B.-C., & Pai, K.-C. (2012). Effectiveness of Automated Chinese Sentence Scoring with Latent Semantic Analysis. *Turkish Online Journal of Educational Technology-TOJET*, 11(2), 80–87.
- Morgan, B., Hampton, A. J., Cai, Z., Tackett, A., Wang, L., Hu, X., & Graesser, A. C. (2018). Electronixtutor Integrates Multiple Learning Resources to Teach Electronics on the Web. *In Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 33:1–33:2). New York, NY, USA: ACM.
- Nye, B.D., Graesser, A.C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427-469.
- Nye, BD, Graesser, AC, Hu, X. (2014b). AutoTutor in the cloud: a service-oriented paradigm for an interoperable natural-language ITS. *Journal of Advanced Distributed Learning Technology*, 2(6), 35–48.
- Person, N. K. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head. *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies*, 97, 47.

- Person, N. K., Craig, S., Price, P., Hu, X., Gholson, B., Graesser, A. C., & Tutoring Research Group. (2000). Incorporating human-like conversational behaviors in AutoTutor. *Proceedings of the Agents 2000 Conference*, 85-92.
- Wallace, P., Graesser, A. C., Millis, K., Halpern, D., Cai, Z., Britt, M. A., Magliano, J., & Wiemer, K. (2009). Operation ARIES!: A computerized game for teaching scientific inquiry. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. C. Graesser (Eds.), *Proceedings of 14th International Conference on Artificial Intelligence in Education. Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 602-604). Amsterdam: IOS Press.

2nd Personalising feedback at scale Workshop: Focusing on Approaches and Students

Lorenzo Vigentini

UNSW Sydney

l.vigentini@unsw.edu.au

Danny Y.T. Liu

University of Sydney

danny.liu@sydney.edu.au

Lisa Lim

University of South Australia

Lisa.Lim@unisa.edu.au

ABSTRACT: After a successful workshop at LAK'18, in which presenters explored tools used to provide feedback at scale, this workshop shifts the attention to data-driven approaches to support the provision of feedback and the students, especially considering how they perceive the feedback and what they do with the feedback received. The workshop aims to bring together scholars and practitioners to find a common ground for showcasing interesting examples of effective feedback and explore what and how data can be used to improve the process and richness of feedback for both learners and educators. Key outcomes will be a better understanding of approaches and existing cases of good practice which will foster discussion and collaboration in the LA community.

Keywords: personalization, effective feedback, student-centered analytics

1 INTRODUCTION

1.1 Background

The provision of effective and timely feedback of and for learning (Brown & Knight, 1994; Hattie, 2008; Hattie & Timperley, 2007; Hounsell, 2003; Sadler, 1989) has been shown to be essential in influencing students' achievement and promoting autonomy and self-regulation (Black, Harrison, & Lee, 2003; Nicol, 2010; Sadler, 2010). Interestingly, feedback is often the lowest rated aspect in terms of satisfaction from graduate satisfaction and/or engagement surveys (Krause, Hartley, James, & McInnis, 2005; McDowell, Smailes, Sambell, Sambell, & Wakelin, 2008; Radloff, Coates, James, & Krause, 2011; Rowe & Wood, 2009; Williams & Kane, 2008). More importantly, while assessment practices have received considerable attention over the past two decades -examples such as REAP, SAFE, 'Transforming Assessment' projects - (Carless, Salter, Yang, & Lam, 2011; Crisp, 2011; Nicol, 2009)-, the focus on feedback to students has remained relatively scarce (Higgins, Hartley, & Skelton, 2002; Orsmond, Maw, Park, Gomez, & Crook, 2013; Rowe & Wood, 2009).

The provision of feedback at scale (Liu, Bartimote-Aufflick, Pardo, & Bridgeman, 2017; Pardo, Jovanovic, Dawson, Gašević, & Mirriahi, n.d.; Vigentini, Liu, Lim, & Martinez-Maldonado, 2018) and the *personalisation* of feedback (using Learning Analytics) has become a sort of holy grail for educators aspiring to improve their students' learning and their satisfaction with the learning experience (Bienkowski, Feng, & Means, 2012; King, Kinash, Kordyban, & Pamentier, 2014). However, students and educators do not hold the same perception of what constitutes quality feedback (Carless, 2006; Forsythe & Johnson, 2017; Hounsell, McCune, Hounsell, & Litjens, 2008; Lizzio & Wilson, 2008; Pitt & Norton, 2017). In most cases, the idea of providing feedback is reduced

to a summative, corrective and transmissive process, which gives a final judgement on students' submitted assignments (Nicol, 2010; Weaver, 2006).

In order to improve the process, some researchers (Forsythe & Johnson, 2017; Pitt & Norton, 2017) have started to reconsider the impact of (or lack of) feedback as currently implemented in Higher Education and, instead, to focus more on the constructive value of a dialogic approach in which both giving and receiving feedback are considered more holistically (Forsythe & Johnson, 2017; Nicol, 2010; Pitt & Norton, 2017; Poulos & Mahony, 2008). This is more akin to the model of continuous feedback which students are used to in schools (Hattie, 2008). Although LA have made tangible connections with critical aspects that can strongly shape learning, such as learning design and self-regulation, the provision of feedback to students has been relatively neglected (Liu et al., 2017; Pardo, 2017). This is despite the affordances of LA to leverage the generation of theoretical and technical mechanisms for understanding and improving learning by "informing and empowering instructors and learners" (Siemens & Baker, 2012). To allow this to happen, teachers need concrete approaches and support mechanisms to bridge the gap between LA research and classroom practice. Newer LA systems are starting to support teachers with means to provide rich feedback beyond typical early warning messages (e.g. SRES or Ontask - Liu et al., 2017; Pardo, 2017; Pardo et al., n.d.; Tempelaar, Rienties, & Giesbers, 2015), but it is clear that there is a need and appetite in the LA community of research and practice to further explore data-informed student-centred pedagogies to provide feedback at scale.

While the first workshop with this topic at LAK'18 focused predominantly on tools and their applications, this second workshop shifts the attention to data-driven approaches supporting the provision of feedback, and encourage submissions to pay more attention to the students, especially considering how they perceive the feedback and what they do with the feedback.

1.2. Scope of the workshop

This workshop brings together scholars and practitioners to explore interesting examples of effective feedback and explore what and how data can be used to improve the process and richness of feedback for both learners and educators. The workshop has three primary goals:

- Provide a multidisciplinary theoretical foundation for practitioners and researchers in LA for the effective provision of data-informed feedback practices in HE;
- Showcase extant or planned approaches that provide feedback to students and consider students' reception of the feedback, with a focus on approaches that are data-driven and personalised;
- Promote reflection on both pedagogical and technological approaches to improve feedback practices targeted at the improvement of student learning and their ability to self-regulate learning.

2. ORGANISATION DETAILS

This half-day workshop is targeted to those who wish to understand and apply principles of feedback of and for learning. Given the explicit multidisciplinary nature of the workshop we expect that it will provide an opportunity to discuss and share innovations, impact on learning, and explore future

directions in the application of learning analytics (LA) to personalisation of feedback. Likely interested participants are:

- Educators/teachers and researchers
- Technologists and educational developers
- Learning scientists and data scientists/analysts
- Academic managers
- and anyone else interested in personalisation of learning and teaching

The organisers welcomed two types of contributions:

- short 1-3 page papers OR extended abstract OR poster to showcase work in progress for successful approaches for the provision of feedback at scale, focusing on students' reception of this feedback
- Short issue papers (max 5 pages) provoking the audience to think about key issues and problems related to scaling feedback and the use of analytics to support the process

The key focus of interest of the workshop include, but was not limited to:

- Overview of tool(s)/approach(es) to personalise feedback
- Implementation process (e.g. infrastructural, staff capacity, etc.)
- Challenges and successes (as well as failures)
- Stakeholder engagement, buy-in, and impact (especially faculty, students)

2.1. Proposed workshop activities

After a brief introduction and conceptualisation of the workshop, three short presentations will showcase two empirical cases and a conceptual piece to give participants a backdrop and provocation to reflect on ways in which we normally provide feedback: this will consider both the typically sparse provision in Higher Education as well as the continuous provision typical of schools. We will look specifically at successful approaches, what they have in common and, most importantly, consider how students receive the feedback and do something with it (i.e. the 'closing the loop').

The first paper by Tomer Gal and Arnon HersHKovitz provides a frank assessment of the challenges which researchers face when studying the effects of response-based feedback at scale. In their study, they present work done with the Khan Academy platform and discuss five major challenges related to the population of the study and its variation, representativeness of the sample, incomplete information/data in the sample, platform/product changes and the effects that UI changes have on the study.

The second paper by Lisa Lim, Hamideh Iraj, Abelardo Pardo and Shane Dawson presents preliminary results from the use of an approach to data-driven feedback in First year courses. The analysis of focus group discussions with students from two different courses, where this data-driven feedback approach was piloted using the OnTask tool, indicates that students acted on the feedback and that the type and level of feedback provided has practical implications for both the affect and self-regulated learning dimensions. This draws the attention to the function of feedback and the ways in which feedback is provided in practice.

The third paper by Lorenzo Vigentini provides a conceptual scaffold for the technical development and implementation of tools supporting feedback at scale. The focus is on the process and the building blocks of the supporting software within the theoretical framework of the provision of feedback at scale. The paper proposes a taxonomy to evaluate the design choices required to turn the model into a working software tool which should be helpful.

After the presentations, in the second half of the workshop, breakout groups will be guided with a semi-structured approach to discuss key themes and issues surfaced during presentations (e.g. use of data-informed feedback by students, types of feedback made possible through data, challenges of faculty professional learning, data sources needed for personalisation, etc.).

A website (<https://sites.google.com/view/lak19workshop/>) has been created to provide access to all contributions and presentations as well as a summary from the organisers after the workshop. The workshop will provide an avenue to continue the conversations beyond the session and open opportunities for further collaborations.

3. INTENDED OUTCOMES FOR PARTICIPANTS

We expect a range of presentations that will cover practical, evidence-based approaches to personalising data-driven feedback at scale. Participants will be able to:

- Obtain a broad perspective of different approaches to using data for personalising feedback
- Enhance their understanding of the forms of feedback that could improve student learning
- Gain an appreciation of the range of contexts where feedback can be valuable, and how data can inform these
- Discuss cases, issues, and potential solutions to implementing LA-enhanced feedback practices
- Connect with researchers and practitioners working to provide personalised feedback, yielding opportunities for collaborating on approaches and tools across attending institutions.

After the workshop, given the commitment to further collaborations, contributors will be invited to consider more substantial submissions with the intention to collate the works into a special issue of journal or an edited book on the topic.

4. REFERENCES

- Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, 1, 1–57.
- Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill Education (UK).
- Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. Psychology Press.
- Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, 31(2), 219–233. <https://doi.org/10.1080/03075070600572132>
- Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, 36(4), 395–407. <https://doi.org/10.1080/03075071003642449>
- Crisp, G. (2011). *Rethinking assessment in the participatory digital world – Assessment 2.0* (National Teaching Fellowship Fil Report).
- Forsythe, A., & Johnson, S. (2017). Thanks, but no-thanks for the feedback. *Assessment & Evaluation in Higher Education*, 42(6), 850–859. <https://doi.org/10.1080/02602938.2016.1202190>
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Higgins, R., Hartley, P., & Skelton, A. (2002). The Conscientious Consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, 27(1), 53–64. <https://doi.org/10.1080/03075070120099368>
- Hounsell, D. (2003). Student feedback, learning and development. *Higher Education and the Lifecourse*, 67–78.
- Hounsell, D., McCune, V., Hounsell, J., & Litjens, J. (2008). The quality of guidance and feedback to students. *Higher Education Research & Development*, 27(1), 55–67. <https://doi.org/10.1080/07294360701658765>
- King, C., Kinash, S., Kordyban, R., & Pamenter, J. (2014). Personalising student learning through education. Bond University. Retrieved from <http://epublications.bond.edu.au/tls/87>
- Krause, K. L., Hartley, R., James, R., & McInnis, C. (2005). *The first year experience in Australian universities: Findings from a decade of national studies*. Centre for the Study of Higher Education, University of Melbourne Melbourne. Retrieved from <http://www.cshe.unimelb.edu.au>
- Liu, D. Y.-T., Bartimote-Aufflick, K., Pardo, A., & Bridgeman, A. J. (2017). Data-Driven Personalization of Student Learning Support in Higher Education. In *Learning Analytics: Fundaments, Applications, and Trends* (pp. 143–169). Springer, Cham. https://doi.org/10.1007/978-3-319-52977-6_5
- Lizzio, A., & Wilson, K. (2008). Feedback on assessment: students' perceptions of quality and effectiveness. *Assessment & Evaluation in Higher Education*, 33(3), 263–275. <https://doi.org/10.1080/02602930701292548>
- McDowell, L., Smailes, J., Sambell, K., Sambell, A., & Wakelin, D. (2008). Evaluating assessment strategies through collaborative evidence-based practice: can one tool fit all? *Innovations in Education and Teaching International*, 45(2), 143–153. <https://doi.org/10.1080/14703290801950310>
- Nicol, D. (2009). Assessment for learner self-regulation: enhancing achievement in the first year using learning technologies. *Assessment & Evaluation in Higher Education*, 34(3), 335–352. <https://doi.org/10.1080/02602930802255139>
- Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5), 501–517. <https://doi.org/10.1080/02602931003786559>
- Orsmond, P., Maw, S. J., Park, J. R., Gomez, S., & Crook, A. C. (2013). Moving feedback forward: theory to practice. *Assessment & Evaluation in Higher Education*, 38(2), 240–252. <https://doi.org/10.1080/02602938.2011.625472>
- Pardo, A. (2017). A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education*, 0(0), 1–11. <https://doi.org/10.1080/02602938.2017.1356905>
- Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (n.d.). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, 0(0). <https://doi.org/10.1111/bjet.12592>
- Pitt, E., & Norton, L. (2017). 'Now that's the feedback I want!' Students' reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education*, 42(4), 499–516. <https://doi.org/10.1080/02602938.2016.1142500>
- Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education*, 33(2), 143–154. <https://doi.org/10.1080/02602930601127869>
- Radloff, A., Coates, H., James, R., & Krause, K.-L. (2011). Report on the Development of the University Experience Survey. *Higher Education Research*. Retrieved from http://works.bepress.com/hamish_coates/79
- Rowe, A. D., & Wood, L. N. (2009). Student Perceptions and Preferences for Feedback. *Asian Social Science*, 4(3), 78. <https://doi.org/10.5539/ass.v4n3p78>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>
- Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550. <https://doi.org/10.1080/02602930903541015>
- Siemens, G., & Baker, R. S. J. d. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). New York, NY, USA: ACM. <https://doi.org/10.1145/2330601.2330661>
- Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47(Supplement C), 157–167. <https://doi.org/10.1016/j.chb.2014.05.038>
- Vigentini, L., Liu, D. Y.-T., Lim, L., & Martinez-Maldonado, R. (2018). Personalising feedback at scale: approaches and practicalities. In *Companion Proceedings of the 8th International Conference on Learning Analytics & Knowledge (LAK'18)*.
- Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, 31(3), 379–394. <https://doi.org/10.1080/02602930500353061>
- Williams, J., & Kane, D. (2008). Exploring the National Student Survey Assessment and feedback issues. *The Higher Education Academy, Centre for Research into Quality*.

Challenges in Studying Feedback in Massive Online Platforms

Tomer Gal, Arnon HersHKovitz

School of Education, Tel Aviv University
{tomergal@mail, arnonhe@tauex}.tau.ac.il

ABSTRACT: The rise in popularity of massive online learning platforms bares the promise of thriving research that builds on large datasets drawn from such platforms. While using such a platform (Khan Academy) to study the effects of elaborated response-based feedback on learning, we have faced a few challenges that are of interest to researchers in this field. Shedding light on these challenges may assist in improving future explorations of data-driven feedback-related explorations. In this paper, we discuss five such challenges: population varies greatly along time; inclusion criteria may harm representativity; incomplete population information; changes in product may interfere with experiment; and the unknown impact of interface. We summarize with recommendations for researchers in the field.

Keywords: feedback, elaborated feedback, massive online platforms, math education

1 INTRODUCTION

Collecting large data drawn from experiments that run on open platforms—like Khan Academy, Coursera, edX, or Scratch—has become a common practice in the Learning Analytics and other communities. These experiments benefit from the ease of recruiting participants, from addressing varied populations in the context of their authentic learning experience, and from being able to run for a rather long time. Some examples for this approach are Machardy and Pardos' study of video use by analyzing two-year data from Khan Academy (2015), or Huang et al.'s study of discussion behavior by analyzing two-year Coursera forums data (2014).

However, there are also some drawbacks to using such large data sets. These Internet-scale experiments are often characterized by anonymity of participants, high attrition, data overload, and threats to internal validity (Stamper et al., 2012). Of course, different concepts may be vulnerable differently to these (and other) disadvantages. Specifically, when studying feedback, students' demographics and other personal characteristics are important factors (Rice & Bunz, 2006; Tangmanee & Nontasil, 2014), as well as interface considerations that may harm or obscure validity (Howie, Sy, Ford, & Vicente, 2000).

In this paper, we draw on our experience in studying feedback in mathematics via the analysis of large data sets from an open learning environment (Khan Academy). We point out five important challenges, demonstrate them in the context of our studies, and report on how we overcame them.

2 THE CONTEXT OF OUR STUDY

Existing research has shown that in computer-based mathematics instruction, elaborated response-based feedback is often more effective than simply noting whether the given answer is correct or

incorrect, or from providing the learners with the correct answer if they were wrong (Van der Kleij, Feskens, & Eggen, 2015). However, there have been only a few attempts to compare between different types of elaborated feedback, and little is known about how to design an effective elaborated feedback. In our study, we aim at bridging this gap. As a first step, we tested for differences in the effectiveness of textual vs. symbolic feedbacks, demonstrating an overall superiority of the latter.

We took a data-driven approach, using randomized experiments in Khan Academy (<https://www.khanacademy.org>), one of the most popular online platforms for learning mathematics. We analyzed data of between 3,023-33,378 learners in each experiment. Findings are reported in the main Conference (Gal & Hershkovitz, 2019). The experiments ran on four exercises (i.e., problem sets) in high-school-level mathematics. Pairs of elaborated feedback messages were written by the authors for the four exercises (that is, to each problem within each exercise). Figure 1 gives an example for textual feedback and symbolic feedback written for the same problem in the exercise "Slope from two points."

The experiment period consisted of two data collection phases of 4 weeks each. During those phases, learners who entered the experiment-exercises were randomly assigned to either experiment or control conditions. Learners in the control group got a simple indicative feedback (right/wrong), while those in the experiment group got either textual or symbolic feedback, depending on the phase. After the duration of the experiment, we used log data to compare feedback effect on success in same and subsequent problem after being exposed to feedback.

<p>What is the slope of the line through $(-1, -7)$ and $(3, 9)$?</p> <p>Choose 1 answer:</p> <p><input type="radio"/> (A) 4</p> <p><input type="radio"/> (B) $-\frac{1}{4}$</p> <p><input type="radio"/> (C) -4</p> <p><input checked="" type="radio"/> INCORRECT $\frac{1}{4}$</p> <p>Make sure the difference in y-values is in the numerator and the difference in x-values is in the denominator.</p>	<p>What is the slope of the line through $(-1, -7)$ and $(3, 9)$?</p> <p>Choose 1 answer:</p> <p><input type="radio"/> (A) $-\frac{1}{4}$</p> <p><input checked="" type="radio"/> INCORRECT $\frac{1}{4}$</p> <p>$\frac{1}{4}$ is $\frac{\text{Change in } x}{\text{Change in } y}$.</p> <p>But we need $\frac{\text{Change in } y}{\text{Change in } x}$.</p> <p><input type="radio"/> (C) 4</p> <p><input type="radio"/> (D) -4</p>
--	--

Figure 1: Examples of textual feedback (left) and symbolic feedback (right)

3 THE CHALLENGES

3.1 Population Varies Greatly Along Time

Big Data's power lies in the number of participating students. For studies that are based on massive online platforms, the number of participants is often a function of time—the longer an experiment runs, the more participants you get. This, however, has a price, as the research population's characteristics may dramatically change over time. As learners' characteristics and their level of knowledge heavily impact their acceptance of feedback (Winstone, Nash, Parker, & Rowntree, 2017), this is a great challenge in studying feedback.

In the context of our study. Our study was conducted in two phases over separate periods. In each phase, the control group received simple feedback (correct/incorrect) while the treatment group received a response-based, elaborated feedback (either textual or symbolic). We were interested in comparing the two treatment groups. However, we discovered that the research populations in the two phases were significantly different in their mathematics level (hence, maybe in other traits as well). In particular, the overall success rates in the first problem of each problem set (which is what the experiments focused on) oscillates along the year, from below 50% to over 70% (see Figure 2). This is probably due to differences in the populations who use the website (e.g., college students during academic semesters and high-school students during school year).

Our solution. In order to overcome this challenge, we decided to normalize our success measures by the average success rate of our population, in each phase separately. We believe that this allows us a better comparison between the research variables.

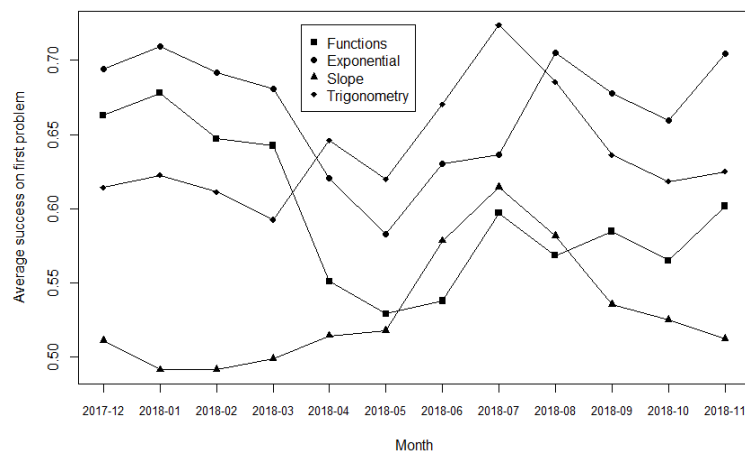


Figure 2: Average success rates on first problem (in four exercises), per month

3.2 Inclusion Criteria May Harm Representativity

Often, in order to be included in a feedback study, learners' behavior must agree with some inclusion criteria. In designed experiments of feedback, one should make sure the participants were exposed to the feedback message, which often—like in the case of feedback for incorrect responses—requires eliminating students based on their responses. For this reason, even though the total number of the platform users may be large, the selected population may be considerably smaller, in a way that may harm representativity (McMahon, 2002).

In the context of our study. To be included in our experiments, a learner had to fulfill a set of conditions: answering first problem using no hints, providing incorrect solution on first attempt in first problem, having at least one more attempt on the first problem, and attempting to answer the second problem. As a result, although we were able to reach final populations of a few thousands users, they were still considerably smaller than the original full populations (see Table 1). This may harm representativity (as we do not assume similar characteristics between the large and the reduced populations), hence may hurt the ability to present generalizable conclusions. Specifically, as the inclusion criteria require that the learner answered the first problem incorrectly and didn't use hints, there may have been an inclusion bias regarding the level of confidence and/or help-seeking behaviors of the final research population.

Our solution. As we had to test for feedback impact, and needed to do it at the very beginning of the topic's learning, we could not compromise our selection criteria. At the very least, we report on the pre- and post-selection population sizes.

Table 1: Initial and final number of students in each experiment

Exercise	Feedback Type	Initial N	Final N
Even and odd functions	Textual	3023	1103
Even and odd functions	Symbolic	2332	744
Graphs of exponential functions	Textual	3684	1063
Graphs of exponential functions	Symbolic	7369	2482
Slope from two points	Textual	22252	7549
Slope from two points	Symbolic	11480	3515
Trigonometric ratios in right triangles	Textual	33378	7605
Trigonometric ratios in right triangles	Symbolic	22490	4542

3.3 Incomplete Population Information

According to the common approach in feedback research, student characteristics have great implications on feedback effect (Shute, 2008); both demographics (Oliver, 2000; Terzis & Economides, 2011; Turner & Gibbs, 2010) and prior knowledge (Fyfe, Rittle-Johnson, & DeCaro, 2012; Smits, Boon, Sluijsmans, & van Gog, 2008) impact how feedback is being perceived and utilized, and how effective it is. In that sense, massive online environments suffer from two main disadvantages. First, they usually do not require students to provide much personal information, if at all, which makes background information inaccessible to researchers (and even if this information is required, one cannot trust its validity.) Second, learning may happen in other platforms in parallel to the use of a given online environment, which makes inferring prior knowledge based solely on the system logs a very difficult task.

In the context of our study. Khan Academy does not require users to provide their age or gender. As a result, we had very incomplete information about our research population. For example, only about 20% of the population had provided their age, and 10% or less had provided their gender; only about 5% or less had provided both (see Table 2). Additionally, we could not rely on log-based measures of prior knowledge.

Our solution. Sadly, we had to give up including age, gender, and prior knowledge in our analysis. In the future, we will seek for means of obtaining this information in a reliable fashion.

Table 2: Number of students sharing their age and gender per experiment

Exercise	Feedback Type	Total students	Provided age	Provided gender	Provided both
Even and odd functions	Textual	1103	233	134	84
Even and odd functions	Symbolic	744	157	84	50
Graphs of exponential functions	Textual	1063	226	126	67
Graphs of exponential functions	Symbolic	2482	529	233	130
Slope from two points	Textual	7549	1377	468	305
Slope from two points	Symbolic	3515	779	308	204
Trigonometric ratios in right triangles	Textual	7605	1475	656	413
Trigonometric ratios in right triangles	Symbolic	4542	795	386	229

3.4 Changes in Product May Interfere with Experiment

EdTech companies strive to survive in an ever changing, highly competitive world. Therefore, their products are constantly changing, in hope to improve the way they promote learning. More often than not, a researcher does not have control on those product changes. As the perception and acceptance of feedback is a complex process (Winstone et al., 2017), this concept is vulnerable to such changes. If such changes happen while an experiment is running, they may have great impact on it, even to a point of rendering the entire study unreliable.

In the context of our study. While our experiments ran on Khan Academy, the company was making adjustments in its mastery system—that is, the mechanics by which learners are recognized for mastery of the skills taught. As a result, the distribution of mastery levels reached by students varied greatly over time, in a way that cannot be explained by changes in population characteristics alone. As we intended to measure the cumulative effect of the type of feedback on knowledge demonstration at the level of a problem set, and as we thought of relying on the system's mastery score for that purpose, this was a major obstacle. Fortunately, these changes were unlikely to affect student behavior during the first couple of problems (which were the focus of our experiments).

Our solution. As a temporary solution, we gave up measuring this cumulative effect for those experiments, and designed a new data collection, after making sure with the company that the mastery scoring mechanism is untouched.

3.5 The Unknown Impact of Interface

Mostly, feedback studies tend to focus on feedback's content or timing. However, in the context of digital learning environments, the user interface may be just as important. One can write the most impeccable feedback message and make sure it appears in perfect timing, but if the user interface does not make it easy to notice or use, the whole endeavor is in vain. Indeed, this issue has been previously acknowledged (Howie et al., 2000). Unfortunately, for many online platforms, researchers

do not have the mandate of changing the feedback interface, which poses a serious restriction on a study's validity.

In the context of our study. Once a learner picks an incorrect answer in a multiple-choice problem in Khan Academy, the feedback message immediately appears below the answer choice, with no separation between the learner's original choice and the message. The feedback message appears in a grey font, which is not very prominent (see Figure 3, left); this design might prevent students from noticing the feedback (and therefore their inclination to use it). Once a learner picks the wrong answer and then picks the right answer, feedback messages for all solutions—correct and incorrect—appear at once, each immediately below the relevant choice (see Figure 3, right); this might have an alienating effect that disengages the student from interacting with the feedback or the website entirely.

Our solution. As we could not alter this mechanism of feedback presentation, we recognize its potential effect on learners.

Figure 3 consists of two side-by-side screenshots of a Khan Academy problem interface. The problem asks: "What is the slope of the line through $(-4, 2)$ and $(3, -3)$?" and instructs the user to "Choose 1 answer:".

The left screenshot shows the user has selected the incorrect answer $-\frac{7}{5}$. The feedback message, displayed in a grey font, states: "INCORRECT (SELECTED)", " $-\frac{7}{5}$ is $\frac{\text{Change in } x}{\text{Change in } y}$. But we need $\frac{\text{Change in } y}{\text{Change in } x}$." Below this, the other three options are listed: (B) $-\frac{5}{7}$, (C) $\frac{7}{5}$, and (D) $\frac{5}{7}$.

The right screenshot shows the user has selected the correct answer $-\frac{5}{7}$. The feedback message for the correct answer states: "CORRECT (SELECTED)", " $-\frac{5}{7}$ ". Below this, the feedback messages for the other three incorrect options are also displayed. For option (A) $-\frac{7}{5}$, it says: "INCORRECT", " $-\frac{7}{5}$ is $\frac{\text{Change in } x}{\text{Change in } y}$. But we need $\frac{\text{Change in } y}{\text{Change in } x}$. Furthermore, $\frac{7}{5}$ is $\frac{3 - (-4)}{2 - (-3)}$. But we need $\frac{(-3) - 2}{3 - (-4)}$." For option (B) $-\frac{5}{7}$, it says: "INCORRECT", " $-\frac{5}{7}$ is $\frac{(-3) - 2}{(-4) - 3}$. But we need $\frac{(-3) - 2}{3 - (-4)}$." For option (D) $\frac{5}{7}$, it says: "INCORRECT", " $\frac{5}{7}$ is $\frac{(-3) - 2}{(-4) - 3}$. But we need $\frac{(-3) - 2}{3 - (-4)}$."

Figure 3: Example of a feedback message for a wrong answer (left) and for a right answer (right)

4 SUMMARY AND RECOMMENDATIONS

In this paper, we drew on our experience in studying feedback in math education via the analysis of large data collected from experiments that ran on a massive online system (Khan Academy). We highlighted five challenges that may be relevant to other studies taking this approach: population varies greatly over time; inclusion criteria may harm representativity; incomplete population information; changes in product may interfere with experiment; and the unknown impact of interface. In order to decrease obstacles for researchers in the field and to increase validity and generalizability of studies as the ones discussed here, we conclude with a few recommendations for researchers and for the community at large.

First, be well familiar with the learning environments from which the data is drawn. This is not an obvious practical statement. For example, it may be tempting to use pre-collected, cleansed data that is shared on open repositories (like on DataShop or in the context of LAK Challenge). Furthermore, this is not an obvious statement at the theoretical level, as often the important role of the digital environment in the learning process is overlooked (Prinsloo, Slade, & Galpin, 2012).

Second, as data may be time-sensitive (e.g., when different populations are represented in different periods)—which may affect operationalization of research variables (Bergner, Kerr, & Pritchard, 2015)—carefully pick experiment's timing, and even better, run the experiment in a few different periods; one can also divide a large dataset into a few parts, based on time, and repeat the analyses on each of these parts separately (or, alternatively, construct training and testing sets based on these sub-datasets).

Third, as with any research, replications are the greatest tool for demonstrating validity, and our community should encourage such studies (Star, 2018). Additionally, we recommend to use the power of large datasets, but to also think of ways to validate the findings in more controlled environments, using various methodologies.

Finally, we recommend that all the challenges a certain data-based study has faced should be transparently reported. This will allow the relevant research community to objectively evaluate not only what was done, but also what may have been missed or misinterpreted.

Studying feedback (as well as many other learning-related phenomena) via large datasets has many advantages as well as inherent challenges. Upon recognizing and reacting to these challenges, researchers may avoid being failed by them, and may leverage the prospect of their studies instead. This, we believe, will benefit not only the researchers, but also the community at large.

REFERENCES

- Bergner, Y., Kerr, D., & Pritchard, D. E. (2015). Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes. *Proceedings of the 8th International Conference on Educational Data Mining*, 234–241.
- Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, 104(4), 1094–1108. <https://doi.org/10.1037/a0028389>
- Gal, T., & HersHKovitz, A. (2019). Different response-based feedback in mathematics: The case of

- symbolic and textual messages. In *The 9th International Learning Analytics & Knowledge Conference*.
- Howie, E., Sy, S., Ford, L., & Vicente, K. J. (2000). Human - computer interface design can reduce misperceptions of feedback. *System Dynamics Review*, 16(3), 151–171. [https://doi.org/10.1002/1099-1727\(200023\)16:3<151::AID-SDR191>3.0.CO;2-0](https://doi.org/10.1002/1099-1727(200023)16:3<151::AID-SDR191>3.0.CO;2-0)
- Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14* (pp. 117–126). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2556325.2566249>
- Machardy, Z., & Pardos, Z. A. (2015). Evaluating the relevance of educational videos using BKT and big data. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 424–427). Madrid, Spain.
- McMahon, A. D. (2002). Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. *Statistics in Medicine*, 21(10), 1365–1376. <https://doi.org/10.1002/sim.1120>
- Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pairwork. *Language Learning*, 50(1), 119–151. <https://doi.org/10.1111/0023-8333.00113>
- Prinsloo, P. ., Slade, S. ., & Galpin, F. . (2012). Learning analytics: Challenges, paradoxes and opportunities for mega open distance learning institutions. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 130–133. <https://doi.org/10.1145/2330601.2330605>
- Rice, R. E., & Bunz, U. (2006). Evaluating a wireless course feedback system: The role of demographics, expertise, fluency, competency, and usage. *SIMILE: Studies In Media & Information Literacy Education*, 6(3), 1–23. <https://doi.org/10.3138/sim.6.3.002>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Smits, M. H. S. B., Boon, J., Sluijsmans, D. M. A., & van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments*, 16(2), 183–193. <https://doi.org/10.1080/10494820701365952>
- Stamper, J. C., Lomas, D., Ching, D., Ritter, S., Koedinger, K. R., & Steinhart, J. (2012). The rise of the super experiment. In *International Educational Data Mining Society* (pp. 196–199). Chania, Greece: International Educational Data Mining Society. e-mail: admin@educationaldatamining.org; Web site: <http://www.educationaldatamining.org>.
- Star, J. R. (2018). When and Why Replication Studies Should be Published: Guidelines for Mathematics Education Journals. *Journal for Research in Mathematics Education*, 49(1), 98–103. <https://doi.org/10.5951/jresmetheduc.49.1.0098>
- Tangmanee, C., & Nontasil, P. (2014). Perception of delay and attitude toward feedback display: An exploration into downloaders' demographics. *International Arab Journal of E-Technology*, 3(4), 242–249.
- Terzis, V., & Economides, A. A. (2011). Computer based assessment: Gender differences in perceptions and acceptance. *Computers in Human Behavior*, 27(6), 2108–2122. <https://doi.org/10.1016/J.CHB.2011.06.005>
- Turner, G., & Gibbs, G. (2010). Are assessment environments gendered? An analysis of the learning responses of male and female students to different assessment environments. *Assessment & Evaluation in Higher Education*, 35(6), 687–698. <https://doi.org/10.1080/02602930902977723>
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1), 17–37. <https://doi.org/10.1080/00461520.2016.1207538>

Student responses to data-driven feedback in two first-year courses

Author: Lisa Lim

University of South Australia
lisa.lim@unisa.edu.au

Author: Hamideh Iraj

University of South Australia
hamideh.iraj@unisa.edu.au

Author: Abelardo Pardo

University of South Australia
abelardo.pardo@unisa.edu.au

Author: Shane Dawson

University of South Australia
shane.dawson@unisa.edu.au

ABSTRACT: The provision of timely and personalised feedback has been shown to be important for students' academic achievement. However, contemporary higher education faces multiple challenges in providing personalised feedback at scale. While learning analytics has been touted as a solution to these challenges, how students respond to this type of feedback remains under-explored. In this paper, we report the preliminary results from the analysis of focus group discussions with students from two different courses, where a data-driven feedback approach was piloted. The results indicate that students acted on the feedback, as measured through affect and self-regulated learning dimensions. From the results, we discuss the possible implications for the deployment of data-driven feedback approaches.

Keywords: feedback, self-regulated learning, learner affect, feedback recipience

1 BACKGROUND

The provision of regular feedback has been recognised as a significant factor in students' academic achievement (Hattie, 2014; Hattie & Timperley, 2007, p. 1). However, research has also shown that the benefits of feedback are not uniform. This is possibly due to the many challenges in providing effective feedback in contemporary higher education (Boud & Molloy, 2013) as well as the way in which students engage with feedback (Winstone, Nash, Parker, & Rowntree, 2017). Learning analytics (LA)-based approaches to feedback such as dashboards and personalised messages present a viable solution in addressing the complexities of teaching at scale. LA automates the collection of learner data and facilitates the provision of personalised, data-driven feedback at scale (Pardo, Poquet, Martinez-Maldonado, & Dawson, 2017). However, knowing *how students respond* to this type of automated feedback remains under-explored. Understanding how students perceive and respond to such feedback is critical to ensure such scaled processes aid student learning and sensemaking. This paper reports the preliminary findings from focus groups with students from two discrete courses in two Australian higher education institutions incorporating automated data-driven approaches to feedback.

2 RELATED RESEARCH

An objective for providing personalised feedback through LA is to foster students' self-regulated learning (SRL). SRL is generally defined as the range of "metacognitive, motivational, and behavioural processes that are personally initiated to acquire knowledge and skill" (Zimmerman, 2015, p. 541). This multidimensional construct has been studied extensively and found to be positively associated with academic success (e.g., Broadbent & Poon, 2015). Although there are several models of SRL (see Panadero, 2017 for a review), all agree that it comprises different yet iterative phases, meaning that what happens in one cycle effects subsequent ones. For this study, we use Zimmerman's (2000) 3-phase socio-cognitive model as a framework to operationalise SRL as it can be applied broadly to describe students' general learning processes. The three phases are:

1. **Forethought phase:** This is the planning and motivational phase, whereby learners set goals and assess their self-efficacy, expectations, and motivation for learning.
2. **Performance phase:** This involves the strategies used for learning tasks, such as the use of imagery, self-instruction, and attention focusing.
3. **Self-reflection phase:** During this phase, learners engage in self-judgment, evaluating the outcomes of efforts in the performance phase, and experience self-reactions such as satisfaction or adaptive/defensive reactions.

LA-based approaches to feedback provide students with their learning data. This point is well noted by Roll and Winne (2015) in stating that: "Learning analytics are reports of analyses of data that describe features of, and factors that influence, SRL" (p.8). This kind of 'data-driven feedback' differs from traditional forms of feedback given by an instructor, in that it is automated, based on rules and/or algorithms that then facilitate feedback provision at scale. Such analytics, for example, may involve reporting student interactions with the online learning environment to aid self-reflection and to foster self-regulated learning processes (SRL) (Pardo, Poquet, et al., 2017). Feedback targeting SRL is effective for facilitating the learning process and improving academic outcomes (Hattie & Timperley, 2007). As proposed by Butler and Winne (1995), externally-provided feedback influences students' self-regulated learning by making them aware of how they are learning (i.e., monitoring), whether they are on the right track, and helping them to know how to adjust their learning strategies to reach learning goals, thereby leading to enhanced achievement. This monitoring provides an internal feedback loop that relies on both internal and external feedback to help students regulate their learning (Winne & Hadwin, 1998). Feedback affects learners' evaluation of the products of their learning and effectiveness of their study tactics and strategies. The evaluation process helps students to know when and how to adjust their learning strategies in order to reach noted learning goals.

While the relationship between feedback and SRL is well noted, the process does assume that students are actively looking for feedback opportunities, and are willing and able to apply the feedback to improve their learning outcomes. However, to date, research on how students respond to data-driven feedback is limited. The growing body of feedback research has seen a paradigm shift: In the old paradigm, feedback was seen as *information* delivered by a teacher to students about the quality of their work and/or performance while in the new paradigm, feedback is a *process* "through which learners make sense of information from various sources and use it to enhance their work or

learning strategies". In other words, the emphasis has shifted from information to action (Carless & Boud, 2018, p. 1). In addition, there has been an increased emphasis on students' sensemaking and incorporation of feedback into their SRL, which is critical to the effectiveness of the given feedback (Price, Handley, & Millar, 2011; Winstone, Nash, Parker, et al., 2017).

Recent research suggests that students are unmotivated to read, understand, and use feedback (see Jonsson & Panadero, 2018 for a review). Interestingly, this phenomenon may also extend to data-driven feedback provided by intelligent tutoring systems (e.g., Harley, Lajoie, Frasson, & Hall, 2017) and technology prompts (e.g., Bannert, Sonnenberg, Mengelkamp, & Pieger, 2015). Although many other student-facing LA-based feedback systems such as dashboards and recommender systems have been developed, there has been limited research examining their impact on student learning (Bodily & Verbert, 2017; Jivet, Scheffel, Drachsler, & Specht, 2018). More recent data-driven feedback systems juxtapose the technology mediated analyses with direct teacher intervention. In essence, a 'human in the loop' involves instructors being able to add customised messages to their students to supplement learner data. Examples of two recent developments are the Student Relationship Engagement System (Liu, Bartimote-Aufflick, Pardo, & Bridgeman, 2017) and OnTask (Pardo et al., 2018). Thus far, published research on initial trials of OnTask have shown positive effects on student satisfaction and academic achievement (Pardo, Jovanović, Gašević, & Dawson, 2017) but again there has been limited work about how students respond to these data-driven feedback messages.

Feedback includes an evaluation of students' work. As such, feedback can be seen to contain elements of judgment (Higgins, Hartley, & Skelton, 2001), which in turn, can elicit strong emotions from recipients (Rowe, 2017). Negative affective responses to feedback may decrease student motivation and uptake of feedback (Pitt & Norton, 2017); this seems to be especially true for low-achieving students (Orsmond & Merry, 2013; Ryan & Henderson, 2017). At the same time, relationships with teachers and fellow students can mediate between feedback and affective responses (Esterhazy & Damşa, 2017). For example if a student receives negative feedback, he/she may feel discouraged, However, care and respect by the teacher can soften the negative feedback and make it more palatable (Fong et al., 2018). In summary, it is worthwhile to understand how students respond affectively to data-driven feedback, especially as such feedback is typically technologically-mediated (through email messages or dashboards) and therefore perceived as neutral and devoid of an interpersonal element.

The foregoing review has highlighted the growing interest in students as active agents in the feedback process as well as a new avenue of exploration carved out by the emergence of data-driven feedback. In view of these research agendas, the present study aimed to understand how students respond to data-driven feedback, by analysing focus-group data from two large enrolment courses which employed data-driven feedback in the form of email messages. This paper reports on the findings related to the following research questions:

RQ1. What are students' affective responses to data-driven feedback?

RQ2. How do students report the impact of their data-driven feedback on their self-regulated learning?

3 METHOD

3.1 Participants and procedure

This study was part of the OnTask project¹, funded by the Office of Learning and Teaching (OLT) of the Australian Government, to design and evaluate an LA-based software tool that assisted instructors to deliver personalised, data-driven feedback at scale. OnTask was trialled in two first-year courses at two Australian universities in 2017. Course A was a biological sciences course with an enrolment of 242, while Course B was a computer engineering course with an enrolment of 601 students. Both courses employed blended learning curricula, with significant portions of online content available for students. In Course A, the OnTask emails were sent out twice in the semester, at Week 5 and Week 9. These emails provided feedback in terms of students' engagement with key learning activities (class attendance, interactions with online content) in the course, as well as performance on the mid-term assessment. In Course B, the OnTask emails were sent out on a weekly basis, providing feedback on students' engagement with the weekly online activities as well as on their performance in the weekly quizzes.

To address the aim of the research, focus group discussions were conducted with 49 volunteers in the final two weeks of the courses, prior to the examination period. ($n_A=25$ and $n_B=24$). The focus group discussions were semi-structured, to allow students to self-report their feelings and reactions freely and to allow the researchers to delve deep into students' feelings and reactions to feedback and find predefined and emergent themes (Yin, 2015). Students were asked about their reactions to the emailed feedback, in particular how they felt upon reading the feedback, and how they acted on it. Examples of these prompts are: "Thinking about those emails, how did you feel when you read them? Why?" and "Did you follow the recommended actions? Why or why not?"

3.2 Data analysis

To address RQ1, qualitative data relating to affect were coded on two commonly applied positive/negative dimensions. For example, this approach is adopted in the circumplex model of emotions by (Pekrun, 2006; Russell, Weiss, Mendelsohn, & Sarason, 1989). The two dimensions are: valence — "positive" (pleasant feelings such as relief, joy) or "negative" (unpleasant feelings such as anxiety, frustration)—and activation—"activating" (increase in physiological arousal) or "de-activating" (decrease in physiological arousal). Positive affect may not be activating, and negative affect may not necessarily be de-activating (Pekrun & Linnenbrink-Garcia, 2012). For example, relief is a positive emotion but not activating, in the sense that it may not galvanise the student to action. In the same way, stress is a negative emotion but it could be activating by creating a sense of urgency in the student to do something in response. To answer RQ2, qualitative data from the focus group sessions were coded deductively, using Zimmerman's (2000) SRL framework (Forethought-Performance-Self reflection; see Section 2 above).

¹ <https://www.ontasklearning.org>

4 RESULTS

4.1 Students' affective responses to data-driven feedback

Students expressed a range of affective responses to their personalised data-driven feedback – positive-activating, such as ‘sense of urgency’, negative-activating, such as ‘guilt’ and ‘stress’, and also negative-de-activating, such as ‘frustration’ and ‘defensiveness’. While there was a dominance of negative affect in response to the feedback (n = 33), this was not necessarily de-activating, but rather, for some students, it created a kind of motivation or a nudge to increase their study efforts; one participant expressed this colloquially as “a kick in the bum”. In terms of positive affective responses, the idea of reassurance featured in a number of responses (n=6). Students felt a sense of comfort or relief when the feedback indicated they were on the right track or doing well, and that this had a positive impact on their motivation to learn in the course.

4.2 How students reported the impact of data-driven feedback on SRL

From deductive coding of statements relating to Zimmerman’s (2000) framework, the data-driven feedback affected all stages of SRL. However, a higher proportion of students reported that the feedback impacted on their *self-reflection* and *forethought* processes.

4.2.1 Forethought

A frequently mentioned theme around forethought was the impact on students’ strategic planning: the emails reminded the students to do what they had neglected, helped them to categorise their tasks, or guided them to turn their weaknesses into strengths. A number of students also noted an effect on their goal-setting. Some students reported that the emails helped keep them on-track for more consistent study, while other students used the information to calibrate their effort to study goals (defined by grades).

Almost half of the students commented that the feedback did enhance their motivation for learning in the course. For many students, the reason for this was the care perceived by the lecturer. Other reasons related to the feeling of greater accountability, being encouraged, peer competition, and the thought that someone was watching them.

4.2.2 Performance

Feedback enabled students to enhance self-control over their learning in the course, by nudging them to “get back to work”. It also drew students’ attention to educational resources such as the course videos, slides and textbooks, and also suggested productive strategies such as doing “short bursts of study”.

4.2.3 Self-reflection

Feedback helped many students reflect on how they were learning, whether they had fallen behind; in that sense it focused on what students needed to do in order to stay on track. Other students reflected on the content of their learning. The feedback informed them of potential topics for further study. As expressed by one of the respondents, “Every week [the instructor] will send you back feedback on whether you need to study more or less for that stuff that you thought was easy or hard”.

In terms of self-reaction, a number of students reported defensive reactions, especially when they felt that their own study methods worked, therefore there was no need to follow the recommendations. Other students also expressed that they were negatively evaluated, e.g., “it was kind of like you’re not doing this, you’re not doing this”.

5 DISCUSSION

This paper presented the preliminary results from a qualitative analysis of focus group data in two courses. The results provide evidence of students’ engagement with data-driven feedback, in terms of affect and SRL. The majority of students were responsive to the feedback and were able to use the feedback, especially in their forethought and reflection. The main contribution of this research is the affective responses of students, which bears implications for the deployment of data-driven feedback. The finding of negative-activating emotions supports the idea raised in Pitt and Norton (2016), that students have the potential to take negative feedback, overcome their bad feelings and use the feedback constructively to improve their performance. However, some students expressed *de-activating emotions*, such as frustration and even a sense of punishment in response to the feedback. An important implication, therefore, is that data-driven feedback should try to reduce negative deactivating affect, and push students toward more activating responses. As a general rule, educators and course designers should minimise ‘emotional backwash’ and maximise student success (Pitt & Norton, 2017) because what students (not educators) perceive as useful helps students to learn (Price, Handley, Millar, & O'Donovan, 2010; Winstone, Nash, Rowntree, & Parker, 2017).

The next steps of this research are to investigate at a deeper level, how contextual factors of the teaching and learning environment—the learning design, perceived course difficulty, student motivational factors—affect the responses of students to data-driven feedback. We will also continue to conduct more case studies in different contexts to investigate the consistency of the results

REFERENCES

- Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short- and long-term effects of students’ self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior*, 52, 293-306. doi:<https://doi.org/10.1016/j.chb.2015.05.038>
- Bodily, R., & Verbert, K. (2017). Review of Research on Student-Facing Learning Analytics Dashboards and Educational Recommender Systems. *IEEE Transactions on Learning Technologies*, 10(4), 405-418. doi:10.1109/TLT.2017.2740172
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698-712. doi:10.1080/02602938.2012.691462
- Broadbent, J., & Poon, W. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1-13. doi:10.1016/j.heduc.2015.04.007
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245-281.

- Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 1315-1325. doi:10.1080/02602938.2018.1463354
- Esterhazy, R., & Damsa, C. (2017). Unpacking the feedback process: an analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, 1-15. doi:10.1080/03075079.2017.1359249
- Fong, C. J., Schallert, D. L., Williams, K. M., Williamson, Z. H., Warner, J. R., Lin, S., & Kim, Y. W. (2018). When feedback signals failure but offers hope for improvement: A process model of constructive criticism. *Thinking Skills and Creativity*. doi:<https://doi.org/10.1016/j.tsc.2018.02.014>
- Harley, J. M., Lajoie, S. P., Frasson, C., & Hall, N. C. (2017). Developing emotion-aware, advanced learning technologies: A taxonomy of approaches and features. *International Journal of Artificial Intelligence in Education*, 27(2), 268-297. doi:10.1007/s40593-016-0126-8
- Hattie, J. (2014). *Visible learning for teachers: Maximizing impact on learning*. Florence, GB: Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. doi:10.3102/003465430298487
- Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: the problem of communicating assessment feedback. *Teaching in Higher Education*, 6(2), 269-274.
- Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2018). License to evaluate: Preparing learning analytics dashboards for educational practice. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia*. New York, NY, USA: ACM.
- Jonsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback*. Cambridge: Cambridge University Press.
- Liu, D. Y.-T., Bartimote-Aufflick, K., Pardo, A., & Bridgeman, A. J. (2017). Data-driven personalization of student learning support in higher education. In A. Peña-Ayala (Ed.), *Learning Analytics: Fundamentals, Applications, and Trends, Studies in Systems, Decision and Control*, 94 (pp. 143-169). Springer International Publishing.
- Orsmond, P., & Merry, S. (2013). The importance of self-assessment in students' use of tutors' feedback: A qualitative study of high and non-high achieving biology undergraduates. *Assessment & Evaluation in Higher Education*, 38(6), 737-753. doi:10.1080/02602938.2012.697868
- Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology*, 8(422). doi:10.3389/fpsyg.2017.00422
- Pardo, A., Bartimote-Aufflick, K., Buckingham Shum, S., Dawson, S., Gao, J., Gašević, D., . . . Vigentini, L. (2018). OnTask: Delivering Data-Informed Personalized Learning Support Actions. *Journal of Learning Analytics*, 5(3), 235-249. doi:10.18608/jla.2018.53.15
- Pardo, A., Jovanović, J., Gašević, D., & Dawson, S. (2017). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*. doi:10.1111/bjet.12592
- Pardo, A., Poquet, O., Martinez-Maldonado, R., & Dawson, S. (2017). Provision of data-driven student feedback in LA and EDM. In C. Lang, G. Siemens, A. F. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (pp. 163-174). Society for Learning Analytics Research and the International Society for Educational Data Mining.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315-341.

- Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 259-282). Boston, MA: Springer US.
- Pitt, E., & Norton, L. (2017). 'Now that's the feedback I want!' Students' reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education*, 42(4), 499-516. doi:10.1080/02602938.2016.1142500
- Price, M., Handley, K., & Millar, J. (2011). Feedback: focusing attention on engagement. *Studies in Higher Education*, 36(8), 879-896. doi:10.1080/03075079.2010.483513
- Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3), 277-289.
- Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, 2(1), 7-12.
- Rowe, A. D. (2017). Feelings about feedback: The role of emotions in assessment for learning. In *Scaling up Assessment for Learning in Higher Education* (pp. 159-172). Singapore: Springer.
- Russell, J. A., Weiss, A., Mendelsohn, G. A., & Sarason, I. G. (1989). Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57(3), 493-502. doi:10.1037/0022-3514.57.3.493
- Ryan, T., & Henderson, M. (2017). Feeling feedback: students' emotional responses to educator feedback. *Assessment & Evaluation in Higher Education*, 1-13. doi:10.1080/02602938.2017.1416456
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, A. C. Graesser (Eds). *Metacognition in educational theory and practice*, 27-30. Mahwah, NJ : Lawrence Erlbaum Associates.
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1), 17-37. doi:10.1080/00461520.2016.1207538
- Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. *Studies in Higher Education*, 42(11), 2026-2041. doi:10.1080/03075079.2015.1130032
- Yin, R. K. (2015). *Qualitative research from start to finish*. New York: Guilford Publications.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82-91. doi:10.1006/ceps.1999.1016
- Zimmerman, B. J. (2015). Self-regulated learning: Theories, Measures, and outcomes. *International Encyclopedia of the Social & Behavioral Sciences (2nd edition)*, 21, 541-546. doi:10.1016/B978-0-08-097086-8.26060-1

It's all in the details: why there are different versions of OnTask

Lorenzo Vigentini
UNSW Sydney
l.vigentini@unsw.edu.au

ABSTRACT: This paper takes for granted that feedback is an integral part of learning, and that generally, feedback is good for learning because it enables learners to adjust their strategies to achieve their learning goals. This is grounded in seminal work in higher education, as well as the conceptual and practical work driven by a recent large research collaborative project (OnTask) to empower educators to use the data available to them and deploy personalised learning support actions (PLSA). The main focus of this paper is specifically on the process and the building blocks of the supporting software, analysing the different ways in which the same conceptual model can lead to different implementations. Additionally, the paper provides a taxonomy to evaluate the design choices required to turn the model into a working software tool. The outcome is an appraisal of three different versions of the software resulting from the project, which demonstrates that development and design choices are an essential aspect of the process to formalise a development roadmap, which current and future partners can relate and align to their contexts.

Keywords: personalising feedback, software development, learning analytics

1 INTRODUCTION

The initial premise of this paper is that feedback is seen as an integral part of learning (Hattie & Timperley, 2007; Boud, 2012). Feedback for learning is intended as any information (this could be teacher-driven, but it could also come from other sources such as resources, other agents such as peers, parents or automated computer tutors, or even self and experience), which is provided to a learner regarding any aspect of one's understanding, performance and achievement (Bloom, 1968; Evans, 2013; Ramaprasad, 1983; Wiliam, 2011). Feedback aims to bridge the gap between the actual level of performance in a task or assessment with the desired learning goal or outcome (Hattie & Timperley, 2007; Poulos & Mahony, 2008; Nicol, 2010). Broadly, the literature presents two main types of feedback: one puts emphasis on the 'telling', feedback as intervention, or the summative purpose of feedback -implying that feedback is used to alert the learner of this gap and must have an impact on the learning process and lead to mastery (Kulik, Kulik, & Bangert-Drowns, 1990; Kluger & DeNisi, 1996). The other focuses on the 'dialogic' function of feedback, in which the focus is on the iterative process of clarifying and resolving misconceptions, and the shared construction of meanings (Nicol, 2010; Carless, Salter, Yang, & Lam, 2011; Carless, 2018). Either way, there seems to be a universal agreement that feedback is good for learning from both the student and the teacher perspectives. However, a key weakness noted is that the provision of feedback at scale has several intrinsic problems in terms of efficiency and effectiveness and Carless has (Carless, 2018; Carless et al., 2011) described at length the challenges for providing effective and sustainable feedback. Significant progress has been made on tackling the challenging of scaling personalised, formative feedback. The pioneering work was at University of Michigan, whose ECoach system has demonstrated the impact of using data from students' online learning activity to give them feedback (Huberth, Chen, Tritz, & McKay, 2015). Similarly, Liu and colleagues (Liu, Bartimote-Aufflick, Pardo, & Bridgeman, 2017; Vigentini et al., 2017; Arthurs et al., 2019) have demonstrated over several years and across multiple institutions that it is possible to deploy feedback at scale with a software tool that leverages on the

data made available during the learning and teaching process ([SRES](https://sres.io) - <https://sres.io>). The data does not need to be sophisticated, often focusing on basic proxy measures of engagement (like attendance or activity) and using specific nudges to encourage students to complete activities and explore resources provided to them. Pardo and colleagues (2018) extended this idea by creating a framework for *OnTask* (<https://www.ontasklearning.org>) and shifting the focus on the purpose of the feedback generated from the data. OnTask is a software tool resulting from the collaboration between several universities funded by the Australian Office of Learning and Teaching, which aimed to empower educators to use learning analytics to drive *Personalised Learning Support Actions* (PLSA). As explained in detail in Pardo et al (2018) the project provided a unique opportunity to work on three aims: 1) develop a conceptual model in which student information (data) is captured in a basic set of rules by the instructor to deploy PLSA; 2) design a software architecture to enable the deployment of PLSA; and 3) the implementation of an open source platform to realise the vision.

2 ONTASK: SCALING THE PROVISION OF FEEDBACK IN DIFFERENT INSTITUTIONS

Although the project builds on incremental advancement in the fields of learning analytics, there are several interesting features which make the OnTask project unique and valuable. First of all, the shared understanding of the potential of effective feedback was an essential driver behind the various stakeholders in the project. Stressing the importance on designing assessment as the best approach for learning as part of a continuing feedback cycle, the most effective form of feedback is a dialogic one in which feedback is most valuable when provided at the *learning process* and *self-regulation* levels (Nicol, 2010; Carless et al., 2011). However, this sort of feedback becomes impractical and extremely challenging to scale to large numbers of students and it is the main driver behind the focus of the project on feedback at scale. The second essential aspect is the shift from the sort of summative feedback on performance, to *Personalised Learning Support Actions* (PLSA). These are focusing on learners' activities that may enable them to change strategies or behaviours, known to be either effective or detrimental for learning, and prompt to action. Secondly, all the partners have an in-depth involvement in the field of learning analytics, with several initiatives already ongoing at their respective institutions. For example, the SRES tool mentioned earlier (Liu et al., 2017) had been successfully implemented in two of the partner institutions (Arthars et al., 2019; Vigentini et al., 2017). Large LA initiatives are ongoing at the University of South Australia, UTS and the University of Edinburgh with leading figures steering learning and teaching, which includes both governance (Tsai et al., 2018) and implementation (Macfadyen, Dawson, Pardo, & Gašević, 2014; Bakharia, Kitto, Pardo, Gašević, & Dawson, 2016; Colvin et al., 2016). In all institutions involved there is a strong push from institutional strategies driving programs and activities (e.g., the [UNSW 2025 strategy](#), the [Edinburgh digital transformation strategy](#), [UniSA](#), or [University of Technology Sydney's learning.futures](#)). In most cases there is also a strong bottom-up interest from instructors (the tinkerers like Jurgen Schulte, Abelardo Pardo, Danny Liu, Lorenzo Vigentini) to actually use data to inform their L&T activities, as well as a top-down support from senior academic managers supporting the initiatives (like Shane Dawson, Simon Buckingham Shum, Dragan Gasevic and George Siemens, as examples of senior academic managers in their respective institutions). Finally, an essential enabling element in the project is the presence of a range of academic roles (from academics in the faculty to senior academic roles, as well as dynamic technically-oriented individuals) which provided a very fertile ground to achieve the set goals of the project. This is discussed further, below.

Although the large collaboration aimed to develop a ‘university agnostic’ software system, in which the focus was predominantly on the conceptual foundation for the deployment of PLSA and the design of a generic architecture which would have enabled the implementation of a software platform capable to delivering this vision, the fact that there are currently 3 main implementations of the tool, reflects the fact that there are important choices to be made to build 6 core blocks of the software architecture (Figure 1).

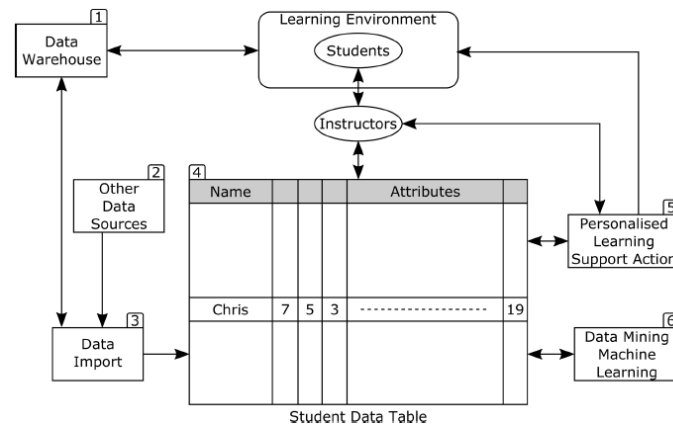


Figure 1. Model for the provision of PLSA to students (reproduced from Pardo et al 2018)

This paper aims to dissect the layers of complexity around the implementation of the tool and provide a scaffold to enable current and future partners to carefully consider the entire design process. It also provides an in-depth description of the development journey and the technical choices that brought UNSW to differentiate from the other two existing versions currently available¹.

3 DESIGN THINKING DRIVING THE CONCEPTUALISATION

As reported in West et al (2016) and Colvin et al (2016), using only the conventional student data available in institutional data warehouses decreases the effectiveness of applications focusing on early intervention and retention: they specified that the database should include a range of data sources, including self-reported student measures. Gasevic and colleagues strongly warned against the use of ‘one-size-fits-all’ adoption and applications of models generated from learning analytics methods without keeping into account the educational context and the instructional design (Gasevic, Jovanovic, Pardo, & Dawson, 2017). Wise and colleagues explored how to implement LA appropriately and presented a model for conceptualizing students' learning analytics use as part of a self-regulatory cycle of grounding, goal-setting, action and reflection putting the students at the centre as users of analytics (Wise, Vytasek, Hausknecht, & Zhao, 2016). Further, a specific framework is needed to include analytics as one of the sources of evidence to evaluate practice and the improvement of teaching systematically and make it another tool in the teacher evaluation toolkit (Vigentini, Mirriahi, & Kligyte, 2016). Within this backdrop of extensive work in LA, it is not a surprise that at the centre of the conceptual model for OnTask is a student *data table* (or matrix), which aggregates key information about the student. Pardo et al (2018) describe in detail all the components and interactions between

¹ Details of the project can be found on <https://www.ontasklearning.org> and additional material associated with this paper will be available on the workshop website: <https://goo.gl/KAf97t>

them (Figure 1, previous page). However, although the model seems to be simple, there are several pragmatic questions which highlight the next tier of questions that surrounds deployment of the model in real platforms: for example, moving beyond the discussion about what type of data is deemed suitable, how does the data get into this *matrix*? There are three obvious sources: 1) a *teacher* user (and a user's role needs to be defined) uploads it into the system, 2) the data is linked directly (requiring integration) from another source or 3) the *student* user can enter the data themselves (which may lead to gaps). Following from this, at what frequency should the data be updated? Who has access to change data? What happens when something is updated?

Very rapidly, the relatively simple concept of the student matrix becomes more complex, with important design and implementation choices required in order to account for data changes, timeliness of the data and accuracy of the data. These choices are non-trivial when one considers the end users, their roles and needs, and applies design thinking to tackle the complexity of the problem in detail. Since the start of the project, partners clearly articulated the type of users/persona who may be involved in the use, dissemination and support of the OnTask tool. Immediately it was clear that the tool required an ecosystem of roles at different levels of institutional organisation with fundamentally different functional goals. Six broad categories were identified: 1) students, 2) academics, 3) educational designers (umbrella term including support roles to academics which could be also labelled as educational developers, instructional designers and academic developers), 4) support teams (including IT, student support, student services etc.), 5) management roles (from data management/governance to senior academic management), 6) researchers (both in the project team as well as future researchers). All these roles are characterised by a wide range of characteristics and we focus here on some specific examples to provide an overview. For example, students are at different levels of their studies, integration, performance and motivation. Describing a range of academics involved in learning and teaching activities is essential: some are champions of Learning & Teaching, while others would prefer doing something else; the level of experience and seniority in the organisation determines their commitment and time-involvement or motivation to innovate or take on the challenges of adopting new systems/approaches. Educational designers have a key role in supporting academics and sometimes take on active roles in actually doing things for the academics requiring a great level of synergy to make the relationship work and ultimately benefit the students. The buy in and uptake of support teams outside the individual courses is paramount: ensuring that IT enables integration with system and processes is as important as ensuring that once an issue is flagged to a student, they can also take action and use the existing support structures already present at their institutions. From this quick overview it is easy to see how the model in Figure 1 may not be descriptive enough to account for the different functions and needs for the range of individuals involved. To take on example, at UTS, the Institute for Interactive Multimedia & Learning (IML) leads all academic professional development around teaching and learning. To scale OnTask briefing and training within UTS, specific staff now have OnTask formally in their learning technology portfolios, working closely with the Connected Intelligence Centre, which leads UTS learning analytics innovation. They support early adopter academics to help maximise the success of their pilots. In parallel, conversations are developing with the IT Division, who need to understand how to embed OnTask within the enterprise architecture to run as a 24/7 production service. With similar conversations unfolding at multiple institutions, a network is forming to share experience and expertise, in order to coordinate technical decisions around open source components that will interoperate in diverse institutional ecosystems (e.g. [OnTask & LA-Architecture Workshop](#), UTS - 18 Feb 2019).

4 A TAXONOMY OF DESIGN CHOICES

From the discussion about the roles and responsibilities involved in the deployment and use of the tool it was apparent at the start that design choices were required to turn the idea into a tool. In order to capture dependencies with contextual enablers, technical choices, and functional design choices a taxonomy has been created around four main categories (Table 1): the first two (integration and technical implementation) are represented as layers of the technology stack, the second two (usability design and functional design) are firmly grounded in design thinking. Each category has several items to capture different aspects (see workshop site for additional material).

Table 1. A taxonomy of design choices in the implementation space for OnTask

Layers of the technology stack	Integration layer (institutional enablers)	Data (institution-dependent data provisioning)
		Messaging (institution-dependent messaging/Customer Relations Management capability)
		Support services (institution-dependent support services)
		Authentication layer (leveraging on federation access, SAML and OAuth)
		Learning Management System integration and data standards (LTI, Caliper, XAPI)
		APIs and connectors (programmability and integration)
	Technical implementation layer	Data models, data standards and databases (SQL vs noSQL)
		Data operation capabilities (data manipulation + libraries)
		Data mining capabilities (including cognitive services)
		Data visualisation capabilities (from chart to dashboards)
		Scheduling
		Backend architecture and scalability of service (cloud-based cluster options)
		Choice of programming language
		Choice of coding frameworks and libraries (Open source preferred)
Aspects of design	Usability design	Good practice: decoupling of frontend/backend
		UX/UI (between aesthetic and functional)
		Input/output flows and action design
		User data literacy
	Function-based design choices	The matrix and the end-user (focus on course, teacher or flexibility)
		Granularity of access (functional roles in the system)
		ITTT (if-this-then-that) rules and filters
		The building blocks for messaging (email/static pages)
		Exportability and sharing (of everything)
		UX/UI (between aesthetic and functional)
		Input/output flows and action design
		User data literacy
	Function-based design choices	The matrix and the end-user (focus on course, teacher or flexibility)
		Granularity of access (functional roles in the system)
		ITTT (if-this-then-that) rules and filters
		The building blocks for messaging (email/static pages)
		Exportability and sharing (of everything)

4.1 Integration layer

This is a core part of the design and development because it provides a set of design requirements without which the software tool would not fit in either the technology stack of a university nor their strategic directions. It also provides the enablers (or hooks) which will allow to have an appropriate conversation with IT roles, data governance roles and other senior academic managers. The specific

thinking around other support services is an essential aspect to create a strategic fit for the tool in what the university is already (or should be) doing.

4.2 Technical implementation layer

This category includes all the technical decisions which specify what the software tool will be able to do. As there are several options, it is essential to accurately articulate in detail what one choice will do (a kind of SWOT analysis) for every step. The most important aspect is that related to data: as it is possible to detect from the list in Table 1, apart for the data models, which is as much philosophical discussion as a pragmatic one, the other items require design thinking, and represent what a user should be able to do. The three slightly different interpretations of what the students' data table looks in the three versions of OnTask is exemplary (details here: <https://goo.gl/KAf97t>). Further, the choice of both programming languages and backend architecture are important enablers which ensure the sustainability of development as well as the placement in individual universities' technology stack.

4.3 Usability design

Although it seems obvious, UX design should be at the centre of tool design, yet, in many cases, and OnTask was no exception, the focus on the key functionality of the system meant that the granularity of the user-stories was probably not detailed enough to drive system design. This was especially clear when the assumptions behind input and output flows as well as the implied technical and data literacy of the end-users led to fairly systematic changes in the UI design of the three versions. While the refinement of the rule creation and action interface was the first of the UI changes, the representation of the student data table also required some major work after the users started to use the system. While the exciting ambition remains of upskilling academics to manage their own OnTask accounts, it is clear from our consultations with them that many do not want to ever see a data table ("the data just needs to be there"), and want to think about nothing more than the rules, with a 'consultant' to drive the OnTask tool, and assist in translating their informal rules into the necessary formal specifications. The user experience and learner experience dimensions to OnTask, and this class of tools more broadly, are of growing interest (e.g. [PLSA UX/LX Workshop](#), UTS - 19 Feb, 2019).

4.4 Functional design

In the functional design category, there are more specific design requirements which enable the end-users to manipulate the data, create actions and determine the rules driving these actions. They also provide some reference to the ideal functionalities which the users would want to use based on their specific user-roles. These are firmly anchored in the software development cycle, requiring specific user input and feedback to align users' needs and the implementation.

5 FROM CONCEPTUAL MODEL TO DESIGN AND IMPLEMENTATION CHOICES

The design taxonomy just described provides a more in-depth set of dimensions to consider turning the conceptual model (figure 1) into a usable and sustainable software tool. Despite the apparently simple model, the set of beliefs grounded in the partners' institutional contexts and understanding as well as their own experience in practice was a key element driving the developments. This became obvious through the emphasis given in the implementation of different stages of the project, which resulted in the three versions of the tool:

- A course-centric approach
- A teacher-centric focus
- A container-centric framework

In the first version of the tool, the course was the key organisational unit around which the software tool was also organised. This approach seems logical as it enables to organise the tool visually as well as enable management of data and users. It also allows for a seamless integration with the LMS and the distinction between the student role and the instructor role (which ideally is also provisioned by the institution identity management system or directly from the LMS). There are two fundamental issues emerging from this approach: there is no intrinsic hierarchy to manage program-level approaches (i.e. think about several first year courses, or a postgraduate program all sending messages to students without a higher level management), and the granularity of access for the different functional roles is also limited (i.e. think about an educational designer helping the lead instructor on a course, or tutors accessing subset of data).

In the second version the focus shifted to the ability for the educator to quickly create actions. In order to do this the focus shifted considerably on their ability to enter, import and manipulate data, from which a decision could be made to create a PLSA. The shift to the Django framework and the intrinsic model-based development in which the underlying data model allows for a quick development of a functional component has the advantage of making the development quick. However, it also creates a fundamental dependency to a specific approach to front-end development which is not ideal in modern MVC approaches. Despite the potential of been deployed in a scalable architecture, there is still no concept of organisational hierarchies which go beyond the course level. Finally, as it was the case for the first version, the data model relying solely on SQL relational databases creates fundamental issues with the scalability and timeliness of the data (including versioning).

The third version of OnTask set out to remediate these issues: starting from the python-based Django implementation of version 2, the front-end and back-end were completely decoupled, enabling the overlay of a React framework which provides great flexibility in UI development with a modern and popular framework. Secondly the shift from SQL to no-SQL data model provides a more flexible approach to data organisation, scheduling and versioning. Removing the dependency to the data model means that some of the advantages coming ‘out-of-the-box’ with version 2 were partly lost. However, refactoring the code enabled a truly scalable architecture which can be easily deployed in a container-based infrastructure: technically this means high portability and scalability which will enable the ‘tinkerers’ to quickly deploy the tool in their environment and reduce the entry level issues which the early adopters have been facing. Further, the decoupling of the backend also means a greater flexibility in the development of APIs that will allow to expand the functionality of the system without affecting the way the application looks like.

Although the tool has seen a wide interest in the community across several institutions, there are still several improvements required. The taxonomy proposed will be useful in guiding the collaborative work which will be key for the further adoption and diffusion of the technology supporting PLSA.

6 REFERENCES

Arthars, N., Dollinger, M., Vigentini, L., Liu, D., Kondo, E., & King, D. (2019). Empowering teachers to personalize learning support. In *Utilizing Learning Analytics to Support Study Success*. Retrieved from <https://www.springer.com/gp/book/9783319647913>

- Bakharia, A., Kitto, K., Pardo, A., Gašević, D., & Dawson, S. (2016). Recipe for Success: Lessons Learnt from Using xAPI Within the Connected Learning Analytics Toolkit. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 378–382). New York, NY, USA: ACM. <https://doi.org/10.1145/2883851.2883882>
- Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–5.
- Boud, D. (2012). *Developing Student Autonomy in Learning* (2nd ed.). New York: Taylor and Francis.
- Carless, D. (2018). Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education*, 0(0), 1–10. <https://doi.org/10.1080/02602938.2018.1531108>
- Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, 36(4), 395–407. <https://doi.org/10.1080/03075071003642449>
- Colvin, C., Rogers, T., Wade, A., Dawson, S., Gasevic, D., Shum, S. B., ... Kennedy, G. (2016). *Student retention and learning analytics: A snapshot of Australian practices and a framework for advancement* (Project report). Canberra: Australian Government Office for Learning and Teaching. Retrieved from http://130.56.250.163/wp-content/uploads/SP13_3249_Dawson_Report_2016-3.pdf
- Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, 83(1), 70–120. <https://doi.org/10.3102/0034654312474350>
- Gasevic, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. *Journal of Learning Analytics*, 4(2), 113–128. <https://doi.org/10.18608/jla.2017.42.10>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-Tailored Student Support in Introductory Physics. *PLOS ONE*, 10(9), e0137001. <https://doi.org/10.1371/journal.pone.0137001>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of Mastery Learning Programs: A Meta-Analysis. *Review of Educational Research*, 60(2), 265–299. <https://doi.org/10.3102/00346543060002265>
- Liu, D. Y.-T., Bartimote-Aufflick, K., Pardo, A., & Bridgeman, A. J. (2017). Data-Driven Personalization of Student Learning Support in Higher Education. In *Learning Analytics: Fundamentals, Applications, and Trends* (pp. 143–169). Springer, Cham. https://doi.org/10.1007/978-3-319-52977-6_5
- Macfadyen, L. P., Dawson, S., Pardo, A., & Gašević, D. (2014). Embracing Big Data in Complex Educational Systems: The Learning Analytics Imperative and the Policy Challenge. *Research & Practice in Assessment*, 9, 17–28.
- Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5), 501–517. <https://doi.org/10.1080/02602931003786559>
- Pardo, A., Bartimote, K., Shum, S. B., Dawson, S., Gao, J., Gašević, D., ... Vigentini, L. (2018). OnTask: Delivering Data-Informed, Personalized Learning Support Actions. *Journal of Learning Analytics*, 5(3), 235–249–235–249. <https://doi.org/10.18608/jla.2018.53.15>
- Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education*, 33(2), 143–154. <https://doi.org/10.1080/02602930601127869>
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13. <https://doi.org/10.1002/bs.3830280103>
- Tsai, Y.-S., Moreno-Marcos, P. M., Jivet, I., Scheffell, M., Tammets, K., Kollom, K., & Gašević, D. (2018). The SHEILA Framework: Informing Institutional Strategies and Policy Processes of Learning Analytics. *Journal of Learning Analytics*, 5(3), 5–20–25–20. <https://doi.org/10.18608/jla.2018.53.2>
- Vigentini, L., Kondo, E., Samnick, K., Liu, D. Y., King, D., & Bridgeman, A. J. (2017). Recipes for institutional adoption of a teacher-driven learning analytics tool: Case studies from three Australian universities. In *ASCILITE Annual Conference Proceedings*. University of Southern Queensland. Retrieved from <http://2017conference.ascilite.org/wp-content/uploads/2017/11/Full-VIGENTINI.pdf>
- Vigentini, L., Mirriahi, N., & Kligyte, G. (2016). From Reflective Practitioner to Active Researcher: Towards a Role for Learning Analytics in Higher Education Scholarship. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, Design, and Technology* (pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_6-1
- West, D., Huijser, H., Heath, D., Lizzio, A., Toohey, D., Miles, C., ... Bronnimann, J. (2016). Higher education teachers' experiences with learning analytics in relation to student retention. *Australasian Journal of Educational Technology*, 32(5), 48–60.
- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
- Wise, A. F., Vytasek, J. M., Hausknecht, S., & Zhao, Y. (2016). Developing Learning Analytics Design Knowledge in the “Middle Space”: The Student Tuning Model and Align Design Framework for Learning Analytics Use. *Online Learning*, 20(2). Retrieved from <http://olj.onlinelearningconsortium.org/index.php/olj/article/view/783>

Exploiting data intelligence in education from three levels: Practice, challenges and expectations

Xiaoqing Gu
East China Normal
University, CHN
xqgu@ses.ecnu.edu.cn

Bian Wu
East China Normal
University, CHN
bwu@deit.ecnu.edu.cn

Yiling Hu
East China Normal
University, CHN
ylhu@deit.ecnu.edu.cn

David Gibson
Curtin University, AUS
david.c.gibson@curtin.edu.au

ABSTRACT: Technological advances have a strong effect on the way instructors teach and students learn. Over the last couple of years, big data has been in and out of the center of focus for emerging technology. There seems to be potential for data intelligence in education to have a big impact on teaching and learning. Not only is it valuable to have large-scale automated data monitoring and reporting, but it is key to have functional capabilities to make decision at all levels of the educational system to expand impact, effectiveness, and efficiencies.

Background

Technological advances have a strong effect on the way instructors teach and students learn. Over the last couple of years, big data has been in and out of the center of focus for emerging technology. The value of big data in education has yet to be unfolded, especially the value of data intelligence in education. There seems to be potential for data intelligence in education to have a big impact on teaching and learning. Not only is it valuable to have large-scale automated data monitoring and reporting, but it is key to have functional capabilities to make decision at all levels of the educational system to expand impact, effectiveness, and efficiencies.

Serious games, smart tools, cloud computing, machine learning, modeling, sensors, and other current and emerging technologies are redefining the tools and capabilities of education. Data intelligence in education will not only provide a stronger functional link to the integration of these technologies but it will greatly support an educational impact at the individual level. Analyzing actual and future capabilities of these teaching and learning technologies involves a complex interplay of technological, pedagogical, and political issues.

There are special interest groups coming up towards the topic of the data intelligence in education, one of which are scholars from a cross disciplinary groups based on the East China Normal University (ECNU) and University of North Texas (UNT) Joint Research Laboratory. Established in 2014, this group has worked to formalize and further collaborate among researchers and scholars in the broad area of learning technologies. ECNU has specific strengths in the area of advanced technologies (e.g., smart technologies) and their application in support of learning and instruction. UNT has specific strengths in the area of research and evaluation of advanced learning technologies. Two groups are working closely together focusing on the learning design, which is afforded by big data and smart technologies.

As the research of big data is unfolding, there are new perspectives of understanding the value of big data. New ways of exploiting the intelligence behind the data is emerging and evidence that this trend of data intelligence, instead of data science, can ask bigger questions and builds models to solve for various complex questions in education.

With data intelligence, questions in education that can and have been answered at three different levels:

Level 1—Intelligent education governing. The current practice that has taken effect in governing is focused on the education equality problems afforded with data intelligence. For example, in Shangrao area, Jiangxi province, where the disadvantaged students are distributed across the area, their status of caregiving, schooling, and their academic performance can be analyzed with data intelligence.

Level 2—Practices of teaching and instruction. When the big data of teaching and learning process and behaviors are systematically tracked and recorded, questions can be answered such as how well courses design, teaching and learning can be discovered. For example, by collapsing students learning data together, and using behavioral modeling for “Classroom orchestration”, the overall teaching quality of the courses, the weakness of the students learning, and the possible relationship among different subjects can be turned out and addressed accurately.

Level 3—Individual learning process. When learning process data is tracked, including the adaptive measurement during learning progress, skill acquisition, and dynamic interaction specific learning interventions can be adapted precisely to the individual level.

Organisational details

Type of event: workshop

Proposed schedule and duration: half-day

The workshop comprises three stages. Stage one features presentations on data intelligence in three different levels. In stage two, group discussions on the presentations ensue with the aim of understanding by invoking the broader perspectives of researchers from the audience. The third stage is to seek consensus on what we have learned and what we can do further to pursue data intelligence in education.

Type of participation: ‘mixed participation’

An open call for paper/participation will be disseminated to invite practitioners, researchers, and industries who are interested in this topic to submit full papers, short papers and/or industrial plans of data intelligence in education. Whereas any interested delegate may register to attend this workshop.

The workshop activities: symposia elements

Planned length of the workshop: ½ day

Expected participant numbers: 50

Planned dissemination activities to recruit attendants: Open call for paper

Full papers should be no more than 10 pages describing original (unpublished) research results, short papers are no more than 4 pages describing ongoing research, and industrial plan are around 4 pages describing the design of data intelligence in a particular educational problem.

Required equipment for the workshop: projector

Target audience

We cordially invite practitioners, researchers, and industries who are interested in this topic to submit full papers, short papers and/or industrial plans of data intelligence in education.

Papers will directly focus on data intelligence in education looking at Practice, Research, and Impact: Sharing Practices of Data Intelligence in Education

- Precise education governing with data intelligence
- Adaptive teaching with data intelligence

- Individualized and adaptive learning with data intelligence

Research on Issues and Challenges of Data Intelligence in Education

- Standard of big data in education
- Data technologies: text, graphic, audio/video, gestures, or sensor technologies
- Data analysis: methods, models, and trends

Envisioning the Impact of Data Intelligence on Education

- Underlying principles of data intelligence in education
- What works with data intelligence in education
- State of the art of data intelligence in education
- Future of data intelligence in education
- Building partnership to increase educational impact

Objectives and intended outcomes

Practice in these different levels are emerging as well as associated challenges, such as the challenge of data tracking, interchanging, analyzing models, and privacy issues. These challenges need to invite researchers and practitioners with multiple perspectives and expertise to have conversations.

The proposed workshop will bring a broad spectrum of stakeholders together to look at the practice, challenges and expectations of data intelligence in education. The workshop will explore such issues with leaders from areas that contribute to, and are impacted by, advances in technology that impacts teaching and learning.

Papers will directly focus on data intelligence in education looking at Practice, Research, and Impact.

Sharing Practices of Data Intelligence in Education

- Precise education governing with data intelligence
- Adaptive teaching with data intelligence
- Individualized and adaptive learning with data intelligence

Research on Issues and Challenges of Data Intelligence in Education

- Standard of big data in education
- Data technologies: text, graphic, audio/video, gestures, or sensor technologies
- Data analysis: methods, models, and trends

Envisioning the Impact of Data Intelligence on Education

- Underlying principles of data intelligence in education
- What works with data intelligence in education
- State of the art of data intelligence in education
- Future of data intelligence in education
- Building partnership to increase educational impact

In many way, the development of new technologies affords an opportunity to enhance student learning across the broad spectrum of educational institutions and educational systems. This workshop will provide attendees with a venue to share their expertise and to network with other professionals to find synergies to build the impact of data intelligence on student learning.

Reference

Baker, R. (2011). Data mining for education. In B. McGaw, P. Peterson & E. Baker (Eds.), *International encyclopedia of education (3rd ed.)*. Oxford, UK: Elsevier.

- Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education*, 13(1), 117–140.
- Gibson, D. C., & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies*, 20(4), 697–713.
- Heitink, M., van der Kleij, F., Veldkamp, B., Schildkamp, K., & Kippers, W. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, 17, 50–62.
- Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. John Wiley & Sons
- Pardos, Z. A. (2017). Big data in education and the models that love them, *Current Opinion in Behavioral Sciences*, 18, 107
- Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4), 49–64.
- Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics—A literature review. *ICTACT Journal on Soft Computing*, 5(4), 1,035–1,049.
- Tang, S., Peterson, J., Pardos, Z. (2017): Predictive modelling of student behaviour using granular large-scale action data. In Lang C, Siemens G, Wise AF, Gaevic D. Alberta (eds), *The Handbook of Learning Analytics*, Canada: Society for Learning Analytics Research (SoLAR); 223-233.
- Zheng, L., Shi, R., Wu, B. & Gu, X. (2017). A Robust Approach of Characterizing Teacher's ICT Usage Trajectories. In Proceedings of EDM 17-10th International Conference on Educational Data Mining.
- Zheng, L., Gong, W., & Gu, X. (2017). Predicting e-Textbook Adoption Based on Event Segmentation of Teachers' Usage. In Proceedings of the 7th International Conference on Learning Analytics & Knowledge.
- Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in learning analytics and educational data mining. Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN2015), Bruges, Belgium. 297–306.

A Model of Persona-based Technological Pedagogical Content Design

Xuanxi Li

Faculty of Education, East China Normal University
xxli@dedu.ecnu.edu.cn

Xiaoqing Gu

Faculty of Education, East China Normal University
xqgu@ses.ecnu.edu.cn

ABSTRACT: This study aims to develop personalized teaching application based on learner personas from data acquisition, data classification and analysis, data application and other aspects, which has a certain practical value. On the basis of previous researches on learner personas, in terms of TPACK theory, this study put forward a model of Persona-based Technological Pedagogical Content Design (PTPCD). Based on the model of PTPCD, designers of personal learning environment are suggested to design and recommend appropriate Technological Pedagogical Content to the target learners with matching persona intelligently or semi-intelligently. This study will be further demonstrated and improved in practice.

Keywords: Personal Learning Environment; TPACK; Learner Persona; Learner Model

1 INTRODUCTION

In recent years, the design and development of personal learning environment has become the focus of attention of many scholars all over the world (Mou, 2016; Sun & Tang, 2017). For distance learners, personal online learning platform can support them in accordance with their aptitude, which is conducive to effective learning (Mou, 2016; Xie & Sun, 2012). Personalized education aims to meet individual demands and provide with customized support for different learners (Cristóbal et al., 2009). According to previous study by Xie and Sun (2012), effective personal learning environment should have the following two characteristics:

First, could support personalized learning, help learners build personalized knowledge, and stimulate learning enthusiasm.

Second, could intelligently or semi-intelligently solve the problems that learners encounter in the process of learning. For example, the frustration of distance learners mainly comes from the lack of timely feedback when they encounter problems in learning. Hence, personal learning environment should play its advantages to help learners reduce such unsuccessful learning experience (Xie, & Sun, 2012).

Some scholars have pointed out that the functions of the current distance learning system in China are relatively strong and comprehensive, including a series of functional applications, such as curriculum design and development, communication, collaboration, reflection, evaluation, etc., but these systems are still difficult to provide with real personalized learning support (Xie & Sun, 2012). Even though the system has a large number of rich learning resources, it is still too suffering for learners to find the content they want, which decreased the learners' learning enthusiasm. One of the reasons is that most

distance learning systems either do not have the support of learner personas, or the learner personas cannot evolutionary survival, which makes the system unable to accurately understand the demands of learners (Xie & Sun, 2012; Yue & Chen, 2017).

According to Brooks and Greer (2014), learner persona is narrative descriptions of typical learners that can be identified through centroids of machine learning classification processes.

However, how to build learners' personas and design learning content for each persona in order to develop an effective personal learning environment? In order to address this problem, this study put forward a model of Persona-based Technological Pedagogical Content (PTPC) Design.

2 PERSONAS AND PERSONAL LEARNING ENVIRONMENT

Persona was proposed by Alan Cooper, an American Software Designer. Alan Cooper has been trying to find ways to make software "easy to use", that is, how to make software meet people's needs. In 1995, when Alan participated in the design of a project, six clients were interviewed to find out the differences among their purposes, skills and tasks when using the software. They were divided into three types, each of which was given a name, Chuck, Cynthia and Rob. It was also the first goal-oriented personas (Cooper, 2006).

Brooks and Greer (2014) described a method of transforming the results of predictive analysis into learner personas. They built a predictive model to identify the kinds of attributes to be included in the model as well as the kinds of learners for whom an intervention is to be created. Personas were separated into two groups: affirmation personas, which describe a target group of learners they were interest in having interventions designed for weak performers and erroneous personas, which describe classifications that the predictive model made that they were not focused on supporting (strong performers). They use Waikato Environment for Knowledge Analysis to build predictive models. However, they did not describe the classification of indicators of predictive model in details. Besides, they did not indicate how to include learner personas into a personal Learning environment as the intervention. Zhang, Zhao, and Guo (2016) built learner personas for college students based on the data of "Tracking Researches on Learning and Development of Chinese College Students" in 2015. Their study analyzed students' family characteristics, pre-university learning characteristics and learning behavior characteristics. The learning behavior indicators include learning motivation, learning strategies, multiple learning styles, core learning behaviors, time allocation, learning outcomes and so on. However, the study did not classify different personas. In addition, the indicators of learning behavior in their study did not have a foundation of theoretical framework.

Some Chinese scholars use term "learner model" or "student model" instead of "persona". For example, Yang et al. points out that learner models are abstract descriptions and representations of learner information (Yang, Wang, & Feng, 2005). Zhuang and He (2015) believe that learner models are the core and key parts of personalized recommendation system for personal learning, which depict relationships. It is the basis of implementing personalized learning recommendation to provide the basis for accurately and appropriately pushing learning resources, learning paths and learning activities for learners.

Huang and Liu (2018) designed student model based on massive open online courses. In their study, first, they collected learner characteristics according to the learning behavior data on MOOC platform; second, a knowledge tracing model based on Bayesian network was constructed. Model parameters were set based on their empirical probabilities and the difficulty of a problem was introduced into the

model. Third, the learners' attitude and enthusiasm were analyzed, and a learning attitude tracing model was designed based on classification algorithms. Based on the experiments conducted on a dataset of MOOC, the predictions of different Bayesian knowledge tracing models were compared. Since Logistic regression algorithm was used as the classifier of students' attitude and enthusiasm, and then the learning attitude could be predicted accurately. The experimental results show that the MOOC learner model is capable of analyzing knowledge and attitude of students. Yue and Chen (2017) proposed to build learner models which is the same meaning to learner personas from four aspects: personal information, learning style, interest model and knowledge model, and further proposed the labels and measurement methods of personal information, learning style, knowledge model and interest model (Figure 1).

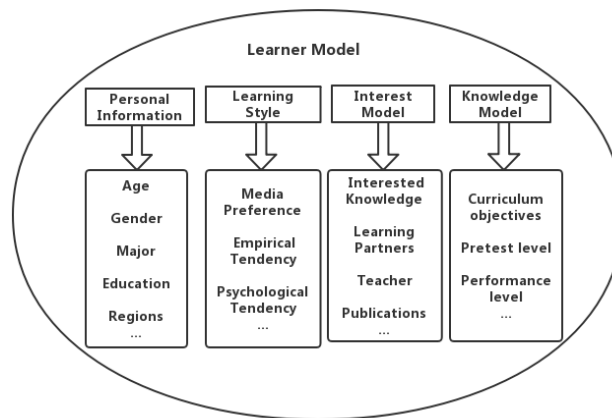


Figure1. Learner Model (Yue & Chen, 2017)

Besides, they put forward a model of Personalized Learning Path Recommendation (PLPR). As shown in Figure 2, the personal learning system obtains learner's learning starting point through diagnosis test, determines learning interest and learning demand through initial interest survey and dynamic analysis of learning process data, determines the next learning content through knowledge performance detection, and combines learner's learning style characteristics (such as sensory preference and collaboration, interactive psychological tendencies, learning tools tendencies, etc.), to customize personalized learning content and learning activities for learners. The learners enter the optimal learning process according to the recommended learning path (Yue & Chen, 2017).

However, the application of learner models (personas) in personalized learning path recommendation is still at the theoretical level (Yue & Chen, 2017), and the model of PLPR put forward by Yue and Chen (2017) is very complicated and sophisticated, it did not illustrate and emphasize the relationship among learning content, resource presentation/learning tools, and learning style. Hence, the model of PLPR is not very easy to be understood and worthy to be further improved, which can be further enriched by using the theory of TPACK.

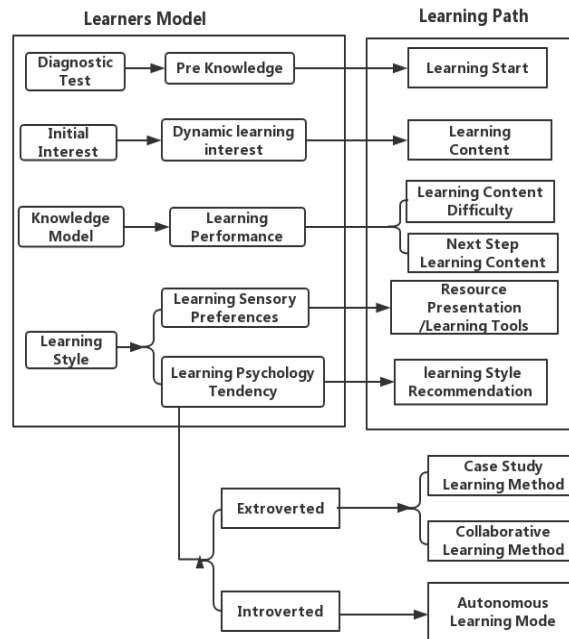


Figure 2. A model of Personalized Learning Path Recommendation (Yue & Chen, 2017)

3 TPACK AND PERSONAL LEARNING ENVIRONMENT

In 2005, Koehler and Mashra of the University of Michigan put forward the theoretical model of TPACK, which integrates educational technology, subject content and general pedagogy. TPACK model contains three core elements, namely subject content knowledge (CK), pedagogical knowledge (PK) and technical knowledge (TK); four complex elements, namely pedagogical content knowledge (PCK), technological content knowledge (TCK), technological pedagogical knowledge (TPK), technological pedagogical and content knowledge (TPACK) (Figure 3) (Mishra & Koehler, 2006). TPACK is different from knowledge of subject experts or educational technology experts, and also different from general pedagogical knowledge that can be used in various disciplines. TPACK is about how to use educational technology to effectively support pedagogy and present content in order to facilitate teaching and learning. Based on TPACK framework, in order to use technology effectively in teaching, teachers must understand the relationship among technology, content and specific teaching methods. Teachers should understand how to use technology to support specific content learning and how specific teaching methods best support the use of technology and promote learning (Padmavathi, 2017).

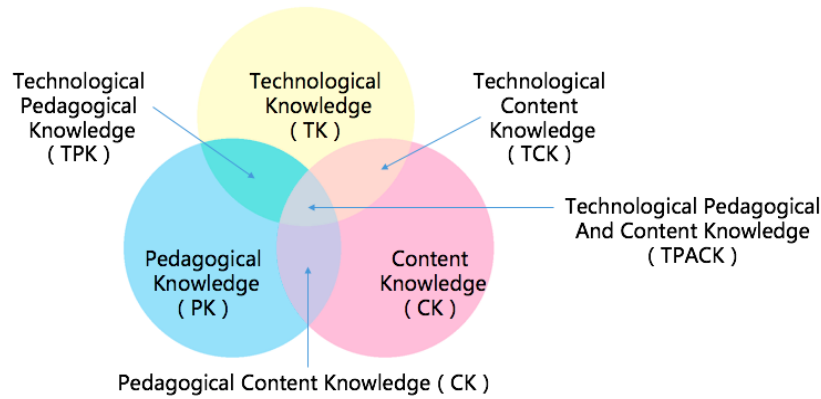


Figure 3. TPACK Framework (Mishra & Koehler, 2006)

Scholars have carried out a lot of theoretical and practical researches on TPACK (Li, 2017). For instance, Li put forward an Advanced Wiki-based Collaborative Process Writing Pedagogy (AWCPWP) to assist kids to write Chinese compositions, which is a paradigm of putting TPACK into action (Mishra & Koehler, 2006).

According to Zhang, TPACK distills a new form of knowledge within the intersection of technology, pedagogy and subject content, which represents a concept of standardizing the application of ICT in education. Therefore, the concept of TPACK is consistent with the requirement of the personal learning environment, which aims to provide with different students with effective learning with appropriately support of technology and pedagogy (Zhang, 2015). Zhang constructed flip classroom teaching mode based on TPACK, and analyzed the path of realizing flip classroom teaching activities. He pointed out, due to the wide application of micro-video technology, the prevalence of flipping classroom teaching mode, and the guidance of TPACK theory, the reform of personalized learning is greatly promoted (Zhang, 2015).

Based on the above statement, TPACK integrates and connects various elements, and establishes a teaching mode combining technology dimension, pedagogy dimension and subject content dimension. Hence, the theory of TPACK can match with learner personas, and together be used to support the design and development of a personal learning environment.

4 PERSONA-BASED TECHNOLOGICAL PEDAGOGICAL CONTENT DESIGN

According to theory of TPACK and previous researches on learner personas, this study proposes a Model of Persona-based Technological Pedagogical Content Design (PTPCD).

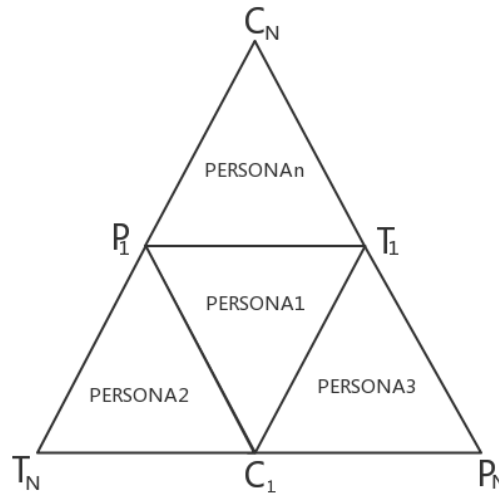


Figure 4. Model of Persona-based Technological Pedagogical Content Design (PTPCD)

As shown in the figure 4, C_1 and C_n represents different learning content, T_1 and T_n represents different educational technologies, P_1 and P_n represents different pedagogies. While the PTPCD has three suggestions for designers:

First, designers are suggested to have the related technological pedagogical and content knowledge (TPACK). They are suggested to be familiar with subject contents, and understand the relationship among content, pedagogy, and technology. In other words, they had better know how to design or select appropriate pedagogies and technologies in order to support different contents, which at last lead to effective learning.

Second, designers are suggested to involve learner personas into the design of personal learning environment. Students can be classified into different personas, for example, persona1, persona2, personaN (Figure 4). Designers can select and decide indicators for learner personas based on different situations. For example, the learner models (personas) composed with indicators of personal information, learning style, knowledge model and interest model, put forward by Yue and Chen (2017), can be used in many cases (Figure 1).

Third, designers are suggested to recommend appropriate technological pedagogical content to the target learners with matching persona intelligently or semi-intelligently, for both customization and personalization of the personal learning platform.

Above all, based on PTPCD, in terms of the same learning content, different learning models with the support of different pedagogies and technologies are assigned to different learner personas (e.g. PERSONA1, PERSONA2, PERSONA_n).

5 DISCUSSION AND CONCLUSION

The advent of E-learning big data era and the development of learning analysis technology have promoted the deepening of personalized service in distance education. The design and development of learner personas and subject contents are the core of large data analysis and personalized teaching.

On the basis of comprehensive analysis of previous learner model and learning paths (Yue & Chen, 2017), in terms of TPACK theory (Huang & Liu, 2018), this study put forward a model of Persona-

based Technological Pedagogical Content Design (PTPCD). Based on the model of PTPCD, designers are suggested to design and recommend appropriate Technological Pedagogical Content to the target learners with matching persona intelligently or semi-intelligently. In other words, when a designer considers how to plan for an instructional activity, it is typically conceptualized around content goals and organized according to learning activities. The application of a technology (integration) should be done so with the intent of enhancing the learning experience through innovative practices. It is also important to build learner personas in terms of different situations, and the learner model put forward by Yue and Chen (2017) can be considered by designers (Figure 1), from which they can choose and decide the indicators of individual information, learning style, learning interest and knowledge model. The learner model put forward by Huang and Liu (2018) can also be considered by designers, which focused on analyzing knowledge and attitude of students.

The focus of this study is to develop personalized teaching application based on personalized learning characteristics from data acquisition, data classification and analysis, data application and other aspects, which has a certain practical value. Of course, this study also needs further demonstration and improvement in practice, specially, apply intelligent algorithms to recommend the appropriate technological pedagogical content to the target learners with matching persona for personalized learning.

ACKNOWLEDGEMENTS

The authors want to thank Dr. Xiao, Jun, the Director of Technical Research, Engineering Technology Research Center of Shanghai Open Distance Education in Open University of Shanghai, for his advice and support on this study.

REFERENCES

- Mou, Z. J. (2016). Solution of Individualized Learning Path Supported by Learner's Data Portrait: Value Enrichment of Learning Computing. *Journal of Distance Education*, 34(6):11-19.
- Sun, Y. H., & Tang, Z. W. (2017). Research on the Construction of Personalized Learning Environment based on Big Data. *Primary and Middle School Audiovisual Education (z2)* :51-54.
- Xie, M. F., & Sun, X. (2012). Research on the Model of Distance Personalized Network Learning System based on Ontology Knowledge Management. *China Educational Technology*, 310, 47-53.
- Cristobal, R., Sebastian, V., Amelia, Z., & Paul, D. (2009). Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. *Computers & Education*, 53(3), 828–840.
- Yue, J. F., & Chen, Y. (2017). Distance learner modeling and personalized learning application based on large data analysis. *Distance Education in China*, 2017(7), 34-39.
- Cooper, A. (2006). *The way of interaction design*[M]. Beijing: Electronic Industry Press.
- Brooks, C. and J. Greer (2014). Explaining predictive models to learning specialists using personas. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, Indianapolis, Indiana, USA.
- Zhang, H., Zhao, L., and Guo, F. (2016) A portrait of first-generation college students in China: an analysis based on the China College Student Survey. *Tsinghua Journal of Education*, 37(6), 72-94.
- Yang, H., Wang, L., & Feng, H. (2005). Research on Two-tier Dynamic Student Model in Intelligent Teaching System. *E-education Research*, 1, 72-75.

- Zhuang, K. J., & He, B. X. (2015). Research on learner model of personalized recommendation system for web-based learning. *China's Educational Technology Equipment*, 372, 67-69.
- Huang D., Liu, X. (2018) . Design and realization of student model based on massive open online courses. *Journal of Computer Applications*, 38 (S2): 327 – 330.
- Mishra, P. & Koehler, M. J. (2006). “Technological pedagogical content knowledge: A framework for teacher knowledge,” *Teachers College Record*, vol. 108, no. 6, pp. 1017-1054.
- Padmavathi, M. (2017). “Preparing teachers for technology-based teaching-learning using TPACK,” *I-Manager's Journal on School Educational Technology*, vol. 12, no. 3, 2017.
- Li, X. (2017) . Putting Technological, Pedagogical, and Content Knowledge (TPACK) in Action - An Advanced Wiki-based Collaborative Process Writing Pedagogy (AWCPWP). *International Journal of Culture and History* 3 (4): 243-249.
- Zhang, S. B. (2015). Construction and Implementation of Flipped Classroom Teaching Model based on TPACK. *China Education Info*, (12), 37-41.

CS-BKT: Introducing Item Relationship to the Bayesian Knowledge Tracing Model

Lingling Meng

Department of Education Information
Technology, Faculty of Education, East China
Normal University
llmeng@deit.ecnu.edu.cn

Wanxue Zhang

Department of Education Information
Technology, Faculty of Education, East China
Normal University
Zhangwanxue0703@163.com

Mingxin Zhang

Department of Education Information
Technology, Faculty of Education, East China
Normal University
2374624930@qq.com

Yu Chu

Department of Education Information
Technology, Faculty of Education, East China
Normal University
1476727343@qq.com

ABSTRACT: Bayesian knowledge tracing model is a typical student knowledge assessment method. It is widely used in the intelligent tutoring systems. In the standard BKT model, all knowledge and skills are independent of each other. However, in the process of student learning, they have very close relation. A student may understand knowledge B better when he masters knowledge A. Therefore, this work introduces a new student model based on BKT. It takes the relationship between knowledge into account. By doing this, the new model proves higher prediction accuracy and performs better.

Keywords: Knowledge tracing. Student modeling. Cross skill

1 INTRODUCTION

With the continuous development of intelligent assistant systems, the evaluation of student learning effects has become increasingly important. It can accurately reflect a series of personalized data such as the learner's learning level and knowledge status. According to the evaluation results, the intelligent systems can determine what the learner needs to practice and chooses the appropriate teaching strategy.

Bayesian knowledge tracing model is a typical student knowledge assessment method. It was first proposed by Corbett and Anderson in 1995(Corbett, A. T., 1995). The model assumes that student knowledge can be represented as a set of binary variables, the learned state and the unlearned state. The usual way is to ask students to answer a series of questions about the knowledge. According to this principle, the model believes that the status of each knowledge point can be inferred from a set of corresponding training results, and the training result is also a set of binary variables, that is, students answer questions correctly or incorrectly. Modeling student knowledge state as a latent variable is one of the more commonly used methods in current research fields. As a potential variable, the student's knowledge status is updated according to the observable variables such as the student's answer to the question. This method is a typical application of hidden Markov model.

2 RELATED WORK

After the BKT model was proposed, the researchers continued to modify the model, hoping to improve the prediction accuracy. Baker and Corbett et al. determined model parameters based on student responses, which can avoid different predictions for students with the same answer. In 2010, Pardos and Heffernan proposed a prior per student model based on individual differences among students (Pardos, Z. A., 2010). They chose three approaches to set the individualized initial knowledge values and analyzed the results. The new model had better performance than traditional BKT. In 2011, they also proposed knowledge tracing item difficulty effect model (Pardos, Z. A., 2011). The model added an extra node to represent item difficulty.

Yudelso et al individualized BKT with student-specific parameters (Yudelso, M. V., 2013). They tested different model variants on different dataset-skill model combinations. The study found adding personalized learning parameter had better prediction accuracy. Mohammad and Rowan suggested a new model named LFKT by integrating potential factor models and standard BKT models (Khajah, M. M., 2014). LFKT calculated the guess and slip probabilities with a new way.

In 2016, Chen Lin and Min Chi proposed an intervention-Bayesian knowledge tracking model by adding a teaching intervention node (Lin, C., 2016). They mainly focused on guiding and explaining the impact of two kinds of teaching interventions on students. The results show that the intervention-bayesian knowledge tracing model has a significant improvement compared with the standard BKT model. Kai Zhang and Yiyu Yao divided the learning process into three learning states (Zhang, K., 2018). They believed there was a learning state between unlearned state and learned state. The new TLS-BKT model improved the prediction accuracy and showed superior robustness.

3 METHODOLOGY

3.1 Bayesian Knowledge Tracing Model

Bayesian knowledge tracing is a very popular model to infer student knowledge. There are four parameters, $P(L_0)$, $P(T)$, $P(G)$, and $P(S)$ in BKT. $P(L_0)$ and $P(T)$ are learning parameters, which are mainly used to indicate the knowledge status of students' learning. $P(L_0)$ represents the initial level of a student's knowledge. When $P(L_0)$ is 0, it means that the student does not know the required knowledge before answering the question. While $P(L_0)$ is 1, the student has mastered the required skill. $P(T)$ is the probability of transition from unlearned state to learned state after studying. $P(G)$ and $P(S)$ are performance parameters. $P(G)$ indicates the probability of answering questions correctly when student does not understand the required knowledge. $P(S)$ indicates the probability of answering questions by mistake if student already knows the required knowledge. When both $P(G)$ and $P(S)$ are zero, the student's answer can fully reflect the knowledge mastery. According to the definition of the parameters, we can get the probability distribution table, as shown in table1.

Table 1: Probability distribution table

Learning State	Wrong answer	Right answer
unlearned	$1-P(G)$	$P(G)$
learned	$P(S)$	$1-P(S)$

The probability of a student giving a right or wrong answer is computed either using the Equation (1a) or Equation (1b). Equation (1c) is used to update the knowledge mastery of student. To compute the probability of student answering the next question correctly, we can use Equation (1d).

$$P(\text{Correct}_n) = p(L_n)(1-p(S)) + (1-p(L_n))p(G) \quad (1a)$$

$$P(\text{Incorrect}_n) = p(L_n)p(S) + (1-p(L_n))(1-p(G)) \quad (1b)$$

$$P(L_n) = p(L_{n-1} | \text{Evidence}_{n-1}) + (1-p(L_{n-1} | \text{Evidence}_{n-1}))p(T) \quad (1c)$$

$$P(C_{n+1}) = p(L)(1-p(S)) + (1-p(L))p(G) \quad (1d)$$

3.2 Cross-Skill Bayesian Knowledge Tracing Model

In the BKT model, all knowledge and skills are independent of each other. However, in the process of student learning, they have very close relation. So we propose a new model called Cross-skill Bayesian knowledge tracing (CS-BKT). This model takes into account the relationship between different knowledge nodes. When students deepen their understanding of a certain skill A, they will also deepen their understanding of skill B to some extent.

Therefore, the model introduces a new parametric matrix, referencing Figure1. In the skill relationship matrix, rows and columns represent different skills or knowledge, and the values of row and column intersections are the probability that one skill affects another skill.

		knowledge		
knowledge		i	j	k
	i	p ₁	p ₂	p ₃
	j	p ₄	p ₅	p ₆
	k	p ₇	p ₈	p ₉

Figure 1: parametric matrix

The new method to calculate the final knowledge mastery of student on one skill should add the influence probability of other knowledge. We can use the Equation as follows. For example, student u answers two questions about skill k. Using the standard BKT model, we can know the mastery of student u on skill k after answering the first question and the second question separately. The difference of this two values reflects the change of student u on skill k. So the product of the probability in skill relationship matrix and the difference of knowledge mastery is the final influence probability that skill k affects skill i.

$$\hat{p}(L_{t+1})_u^k = p(L_{t+1} | obs)_u^k + (1-p(L_{t+1} | obs)_u^k) \cdot p(T)^k \quad (1e)$$

$$\Delta p(L_{t+1})_u^k = \hat{p}(L_{t+1})_u^k - p(L_t)_u^k \quad (1f)$$

$$p(L_{t+1})_u = p(L_t)_u + R_k \cdot \Delta p(L_{t+1})_u^k \quad (1g)$$

Here is the structure of CS-BKT model, referencing Figure2. The circles labeled by k stand for learning states. The circles labeled by o stand for answering results. The parameter $p(L)$ will be continuously updated, as the answer changes. R_{kl} is the influence of knowledge k to knowledge l.

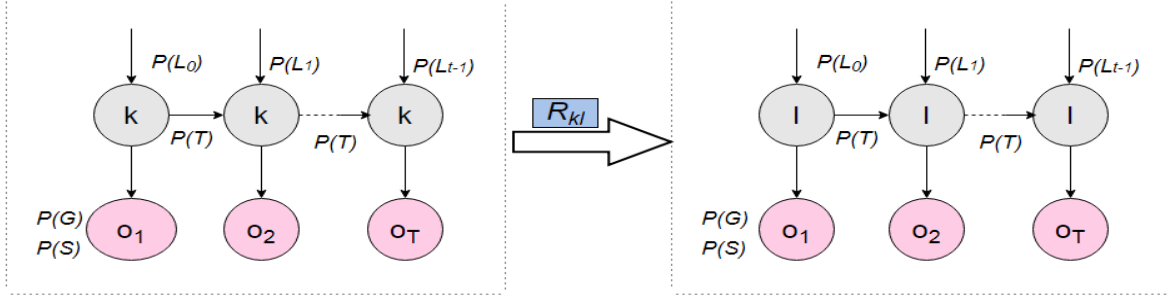


Figure 2: CS-BKT model

4 EXPERIMENT

The data set for this study was from an online education system at Bridge to Algebra, one of the competition data for the 2010 KDD Cup. We selected four commonly used model evaluation indicators to evaluate the CS-BKT model, which are accuracy, AUC, root mean square error (RMSE) and loss rate.

The training results are shown in the figures below. The orange lines represent CS-BKT, and the blue lines represents BKT model. We can know the accuracy and AUC of CS-BKT are higher than BKT while the RMSE and loss rate are lower. So the new model has better performance than the standard model.



Figure 3: The result of accuracy

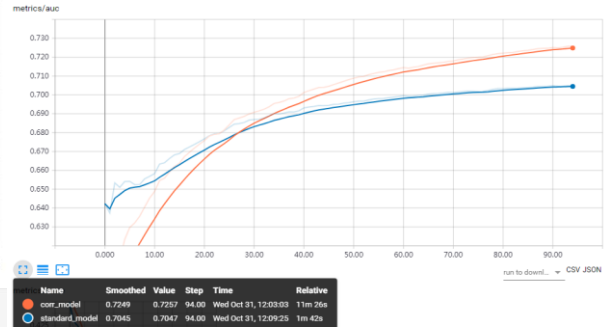


Figure 4: The result of AUC

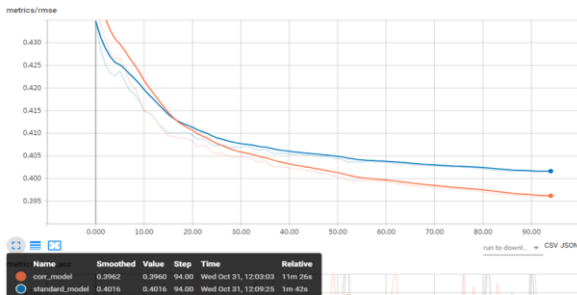


Figure 5: The result of RMSE

Figure 6: The result of loss

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a new model called CS-BKT. Considering the relationships between knowledge, we introduced a new matrix. Each knowledge has different influence probabilities on other knowledge. After using the KDD data set, we proved the new model has higher prediction accuracy and performs better.

One direction for future work about the model would be to exploration for more accurate probability matrix to show the relationship of different skills or knowledge. A good way to get the matrix is to ask several experts to give the answers they think and then take the average. Or we can tell the experts to discuss the influences of knowledge and then give the final result.

Another area of interest would be to personalize the matrix. As we know, every student has his own learning level, so the influence of knowledge is also different. Researchers should give unique matrixes for different students so that they can predict student's performances more accurately.

REFERENCES

- Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. *Lecture Notes in Computer Science*, 5091, 406-415.
- Corbett, A. T., & Anderson, J. R.. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253-278.
- Khajah, M. M. , Wing, R. M. , Lindsey, R. V. , & Mozer, M. C. . (2014). Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. *Journal of Nursing Education*, 39(9), 409-11.
- Lin, C., & Chi, M. (2016). Intervention-BKT: Incorporating Instructional Interventions into Bayesian Knowledge Tracing. *Intelligent Tutoring Systems*.
- Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. *International Conference on User Modeling*.
- Pardos, Z. A. , & Heffernan, N. T. . (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. *User Modeling, Adaption and Personalization*. Springer Berlin Heidelberg.
- Yudelso, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. *International Conference on Artificial Intelligence in Education*.

Classification of At-Risk Students Using Student Interaction Information in Learning Management Systems (LMS)

Zhouxiang Cai
George Mason University
zca2@gmu.edu

Zhiyun Ren
George Mason University
zren4@gmu.edu

Huzefa Rangwala
George Mason University
rangwala@cs.gmu.edu

ABSTRACT: Distance education offerings in the form of massive open online courses (MOOCs) and traditional universities have seen a surge in enrollments from students across the world. While enrollments are large it is not clear if these online offerings are able to achieve the desired learning outcomes as in the case of in-class face-to-face learning. Learning management systems (LMS) aid both online and in-class course offerings by providing content and collaborative tools between students and instructors. Learning analytics seeks to analyze the data extracted from LMS server logs to identify student learning behaviors and engagement characteristics to further help the students in achieving academic success. Analysis of these datasets can also help the instructor in designing improved course content, identifying common challenges across students and improve overall pedagogy. The objective of this study is to develop and assess machine learning methods that use features extracted from LMS server logs to perform early and real-time prediction of student performance within a course. Leveraging data across multiple courses taken by a given student, the engineered features capture student interactions and course characteristics. We performed a comprehensive evaluation using the de-identified data obtained from Canvas Network open courses. Our experimental results show that we can predict the student final learning outcomes with high accuracy.

Keywords: Early Warning, Learning Analytics, Regression, Classification, Student Behavior

1 INTRODUCTION

With the advancement in learning technologies, education institutions increasingly rely on the online sources for delivering educational content and achieving learning outcomes (Na & Tasir, 2017). Online or distance education can be synchronous i.e., in conjunction with a brick-and-mortar class happening at the same time, asynchronous or hybrid where the online material supplements traditional in-class material. Massive Open Online Courses (MOOCs) since their inception have promised the opportunity of delivering low-cost (free) educational resources to thousands of students across the world (Ren et al., 2016).

Both, brick-and-mortar educational institutions and MOOCs use learning management systems (LMS) or course management systems. Prime examples include Blackboard (blackboard.com), Canvas (canvas.net) and Moodle (<https://moodle.org>) for online access to course content. These systems allow for collaboration and communication amongst the different stakeholders within a course: (i) instructors, (ii) students and (iii) teaching assistants. The server logs serve as a source of student-interaction data with the LMS that can be used to identify student engagement and learning behaviors for a given course. Learning analytics researchers have developed several different approaches to analyze this interaction

data for several purposes: (i) improving content and learning outcomes by identifying for the instructor, course content that several students face difficulty in mastering, (ii) predicting students' future academic performance to facilitate better degree planning and advising (iii) early identification of students who may be at risk of failing a class and would benefit from attention/intervention by course staff and (iv) identifying successful pedagogical approaches that helps students learn better.

Learning management systems utilized for MOOCs provide students and instructors with a collaborative way of overcoming the limitations of traditional classroom space while also saving time. Using server logs, various student engagement and interaction features can be derived. Examples include the amount of time required for studying individual chapters, completing quizzes and wrapping-up assignments (Ren et al., 2016). By evaluating these student interactions as well as the course information, learning analytics can identify patterns associated with student learning. Instructors, as well as other stakeholders provided with access to these data analytics approaches can identify if students are achieving the class learning goals in a timely manner and provide interventions and personalized feedback (Kotsiantis et al., 2014).

In this paper, we implement machine learning methods to identify students who are at risk of falling behind in a timely fashion. We simulate two real-world scenarios of students enrolled within MOOCs. Specifically, we name these approaches as **Student-Specific** and **Course-Specific**. We seek to perform the task of in-class prediction i.e., using interaction data extracted from LMS server logs to predict the final grade for a student and identify students who are at risk of failing a course. Another key objective of our proposed methods is to identify these students earlier in the semester (also used interchangeably with the term). We evaluated the proposed methods on a Canvas that is comprised of de-identified data from 376 Canvas Network open courses which are also MOOC offerings. Our results highlighted the strengths of the proposed approaches in predicting students who are at at-risk of not passing the class using features derived from LMS data.

2 LITERATURE REVIEW

In order to improve the retention, several researchers have focused on the analysis and prediction of student's performance based on student's past learning related habits and aptitudes. Romero et al. (Romero et al., 2008) evaluated various data mining techniques to classify students as high and low performers based on their LMS usage data. Ren et al. (Ren et al., 2016) developed a multi-regression based model to predict the performance of a student per assessment (HW) based on student interaction data for several MOOCs. Devasia et al. (Devasia et al., 2016) predicted students' performance best by analyzing student social features like that of gender and lifestyle habits. Besides in-class performance prediction, an understanding of suitable approaches and theories of learning analytics is also required for further examination of learning behavior (Lave et al., 1991). Pittman (Pittman, 2008) compared data mining techniques used to predict student retention and found that logistic regression was the most suitable. Boroujeni and Dillenbourg (Boroujeni & Dillenbourg, 2018) discovered some common study patterns based on the MOOC interaction sequence and found that these study pattern transitions probabilities correlated with different learners. Zhang and Rangwala (Zhang & Rangwala, 2018) developed an Iterative Logistic Regression method to address the challenge of early predictions and got a much more precise answer than results obtained from standard logistic regression.

In this paper, we study the application of machine learning as it relates to early in-class student grade prediction. Similar performance prediction techniques have been explored in different settings. Several prior studies (Ren et al., 2016) use MOOC server logs to predict homework grades or dropouts. He et

al. (He et al., 2015) investigated the early warning signs of students at risk of failing a MOOC by evaluating multiple offerings under potentially non-stationary data. They built predictive models weekly based on the numerous offerings of a course. Jokhan et al. (Jokhan et al., 2018) designed an early warning system based on the students' features such as gender, age, social status and engagement features to achieve a 60.8% accuracy based on that particular model. Due to the absence of data from previous classes, Hlosta et al. (Hlosta et al., 2017) developed a 'self-learner' method which used current course data as the training set to identify the at-risk students. Le et al. (Le et al., 2018) applied the state-of-the-art in recurrent neural network classification to predict students learning status based on 20 MOOC courses. Whitehill et al. (Whitehill et al., 2015) designed a MOOCs stopout detector and used the stopout detector to conduct an intervention.

Contributions: (1) We seek to identify students' at-risk of failing a course by using LMS-derived features within standard machine learning models. Most prior research on identify at-risk student focuses on training for a small MOOCs database (usually no more than 20). In order to get a generalizable finding, we stem from benchmarking and leveraging data across more than 300 courses (rather than a single course) for a given student and focusing on the early identification of at-risk students. (2) We simulate two real-world scenarios for in-class final performance prediction, first centered around students enrolled in multiple MOOCs and second involved developing course-specific models that assume multiple offerings for a given course across different terms.

3 PROBLEM DEFINITION

Given a database about the interaction of students with a learning management system for a given course, the objective of this study is to develop classification methods to identify students (early on) who will perform well in a target/current course. The set of interaction features capturing student engagement and learning habits extracted from the LMS is denoted by \mathbf{F}_i^j for the i -th student and j -th course. Formally, the objective of the classifier is to learn a mapping function $f: \mathbf{F} \rightarrow \{0, 1\}$ that takes as input the feature from the current class \mathbf{F}_i^c and output 0 (representing passing a course) and 1 (representing failing a course). Additionally, the proposed algorithms seek to make these performance predictions early on to assist the student (who are at risk of failing) do better. As such, we assess the performance of the proposed algorithms by using features extracted from the first few days or weeks of the course. We encode this by extracting features only from the first 10%, 20%, 30% and 40% of the course during training and testing. Figure 1 shows the student interaction data for a typical student. We can view students' various activities at different timestamps within the Figure. The Y-axis shows the number of requests made per day by the student. The dots along the time-series indicate specific course-related events i.e., submission of quizzes or assignments made by this student. The percentage value indicates the score earned by the student on the particular graded activity. Along the top, we highlight the feature extraction from the start of the semester based on the amount of time we want to consider. For the given course we show in Figure 1 X set to 0.1 indicates the first 10% features of the class $C1$ will be used. In our study, we set X to a 10%, 20%, 30% and 40% to catch student features towards the beginning of the course, as we defined as **Early Stage Feature**.

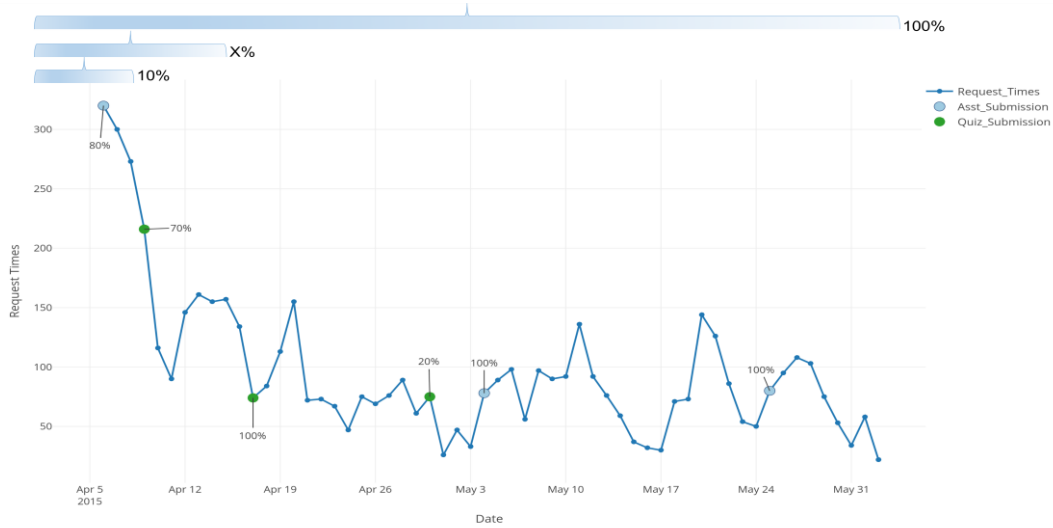


Figure 1: A sample student engagement time-series data

We simulate two common real-world scenarios centered around a student and course, described in detail below.

3.1 Student-Specific Approach

In the *Student-Specific* approach, we simulate the real-world scenario of students enrolling in multiple courses over time. Each enrollment record associated with a student-course pair is stored in a database and we call it *Student-Courses records*. We seek to predict the performance of a student within a given class based on performance/interaction within prior Student-Course records. Specifically, we are predicting the final grade of a student in a current or active class using interaction data from the first few weeks of the current/target class.

The graphical representation of this approach shown in Figure 2. We divide the Student-Course dataset into a training set and a testing set. The training set is regarded as the set of courses completed in the past and the test set is the set of active/current on-going courses. We split the data such that training set accounts for 90% of the dataset. We also seek to understand the relationship between prediction accuracy and amount of data in terms of time/weeks needed for deriving the features and hence the predictions. For the test set we predict within the first few weeks by setting the parameter X in Figure 2 to smaller values such as 10%, 20%, 30%, 40% to evaluate this early warning approach. We combine both student- and course-related features as described below for predicting the final grade. Lastly, we input both training and testing data into the three machine learning methods.

3.2 Course-Specific Approach

Educational institutions usually offer the same course across different semesters. We also consider an alternate way of identifying possible at-risk students in a course by comparing student's performance in previously offered course (completed course). The graphical representation of the *Course-Specific* approach is shown in Figure 3. Our data only has the course discipline information, and we cannot identify the previously offered courses in our dataset. We simulate this by sampling the training and testing data from the same course. We use 10% of students as a testing set presented as $\{F_1^c, \dots, F_m^c\}$ in Figure 3. We assume this 10% of students would take this course next semester. The remaining students are training set and presented as $\{F_1^p, \dots, F_n^p\}$. For the Course-Specific Approach, we only use the student features because the course features (like CourseLen) would remain constant for the same class. Unlike

the student-specific approach we train multiple course-specific models. We applied this approach to a total of 107 courses and averaged the accuracy along with the final scores for each experiment. The features extracted are similar for the student-specific and course-specific approach.

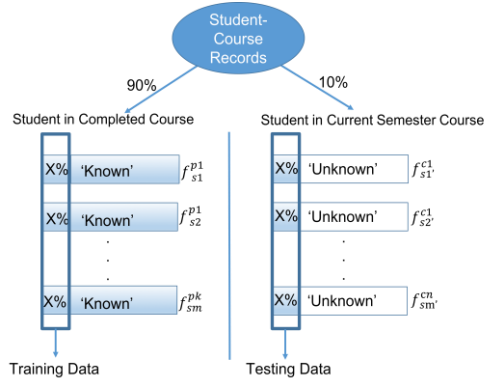


Figure 2: Student-Specific Approach

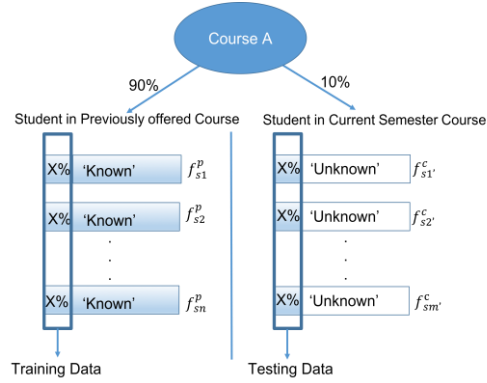


Figure 3: Course-Specific Approach

4 FEATURE DESCRIPTION

Logistic Regression (LR), Random Forest (RF) and K-nearest Neighbors Algorithm (Knn) are used in student performance analysis. To find the best possible learning related patterns for each student, we extract 13 features grouped into the following four categories: (i) course feature, (ii) quiz feature, (iii) assignment feature, (iv) access times feature. Description of each of extracted features are as follows:

(i) **Course Features:** These features capture course statistics including discipline:

- **CourseLen:** denotes total consecutive time duration of all course meetings. To define an early in-class feature, an experimental course must have clear start and end date. This feature may be a good predictor if students adapt to a specific length of a particular course over another.
- **Type:** There were 12 different discipline courses in the dataset used in this study. We include this feature to capture a student's interest/aptitude within a specific discipline over another.
- **Size:** denotes the number of students enrolled for a given course. This feature will be a good predictor of if students tend to concentrate better within a smaller-sized class over one that is larger and more densely-packed.

(ii) **Quiz Features:** These features seek to capture performance of students based on quiz submissions and trials:

- **#Q:** denotes total number of quizzes offered over one course. In our database, there are some quizzes and assignment with 0 possible_points. We did not include practice quizzes having a raw score of 0 possible_points because practice quizzes bear no effect on course grades.
- **QSubmission:** is the number of quiz submissions made by a student before a given cut-off period. We normalize this feature value by comparing it to average submission of the class. QSubmission is an important engagement feature in course management system and could be a strong predictor. A passing student completes most quizzes and assignments on time.
- **QScore:** sums the raw scores of all quizzes taken by a student and is calculated from each quiz submission and then normalized by comparing to the average quiz score of that of the entire class. QScore is one of the most important feature aspect of grade prediction.

- **QAttempt**: is the average number of attempts of quiz submissions made by one student. Certain quizzes in our database allowed for multiple submissions and LMS retains the highest score. We believe this feature to be a success indicator since the more attempts made by a student indicate the willingness to learn and work hard.
- **QTime**: is the average time a quiz has remained opened before submission by one student. Based on the student, a longer quiz duration may indicate a student rechecking final answers before the final submission.

(iii) **Assignment Features**: These features seek to capture performance of student on graded assignments.

- **#A**: is the total number of assignments pertaining to one course. Assignments having a raw score of 0 possible_points were not counted as they bear no relevance to final student grade.
- **ASubmission**: is the amount of assignment submissions for each student over a specific time duration. The intuition for choosing this feature is identical to QSubmission.
- **AScore**: is the total of assignment scores normalized by class average.

(iv) **Request Features**: Request features seek to capture student engagement.

- **AvgLoginHour**: is the average number of hours logged by a student per day over the entire course period. To filter useless requests, we set the evaluate scale to one hour. As long as a student requests LMS in one hour that hour is considered to be a "Working Hour." The working hour rate and CourseLen can further display a student's engagement characteristics. The formula for this feature is given by:

$$\text{AvgLoginHour} = \text{WorkingHour} / \text{CourseLen}$$

- **AvgLoginDay**: captures the fraction of 24-hour cycles where the student has a request from the LMS over the entire course period. As long as a student make requests in a single day, we consider that day to be a "Working Day." The rate of the working day and CourseLen can demonstrate a student's engagement characteristics. We can view student engagement features from a various view using different evaluation scales (Hour, Day). The formula for this feature is shown below:

$$\text{AvgLoginDay} = \text{WorkingDay} / \text{CourseLen}$$

Final Student Performance

We discretized the final passing grades using a binary output. If the student's grade was greater than the average score of the class or the student's final grade was greater than 60, we assume this student to receive a passing grade for the course. We also consider CourseLen, Type, Size, #Q and #A as Course features. These are identical for every student within a class. The rest of the features are considered as Student features and are unique for every student and are associated with a time-stamp parameter as described previously. We also combine the proposed student-features and course-features in our study. The course features though unique for every student within a course differ as the student takes different courses and seek to capture patterns about course that correlate with student performance.

Table 1: Course Principle Distribution

Discipline	#Courses
Business and Management	47
Computer Science	12
Education	85
Humanities	46

Interdisciplinary	22
Life Sciences	5
Mathematics, Statistics	18
Medical Pre-Medical	12
Other or Interdisciplinary	6
Physical Sciences	7
Professions & Applied Sciences	104
Social Sciences	12

5 EXPERIMENTS

5.1 Datasets

We performed empirical evaluation on dataset obtained from the Canvas Network Person-Course De-identified Open Dataset from 1/2014 to 9/2015 (<https://dataverse.harvard.edu/dataverse/cn>). This dataset consists of 376 courses across different disciplines listed in Table 1. To evaluate our proposed approaches, we sample the courses using the following three criteria: (i) Courses should have a start and end date because we can only identify the early stage of a course if the course has a distinct period. (ii) A suitable course should have a meaningful grade distribution. We filter out courses which do not report a final grade i.e., all the entries are 'N' or 'O' in the final grade. (iii) The server logs have student request logs. Several of the courses within the Canvas dataset does not have any information pertaining to student enrolled within a course. This results in a total of 221 courses that satisfy all three conditions. These courses are referred by **Eligible-Courses** in the paper.

5.2 Data Pre-processing

5.2.1 Filter for Student-Specific Approach

For the Student-Specific approach, we sample students who have completed enough courses and with variance in their performance across the different courses. We choose students with a greater than a four-course history and with at least two of the courses have passing and failing grades within the **Eligible-Course**. We found 586 students matching these criteria and their distribution is shown in Figure 4 with the **4363** student course performance data. We used the stratified shuffle split method to achieve cross validation and the parameters used for this method list are displayed in Table 2. In the stratified shuffle split method, if we set the test set size to 0.1, this indicates that the testing set size is 10% of the overall dataset and remaining 90% for training data in each round. In each round, the training and testing dataset are randomly picked and this is no overlap between them. We set the test size to a smaller value because we want to set a prediction for a current student's course enrollment based upon their whole recorded course histories. A total of 436 testing data and 3927 training data records were used in this experiment.

Table 2: Stratified Shuffle Split Parameters Table

Parameter	Description	Value
n_splits	Number of splitting iterations.	20/20
test_size	The size of the testing set.	0.1/0.1

5.2.2 Filter for Course-Specific Approach

For Course-Specific approach, we choose courses from eligible course pools having more than 100 students with a final grade greater than 0. This selection resulted in 107 courses. The classroom size distribution is shown in Fig 5. We use stratified shuffle split method to split training and testing data.

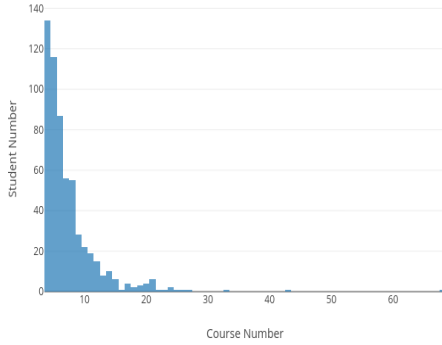


Figure 4: Course Number Distribution

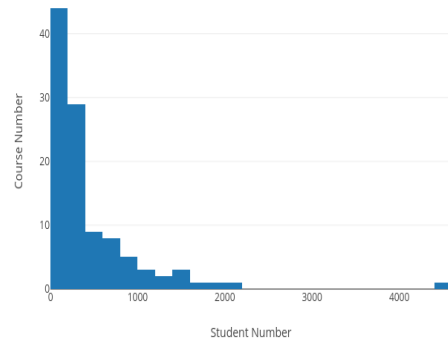


Figure 5: Student Number Distribution

6 RESULTS AND DISCUSSION

6.1 Student-Specific Approach

Figure 6 shows the accuracy and f1 scores for the classification performance varying the amount of data seen for training and testing. Results are reported for three different machine learning algorithms. The x-axis shows the percentage of data considered. For example, 10% denotes that we use the first 10% of the feature of training and testing data. We also evaluate the three types of features created: (i) course-feature, (ii) student-feature and (iii) hybrid features. The Figures show that the f1 score and accuracy increases from 70% to 80% as we increase the amount of data used for features from 10% to 40%. It is expected to determine the student's final grade and risk of dropping a class as the course draws to a close. The strong performance of the prediction methods early on shows the promise of making intervention decision early. We also observe that Random Forest methods outperform the logistic regression and nearest neighbor algorithms. We also report accuracy results using both, the course and student features. These results mirror the experiment conducted by Hlosta's team (Hlosta et al., 2017), which show far more accurate prediction methods by addition of features beyond grade-related within the prediction framework. In summary, given 40% of data uses (indicates course already passed 40%), we report 82.7% accuracy and 80.9% f1 score. Even for 10% of observed data, we report 71% accuracy and 67% f1 score. For the student-specific approach, the hybrid features within the random forest framework proves to be the winner.

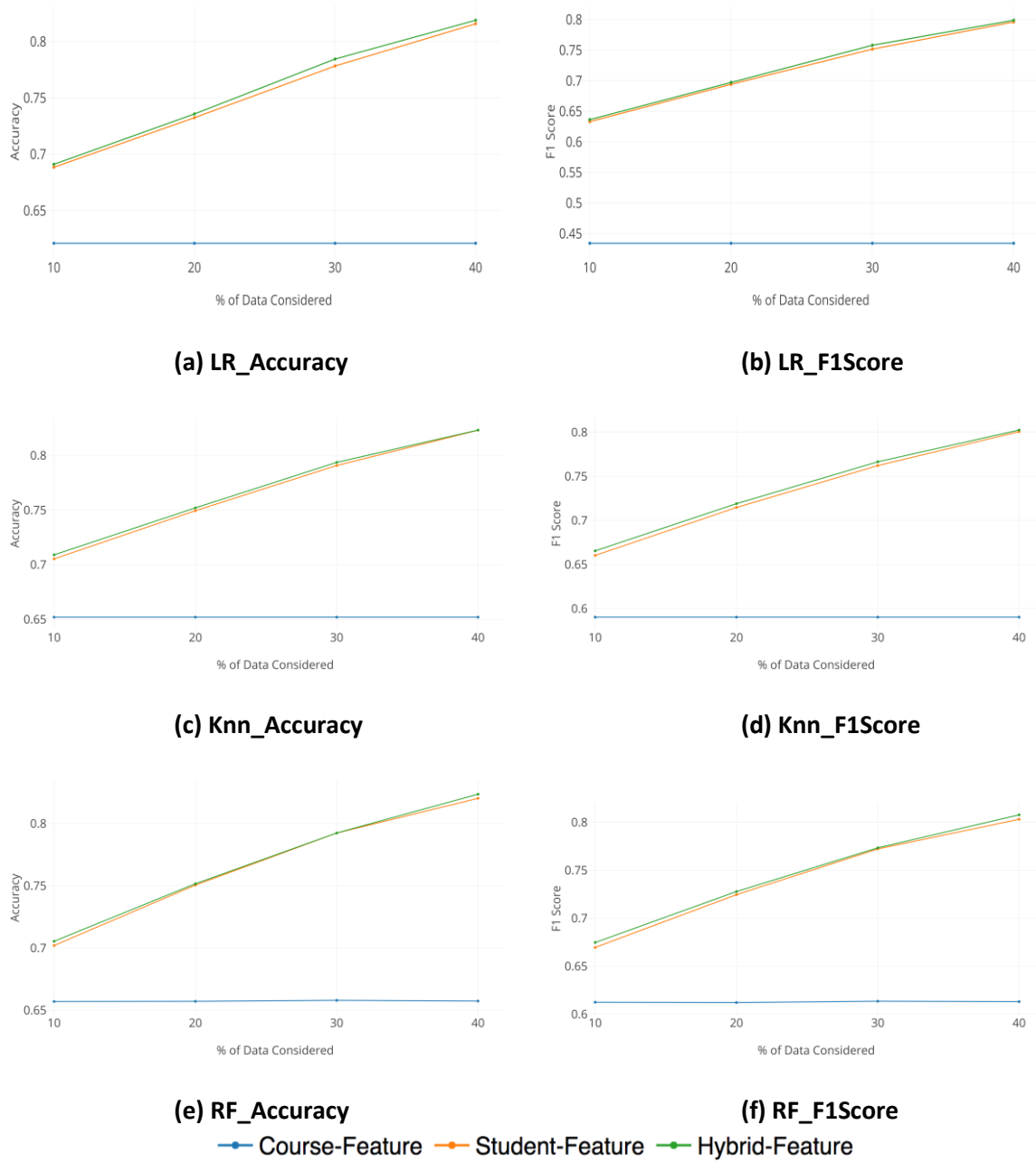


Figure 6: Accuracy and F1 score using student, hybrid features respectively for three methods

6.2 Course-Specific Approach

Figure 7 shows the performance of the methods with the Course-Specific Approach. We report good accuracy and F1 scores with the Course-Specific approach. Specifically, we observe approximately 95% accuracy and 90% f1 score using just the first 10% of the data/features. As we increase the amount of data used for making predictions (from 10% to 40%) for the final grade we do not see a substantial change. For a given course, the features related to interaction patterns computed for a given student are similar as time progresses within a semester. As such, there is little variance between early stage features and final stage features. The RF was the best performing method in the Course-Specific approach, followed by LR and Knn. We noticed a similar trend for the Student-Specific approach as well. As noted before this approach still has limitations. For now, we only conduct experiments in the same class with part of the students from this course are simulated as students enrolled in the particular course for

the first time. In real-world scenario, it is not guaranteed that two courses offered in different semesters would share the same feature types and can vary from semester to semester. The Course-Specific Approach though shows better results than the Student-specific approach should account for the assumptions discussed above. The Student-Specific approach only works well if a student has a longer course history (more courses are taken over a longer period). If we have enough have detailed enough students course performance records, we will consider the Student-Specific Approach to be the best predictor for the outcome.

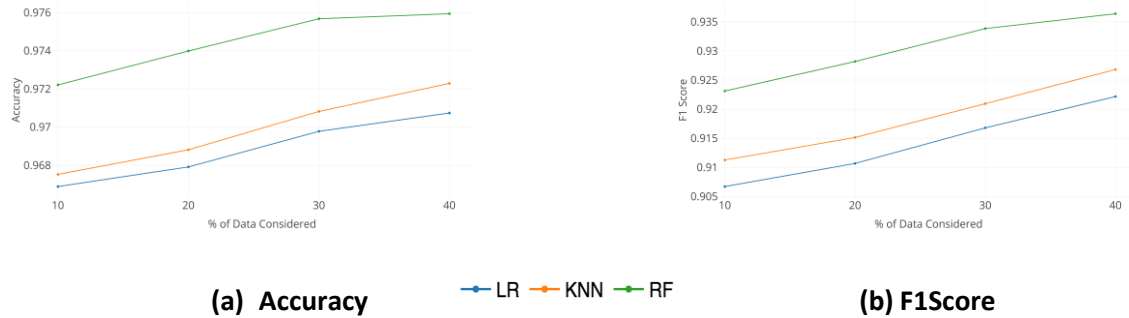


Figure 7: Average accuracy and F1 score for Course Specific Approach

7 CONCLUSION AND FUTURE WORK

In this study, we developed a framework to predict the student's final class performance based upon features extracted from LMS and especially from the first few weeks of the semester. In our Student-Specific approach, even though the prediction based on the early feature did not perform as well as the complete features, the approach still achieved close to 83% accuracy over using 40% feature for current courses. Random forest approach was found to be the most suitable for the algorithm. We propose to consider non-grade related features such as gender, citizenship and professor/instructor in the future. Regardless we plan to enhance our approach to handle the student with the fewest records. In the Course-Specific approach, the machine learning methods achieved nearly 95% accuracy using the first 10% features.

ACKNOWLEDGEMENT

This work is supported by an NSF REU Site on Educational Data Mining at George Mason University; Grant No. 1757064.

REFERENCES

- Boroujeni, M. S., & Dillenbourg, P. (2018, March). Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 206-215). ACM.
- Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In *Data Mining and Advanced Computing (SAPIENCE), International Conference on* (pp. 91-95). IEEE.
- He, J., Bailey, J., Rubinstein, B. I., & Zhang, R. (2015, January). Identifying At-Risk Students in Massive Open Online Courses. In *AAAI* (pp. 1749-1755).
- Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017, March). Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 6-15). ACM.

- Jokhan, A., Sharma, B., & Singh, S. (2018). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education*, 1-12.
- Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5), 411-426.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.
- Le, C. V., Pardos, Z. A., Meyer, S. D., & Thorp, R. (2018, June). Communication at Scale in a MOOC Using Predictive Engagement Analytics. In *International Conference on Artificial Intelligence in Education* (pp. 239-252). Springer, Cham.
- Na, K. S., & Tasir, Z. (2017, November). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In *Big Data and Analytics (ICBDA), 2017 IEEE Conference on* (pp. 118-123). IEEE.
- Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention*. Nova Southeastern University.
- Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting performance on MOOC assessments using multi-regression models. *arXiv preprint arXiv:1605.02269*.
- Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In *Educational data mining 2008*.
- Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. In *Educational data mining 2015*.
- Zhang, L., & Rangwala, H. (2018, June). Early Identification of At-Risk Students Using Iterative Logistic Regression. In *International Conference on Artificial Intelligence in Education* (pp. 613-626). Springer, Cham.

Personalized Learning Path Recommendation Based on Knowledge Structure

Lingling Meng

Faculty of Education, Department of
Education Information Technology, East
China Normal University
llmeng@deit.ecnu.edu.cn

Mingxin Zhang

Faculty of Education, Department of
Education Information Technology, East
China Normal University
2374624930@qq.com

Wanxue Zhang

Faculty of Education, Department of
Education Information Technology, East
China Normal University
Zhangwanxue0703@163.com

Xueyu Shi

Jinshan Elementary School, Liangjiang New
District, Chongqing
sxyjay1991@163.co

ABSTRACT: With the rapid advancement of education, personalized learning has gained a considerable attention in recent years. How to select the suitable learning objects (LOs) and provide a personalized learning path for learners is a difficult task. In this paper, we propose a generation method based on knowledge structure and learning diagnosis. First, we construct the knowledge model by marking the precursor knowledge and subsequent knowledge. Building a new student model according to the student's ability and learning status. Last, we use Euclidean distance to calculate the similarity between student's ability and the difficulty of the knowledge, and according to the similarity to select next knowledge for learners. Thus the system adaptively generates a learning path for the student in the light of their knowledge level. From the experimental results, it can be concluded that the proposed approach shows high adaptability and efficiency in e-learning system.

Keywords: Personalized learning path, knowledge structure, learning diagnosis

1 INTRODUCTION

With the advance of technology, people want to adaptively recommend learning object for learns. Most researchers use collaborative filtering to recommend the next knowledge depending on their interest. However, in online education, their interest could not reflect the relationship of knowledge and fit their need of learning. The most important thing is to build the net of knowledge, then select the best suitable knowledge for them based on their knowledge status and cognitive level.

In the past, the task of constructing the net of knowledge was completed by teachers, and teachers usually used score to decide the status of the learner, which were not very accurate. But now, schooling from "speaking with experience" to "using data-driven to manage and innovate" (Shengquan Yu & Xiaoqing Li, 2017). By analyzing a large amount of data, it is possible accurately build the relationship of knowledge and identify student's knowledge status and learning level, predicting learning outcomes, and provide personalized learning intervention and guidance to achieve a more flexible and convenient online learning.

Therefore, we can rely on learning data to construct the relationship of knowledge and analyze student's learning level and knowledge states, trying to adjust the learning progress to respond to their learning needs. From the perspective of data mining, the research builds knowledge model, establishes students' model through their learning data, and adaptively select next knowledge based on the similarity of knowledge and student's learning level, so as to generate the personalized learning path for students.

2 RELATED WORK

Personalized learning path is a method to design the sequence of knowledge for learners. In the area's researchers are focussing on the following two aspects: First, how to characterize the feature of students; second, how to generate different learning paths for students.

In the learning path, a large number of scholars are devoted to exploring what kind of information can characterize students. Brusilovsky first proposed representing students' learning information from the two dimensions: learning objectives and knowledge levels (Brusilovsky, 2003). His method was oriented by student's target, choosing next knowledge based on students' knowledge level. This method only interpreted learning need and lacked the personalized information of students. Graf and Kinshuk proposed that if the current learning content of the student was contrary to his or her learning style, the learning would be difficult to develop smoothly, so the course recommendation should be based on the student's learning style (Graf & Kinshuk, 2009). But in recent research, the proposal of this method has been doubted by many scholars. Paul and Jeroen thought learning style was a legend and received very little support from objective studies (Kirschner & Merriënboer, 2013). And some other scholars also believed it lacked scientific evidences (Kirschner & Paul, 2017). Therefore, researchers wanted to find other aspects to describe learners. Such as Peng generated learning paths based on interest (Jianwei Peng, 2009) and Jiang selected learning resources based on behavioral tendencies (Qiang Jiang, 2012). Dwivedi and Bharadwaj utilized the student's interest to cluster student groups and collaboratively generate learning paths (Dwivedi & Bharadwaj, 2018).

In addition, some scholars give attention to how to design algorithms to generate the personalized learning path. Similarity algorithm is the most widely used matching algorithm. Huang used the similarity to calculate the degree of matching between the students' cognitive level and the difficulty of knowledge (Zhifang Huang, 2015). Jin used cosine similarity to judge the degree of matching between knowledge and utilized fuzzy synthetic evaluation algorithm to evaluate students' mastery from multiple angles (Muxin Jin, 2017). Others used the nearest neighbor algorithm to calculate the similarity of learning resources and designed a recommended learning resource based on the bipartite graph (Zongbao Liu, Hua Li & Wenai Song, 2018).

But these methods did not consider the relationship of knowledge, and the next learning object could not fit the student's need. Therefore, this paper wants to develop a strategy of learning path generation by linking the student's performance to the knowledge components.

3 METHODOLOGY

3.1 Student Modeling

Adaptive systems commonly used overlay models, perturbation models, and cognitive models to describe students. These models have their feature, but it is necessary to reflect the student's knowledge status in online education. Therefore, we propose that a new model characterizes the student from two aspects: behavioral information and status information (as figure 1).

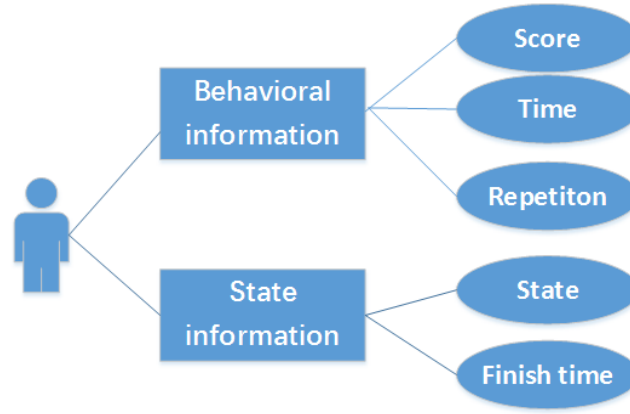


Figure 1: Student model

Behavioral information mainly records the academic performance, time and repetition of the student i at the knowledge point j . Using the form of $S_{ij} = \{Score, Time, Count\}$ to collect this information. The $S_{ij} = (State, Time)$ is used to store the student's learning status of knowledge and records the time of accomplishing.

3.2 Knowledge Modeling

Knowledge does not exist independently, so Acampora and Loia defines the relationship of knowledge as an inclusion relationship, a priori relationship and weak relationship when establishing the knowledge model (Acampora & Loia, 2011). Based on this, we also consider the relationship of knowledge, constructing the knowledge model from the basic information and relationship information (as figure 2).

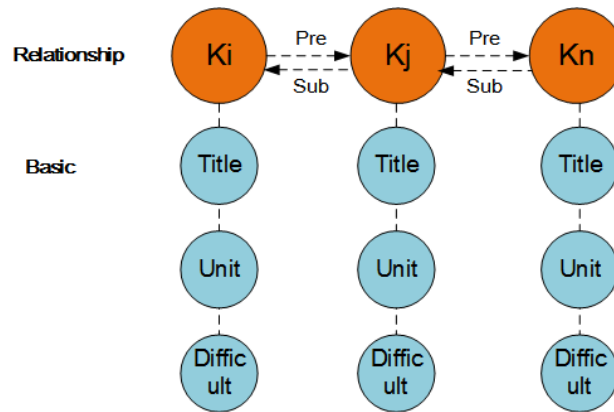


Figure 2: Knowledge model

Basic information includes the title, unit, and difficulty, recorded in the form of set $K_i = \{Title, Unit, Dif\}$. The relationship is defined as the predecessor relationship (pre-knowledge) and the subsequent relationship(sub-knowledge), which represented by the set $K_i = \{Pre, Sub\}$.

3.3 The Method of Generation

3.3.1 Estimation of learning ability

Student ability is the degree of ability that students reflect in learning knowledge, and their ability level is constantly changing in the process of learning knowledge. Therefore, in this research, the student's ability level is regarded as a dynamic value. It is measured from two dimensions: knowledge and behavior. The academic achievement is reflected in the student's knowledge level; at the behavioral level, the repetition of students learning resources in the platform and the learning time (Ballera, 2018) are used to evaluate their abilities by mining the log files. The specific calculation method is as shown in formula 1:

$$Ability_{ij} = p * \frac{Score_{ij}}{S} + q * \frac{(T - Time_{ij})}{T} + (1 - p - q) * \frac{N - Count_{ij}}{N} \quad (1)$$

Where S, T, N are the threshold of score, time and count, the p, q and (1-p-q) are the weight of score, time and count to calculate the ability of students.

3.3.2 Calculation of knowledge level

The level of knowledge reflects the difficulty of knowledge and what kind of ability the student should master. Difficulty value is typically used as a reference for knowledge levels. However, the difficulty value is only assessed by the senior teachers, not adjust the value according to the specific learning situation of the students, so it's hard to embody personalization. Therefore, in this study, the difficulty and student performance are simultaneously used as variables for calculating the level of knowledge. Considering the knowledge has been divided into pre- knowledge and sub-knowledge, the pre-knowledge has historical achievements of students, but the sub-knowledge has no student's academic performance. Therefore, the calculation methods of the two are very different.

The knowledge level of the predecessor knowledge is calculated on the difficulty given by the teacher and the student's achievement. The calculation is formula 2:

$$Level_{ij} = p * Dif_i + (1 - p) * \frac{S - Score_{ij}}{S} \quad (2)$$

Where S is the threshold of score, the p and (1-p) are the weight of difficult and score to decide the influence in the level of knowledge j.

If the system is in the initial state, the formula 3 for calculating the knowledge level:

$$Level_{ij} = \frac{Dif_i}{5} \quad (3)$$

If a student has entered the system to study, the formula 4 for calculating the knowledge level of subsequent knowledge

$$Level_{ij} = p * Dif_i + q * (1 - \frac{\sum_{i=1}^n Score_{ij}}{N}) + (1 - p - q) * (1 - \frac{\sum_{j=1}^n Count_j}{N}) \quad (4)$$

Where N is the number of students who have learned, we calculate the total of score and count, then take the average of them. The p, q and (1-p-q) are the weight of difficult, score and count to decide the influence in the level of knowledge j.

3.3.3 The strategy of choose

After calculating students' learning ability through the calculation 1 and use the feature of the knowledge to correct the knowledge level of the current knowledge by calculation 2, 3, or 4, standardize the value of two. Then, using the Euclidean distance formula to calculate the similarity between them. The calculation formula is 5:

$$Sim_{ij} = \sqrt{|Ability_{ij} - Level_{ij}|^2} \quad (5)$$

Thereby obtaining the distance between students and knowledge and select the knowledge which has the shortest distance as the next.

3.3.4 Generative process

Step 1: Enter the student ID number

Step 2: Sort the unit based on the student's initial score

Step 3: Select the initial knowledge of each unit according to the obtained unit sequence

Step 4: Select the best suitable knowledge to the students

Step 5: Judging the learning state of the students, which are consisted of three: return state, adaptive state, and update state (as figure 3).

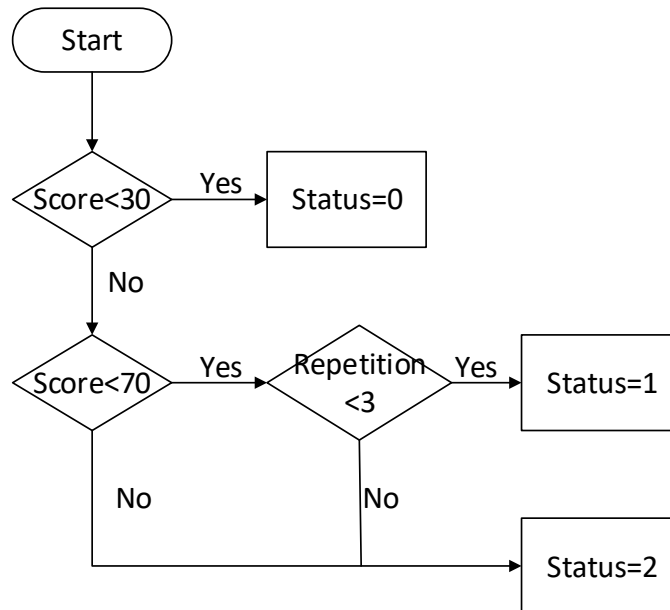


Figure 3: Learning status

If the student's state is returning, calculate the ability level of the student according to the formula (1), and select the most familiar pre-knowledge as next; If the student's status is adaptive and the repetition of learning is less than 3, returning the current knowledge; If the student's status is updated, calculate the learning ability of the student and return the most suitable sub-knowledge as next.

Step 6: Judging whether the learning is complete. If it is finished, step 7 is performed; if the learning is not done, step 5 is performed.

Step 7: Output the student's learning path.

4 EXPERIMENTAL ANALYSIS

We take the sixth-grade mathematics as the knowledge content, and select a senior math teacher in Chongqing to determine the difficulty and relationship of each knowledge, initially generating the relevant values of the knowledge model. The learning data of 24 students (1026 in total) was used as the experimental data. Their learning path was generated as figure 4 and figure 5.



Figure 4: student 1 learning path

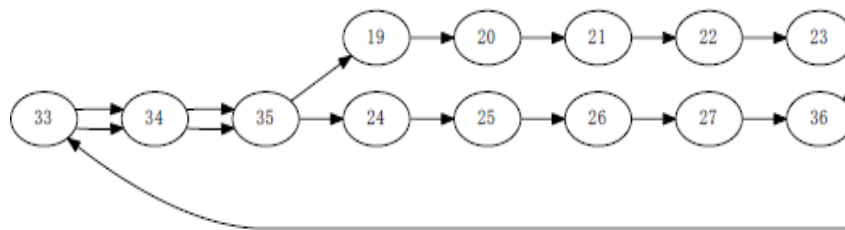


Figure 5: student 2 learning path

In figure 4 and figure 5, the number is the label of knowledge. As those figures show that different student's learning path is determined by the student's learning ability. If the student can quickly master the knowledge, the next will be difficult and some knowledge will be passed. To prove this method, the accuracy and diversity of the generation were evaluated. The test results are shown in Figure 6 and Figure 7.

In Fig 6, the abscissa indicates the number of test data, and the ordinate indicates accuracy. It can be concluded that the accuracy value fluctuates between 0.3 to 0.6, where the minimum value is 0.37 and the maximum value is 0.56, indicating that the next knowledge is within the learner's ability and do not exceed the student's ability threshold. But the accuracy is not very high, maybe the knowledge needs to divide more specific that fit most of student's need.

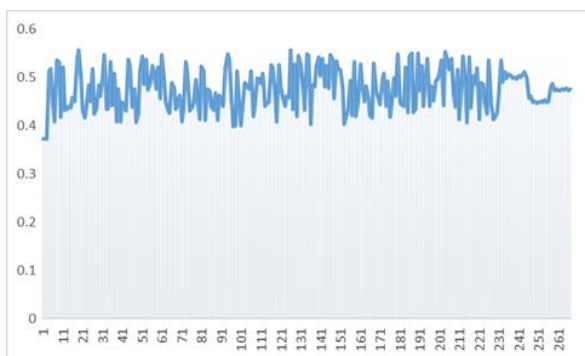


Figure 6: Accuracy

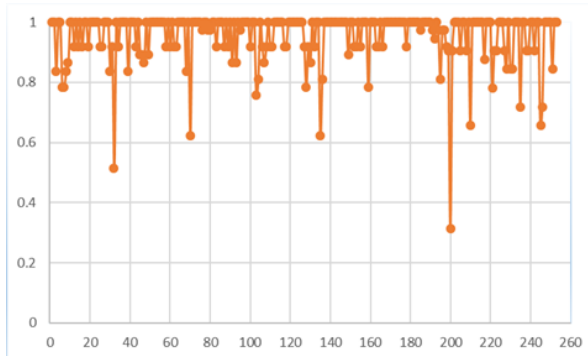


Figure 7: Diversity 6

In Fig 7, the abscissa also represents the number of test data, and the ordinate is the diversity. As shown in Figure 7, the diversity is mostly close to 1, with a minimum of 0.32 and a maximum of 1. According to statistics, the number of value, less than 0.5 only one, and the overall average is 0.95, indicating that the diversity of the method is high. In the method, a few learners' learning path repetition rate is high because the ability of the two is similar, resulting in a high similarity of the learning paths; but other learners' learning path is diversified and can be adapted according to the personality of the students, reflecting the personalization of the method.

5 CONCLUSION AND FUTURE WORK

This paper constructs the knowledge model by marking the precursor knowledge and subsequent knowledge. Constructing the student model according to the student's ability and learning status. Then the similarity calculation is carried out by the ability level of the students and the knowledge level in the two models. Thus, adaptively recommend next knowledge according the student's mastery degree. Finally, generating personalized learning path by the sequence of learning knowledge. Compared with other papers, the learning path of this paper estimates the students' ability, selecting the subsequent knowledge and generates the learning path, which not only considers the order among the knowledge but also satisfies the knowledge level of the students to adapt the speed of learning. Finally, accuracy and diversity of the method are verified by data.

It should be noted that the method only considers the student's learning characteristics, including the student's learning ability and learning status, and does not consider other aspects of the student, such as learning objectives, information processing, and memory threshold. So future research may focus on those aspects to generate learning path.

REFERENCES

- Acampora, G, Gaeta, M, & Loia, V. (2011). Hierarchical optimization of personalized experience for e-Learning systems through evolutionary models. *Neural Computing and Applications*, 20(5), 641-657
- Ballera, M, Lukandu, I. A, & Radwan, A. (2015). Personalizing E-learning curriculum using: reversed roulette wheel selection algorithm. *The International Conference on Education Technologies and Computers*(pp.91-97). IEEE.
- Brusilovsky, P. (2003). From adaptive hypermedia to the adaptive web. In *Mensch & Computer 2003* (pp. 21-24).
- Dwivedi, P.Kant, V.& Bharadwaj, K. K. (2018). Learning path recommendation is based on modified variable length genetic algorithm. *Education and Information Technologies*, 23(2), 819-836.

- Graf, S, & Kinshuk. (2009). Analyzing the Behaviour of Students in Learning Management Systems with Respect to Learning Styles. *Advanced in Semantic Media Adaptation and Personalization*.
- Jianwei Peng. (2009). Personalized learning path recommendation based on memetic algorithm. (Doctoral dissertation, Hunan University).
- Kirschner, P. A., & van Merriënboer, J. J. (2013). Do learners really know best? Urban legends in education. *Educational psychologist*, 48(3), 169-183.
- Kirschner, & Paul, A. (2017). Stop propagating the learning styles myth. *Computers & Education*, 106, 166-171.
- Muxin Jin. (2017). Research on the design of the adaptive navigation for primary media literacy course based on EAHAM. (Doctoral dissertation, Yunnan Normal University).
- Qi Liu, Peng Ding, Xiaoqing Huang, Jingjing Dong(2018). Research on Personalized Learning Recommendation System Based on Test Network [J]. *Modern Educational Technology*,28(06):11-16
- Qiang Jiang. (2012). Research on Supported Model and Implementing Mechanism for Adaptive Learning System. (Doctoral dissertation, Northeast Normal University)
- Shengquan Yu, Xiaoqing Li. Research on the Analysis and Improvement of Regional Education Quality Based on Big Data [J]. *e-Education Research*,2017(7):5-12.
- Zhifang Huang. (2015). Dissertation research on adaptive learning path recommendation for E-learning-examined by junior middle school physics. (Doctoral dissertation, Central China Normal University)
- Zhongbao Liu, Hua Li, Wenai Song, Xiangyan Kong, Hongyan Li,Jing Zhang. Research on Mixed Recommendation Method of Learning Resources Based on Bipartite Network[J]. *e-Education Research*,2018,39(08):85-90.

PBL Discourse Analysis in Ill-constructive Field Based on ENA ——A case study of Chinese medicine education

Bian Wu

Department of Educational
Information Technology,
East China Normal
University
bwu@deit.ecnu.edu.cn

Fengfeng Du

Department of Educational
Information Technology,
East China Normal
University
1359357060@qq.com

Yiling Hu

Department of Educational
Information Technology,
East China Normal
University
ylhu@deit.ecnu.edu.cn

ABSTRACT: Problem-based learning (PBL) brings new opportunities for the resolution of ill-structured problems in education. However, in the process of PBL, the procedural evaluation is neglected, and the individual ability of the students cannot be evaluated. In addition, the different guiding styles of teachers are used to study the quality of PBL discussion and the development of students' thinking ability. Therefore, this study takes traditional Chinese medical(TCM) education as an example, and selects 19 graduate students and 2 teachers. The two groups of students are randomly divided into two groups (D group and F group), which are controlled by two different modes (facilitative tutor guidance and directive tutor guidance). The facilitating teacher guides the discussion process. A qualitative data quantification modeling method called Epistemic Network Analysis (ENA) was used to model the students' TCM thinking ability and the teacher's guiding mode. The results show that there are significant differences in the ability networks of the two groups of students. Compared with the guidance and control type, the empowerment-promoting guiding style is more effective for students' social knowledge construction and disciplinary thinking ability. Finally, the innovative quantitative ethnographic method used in this study can effectively solve the problems of modeling and evaluation in social learning activities.

Keywords: PBL, Quantitative ethnography, ENA, Chinese Medicine Education

1 INTRODUCTION

There are many ill-structured problems in educational research. In the specific situation, the definition, concept, rules, principles and solutions of ill-structured problems are difficult to determine, and the evaluation process is also very inconvenient (Luszcz,1984; Wood,1993). PBL (problem-based learning) brings new opportunities for the resolution of ill-structured problems. PBL was pioneered by American professor of neurology Barrows in 1969 at McMaste University in Canada. PBL is a real-world, student-centered approach to teaching students' learning by analyzing and solving real-world problems. In general, PBL emphasizes that learning is placed in complex and meaningful problem situations, and the real problems are solved through the cooperation of learners, so that the scientific knowledge hidden behind the questions is learned, and the skills and skills of students to solve problems are cultivated. The ability to learn independently. Although PBL is different from traditional subject-based pedagogy, there are still many problems in practice. In the PBL, the teacher is the leader of the teaching, and the teacher will guide the discussion process of the students. Regarding the effective guidance in the PBL process, the researchers did not reach consensus on “facilitative tutor guidance” and “directive tutor

guidance". For example, some researchers believe that experienced teachers should include the following skills: supporting student self-directed learning, questioning, pushing student for explanations, revoicing and summarizing, and supporting hypothesis generation (Hmelo-Silver et al., 2006; Aarnio et al., 2014). But Bude (2011) and others believe that the directive tutor guidance in PBL can improve students' understanding ability compared with traditional counseling methods. Papinczak, Tunny and Young (2009) believe that the essence of promoting student learning is to strike a balance between the two guiding methods, but it remains unclear how these two guiding methods affect the PBL process. In addition, in disciplines such as medicine with complex knowledge structures and domain problems with non-constructive features, modeling and evaluating students' professional thinking skills requires quantitative modeling and meaning construction based on a deep understanding of professional practice activities. It is necessary to provide rooted and statistically valid evaluation evidence. In order to assess students' professional skills in the PBL process, researchers not only need to understand how professional knowledge develops in a particular field, but also the learning effect of PBL is the result of collaborative learning. Researchers also need to conduct a process evaluation of the discussion of PBL. Leung (2012) believes that ethnographic research helps us understand the social learning processes in PBL, such as student group discussions and teacher-student interactions, and improve our educational practices. Therefore, to analyze the process of learning and professional development in the context of PBL, dialogue research is the core of ethnographic research. However, ethnographic research may face the challenge of theoretical saturation when analyzing a large number of qualitative data and clarifying meaningful teaching models and evaluation models. To support the basic understanding of ethnographic analysis of discourse data, Shaffer et al. (2007) proposed an ENA (epistemic network analysis) method to quantify expertise using a network model. In a specific discourse segment, ENA can express meaningful cognitive connections between cognitive framework elements through the co-occurrence of concepts (Landauer, McNamara, Dennis, & Kintsch, 2007; Lund & Burgess, 1996).

Therefore, we explore the application of ENA in PBL education of Chinese medicine, explore the influence of teachers' guidance mode on students' professional ability, solve many challenges faced by students in medical education, and deeply understand the development of students' ability in TCM education. We have proposed the following two research questions to analyze the group discussion and provide experience for the future development of PBL: Question 1: in the PBL discussion of ill-structured problems, how do the two different guidance methods affect the development of TCM competence of the two groups? Question 2: In the two groups, what are the different regulatory modes of the facilitative teacher and the directed teacher?

2 LITERATURE REVIEW

2.1 Ill-structured Problems

In a specific situation, it is generally difficult to clearly define these ill-structured problems, and the statement of the problem does not help the problem (Chi & Glaser, 1985). In the process of solving ill-structured problems, the number of targets is difficult to clearly define, and the information that is good for the solver is usually incomplete, incorrect, and vague. At the same time, in the process of solving, the concepts, rules and principles needed to solve ill-structured problems are uncertain, and the concepts, rules and principles are inconsistent and contradictory. Many researchers use qualitative methods to understand the solution process of ill-structured problems. In fact, some ill-structured problems may have multiple solutions, and some even find a suitable solution (Ward & Woisetschlaeger, 1983r). Given that solutions to ill-structured problems are difficult to achieve universally, there are multiple criteria

for evaluating solutions to ill-structured problems. In addition, since the elements interact in a specific problem situation, the elements of the problem will be very different in different problem situations. These factors have brought great inconvenience to the successful evaluation of ill-structured problems. PBL refers to the use of knowledge and skills to solve a series of practical problems to achieve the purpose of constructing experience (Bligh, 1995), a method to solve ill-structured problems. PBL emphasizes the active learning of students; PBL links learning to larger tasks or problems, and puts students into the problem; PBL designs authentic tasks, emphasizing the setting of learning into complex and meaningful problem scenarios. Solve problems through students' independent inquiry and cooperation, so as to learn the scientific knowledge hidden behind the problem, to form problem-solving skills and self-learning ability.

2.2 Teacher's Guiding Style in PBL

PBL contains three elements: questions, students and teachers. Problem is the driving force of PBL, students are the problem solvers, while teachers are the helpers and promoters of students' learning. Teachers' main responsibility is to guide students to acquire problem-solving strategies. Therefore, in PBL, teachers' ability to use promotional guiding skills plays a decisive role in the effect of PBL. The process of PBL learning is not only the discussion among students, but also the interaction between teachers and students, which is conducive to the development of students' professional knowledge. There is no consensus among researchers in favor of accelerative guidance or direct guidance on the effective counseling methods in the PBL process. Barros (1988) describes the role of the teacher as a learning facilitator and believes that the skills that an experienced learning facilitator should have include: supporting students to learn independently, asking questions, facilitating student interpretation, retelling, summarizing, and supporting hypothesis generation. However, Budi et al. (2011) believe targeted questions and guidance of domain experts in PBL can promote students' deep understanding. Pazarzak et al. (2009) believe that instructor is essentially balancing the two guiding styles, but how these two guiding methods affect the interactive process in social learning like PBL.

2.3 Quantitative Ethnography

From the perspective of social constructivist learning theory, in order to capture the essence of social behavior, Leung (2012) believes that ethnographic research methods can better help us understand the social learning process, such as understanding student group discussions and the interaction between teachers and students. Thereby improving our educational practice. Therefore, in order to analyze the development process of learning and professional knowledge in the context of PBL, dialogue research is the core of ethnographic research. However, ethnographic research may face the challenge of theoretical saturation in analyzing a large number of qualitative data, clarifying meaningful teaching models and assessment models. To support basic understanding of ethnographic analysis of discourse data, Shaffer et al. proposed an epistemic network analysis (ENA) method to quantify expertise using a network model. ENA has been used to model and evaluate complex problem-solving activities in the STEM field, such as engineering education (Chesler et al., 2013) and urban planning (Nash & Shaffer, 2011; Schafer & Bagley, 2015). In a particular discourse segment, ENA is able to express meaningful cognitive connections between cognitive framework elements through conceptual co-occurrence (Landauer, McNamara, Dennis, & Kintsch, 2007; Lund & Burgess, 1996). In simple terms, the ENA analysis first constructs an adjacency matrix for each discourse window to quantify the co-occurrence of coded elements, accumulate each adjacency matrix, and normalize it into a high-dimensional space. Then, the singular value decomposition (SVD) is used to reduce the dimension in the high-dimensional space. Finally, an optimization algorithm is used to place the nodes of the network model (i.e., complex

collaborative thinking coding elements) in the first two dimensions of the SVD, so that each the centroid of the network model corresponds to the position coordinates of the network space after dimension reduction. Finally, two analysis results are generated: (1) the position of each network in the projected metric space; (2) the weighted network map of each network to explain the reason why network model is at the current location.

3 METHODS

3.1 Participants

In this study, 19 graduate students and two instructors from the Acupuncture Medical Book Reading course at a Chinese medicine university in eastern China were selected. The focus of this course is to enable students to learn to choose TCM books related to acupuncture. This study focuses on one module of the course. The course's professors use the PBL method. Participants were randomly divided into two groups, one teacher per group, one tutor using an empowerment-promoting strategy, the group name was F group (N=10); the other teacher used a guidance-controlled strategy, group name for group D (N=9). The difference in coaching between the two teachers was recorded by the teacher's self-reflection and the investigator's classroom observations.

3.2 Procedure

The course module of this study is a typical three-stage PBL case written by the teacher, which describes a patient with traumatic brain-complement syndrome who needs acupuncture treatment. The course week for this module is three weeks. The first week and the third week are 1.5h classroom discussions, and the second week is self-learning week. The content of the first two phases of the case needs to be completed in the first week of the course. After the first week of the course, the two groups will ask a series of questions for the second week of the autonomous learning course, in the third week of the course. The two groups will share the knowledge they have learned, as well as their current views on the case and the diagnosis and treatment options. In the middle of the second self-discussion, the two teachers gave different guidance to facilitate their discussion until the two groups reached the final conclusion. Finally, the two groups will share the solution to each other. Throughout the process, the two teachers will rate their scores based on their performance.

3.3 Data Collection and Encoding

The researchers recorded and transcribed the first and second phases of the two groups (Group F & Group D) for a total duration of 6 hours. In this study, the discourse, that is, the rotating speech segments of the participants in the conversation, is used as the unit of coding analysis, and is coded from the perspectives of the students' TCM ability and the teacher's TCM regulation strategy. With regard to the TCM capability perspective, based on the grounded theory method, the axial and selective coding is used to simplify the coding scheme, and two dimensions are added to the selected six coding dimensions to extend the research of the previous Western medical PBL competence assessment to TCM fields such as Chinese medicine concepts and acupuncture concepts. For the coding scheme of the regulation strategy, this study draws on the six dimensions (positioning, planning, execution, monitoring, evaluation and interpretation) proposed by Lajoie et al., taking into account the two dimensions of implementation and evaluation included in this study. The data is limited, so delete these two dimensions. See Table 1 for definitions and examples of coding schemes for TCM capabilities and regulatory strategies. The transcription data was encoded by two independent researchers and the

coding results were well-consistent ($\kappa = 0.74$). Finally, the problem of score difference was solved through discussion.

Table 1: Coding scheme of TCM competence and PBL regulation

Two dimensions	Codes	Definition	Examples
TCM competence	Clinical concept	Understanding of clinical concepts, such as symptoms, signs, and clinical diseases.	cervical spondylosis, traumatic brain injury, nodules, arthritis
	Concept relation	Relations between different clinical concepts, TCM concepts and meridians concepts.	prototype, relationship, difference, type, connection
	Reasoning & Justification	Use evidence to support/refute hypothesis, and elaborate the reasoning process	possible, cause, relate to, lead to, match, rule out, affect
	Clinical action	Actions in relation with clinical diagnosis, treatment, and other management procedure	check, examine, solve, treatment, inquiry, clinical, fMRI, X-ray
	Diagnostic conclusion	Draw diagnostic conclusions of patient problem	tend to, diagnose, problem, disease, judge
	TCM concept	General TCM theory such as Zang Fu organ theory, Qi forms and functions	energy(qi), pulse, sea of marrow, deficiency and excess, pattern, Bi syndrome, blood stasis, fire, pattern differentiation
	Acupuncture concept	Theory about Meridians and Sub-Meridians and treatment method of acupuncture and moxibustion	meridian, needling, acupoint, tender spots (Ashi point), points, channel, tendino-muscular meridians
PBL regulation	Orientation	Activate prior knowledge; Establishing task demands; Revoicing; Hypothesizing	Any other opinions?
	Planning	Looking for particular information; Subgoalting; Using external source to get explanation; Forming action plan	Can we summarize our conclusions?
	Monitoring	Claiming understanding; Error detection; Noticing inconsistency; Noticing unfamiliar terms	Did you just say he was...?
	Explanation	Elaboration; Justifying; Verifying; Summarizing	What's the difference between ...?

3.4 Data Analysis

By coding the conversational content in the PBL discussion in the first week and the third week of the two PBL groups, and then modeling the cognitive network of the encoded data, we obtained the clinical thinking of TCM based on two discussions before and after the two PBL groups. Cognitive network maps, and cognitive network diagrams of the teaching scaffolding strategies of two PBL instructors. We quantitatively analyze and compare the network diagram structures of the two groups and the

corresponding two tutors, and select typical discussion segments for qualitative conversation analysis to provide qualitative evidence support for the quantitative analysis results of cognitive network graphs.

4 RESULTS

4.1 Thinking Ability Modes of Two PBL Students

The PBL discourses in the first and second discussions of Groups F and D were coded according to seven TCM capability components, and their respective plotted maps and networks were generated using ENA. Each point in the plotted point map represents each student's TCM capability network, and the squares mark the average network location of all students in the same group of PCL sessions. The box around the box represents the 95% confidence interval for the group, and the average network is the weighted network for the entire group, represented by the same color in the same predicted TCM capability space.

In order to explore the differences between the two groups of students in the two discussions, the results of the independent sample T test are shown in Table 2. From Table 2 and Figure 1-a, there is no significant difference between the first and second discussions of the F group in the X and Y dimensions; at the 0.01 significance level, the first time in the D group. The discussion and the second discussion were significantly different in the Y dimension (Traditional Chinese Medicine Problem Solving and Clinical Problem Understanding) ($t=-3.503$, $p=0.006$). The interpretation of the X and Y dimensions in Figure 1-a is based on a weighted network structure, as shown in Figures 1-b to 1-g, since the plotted point map and the network are all in the same projection space. In addition, at the level of 0.01 significance, in the first and second discussions of the D and F groups, there was a significant difference in the X dimension (acupuncture and Western medicine) (D: $t = -5.515$, $p = 0.001$; F: $t = -6.141$, $p = 0.001$). However, in the second discussion, there was only a significant difference in the Y dimension (the solution of TCM problems and understanding of clinical problems) ($t=-6.213$, $p=0.022$).

Table 2: Descriptive statistics and t-test of centroids of two groups in two PBL sessions

Comparison criteria	X dimension (acupuncture versus western medicine)			Y dimension (TCM problem solving versus clinical problem understanding)		
	Mean	t-test	p	Mean	t-test	p
F-Group in Session 1	-.219	-5.515	.001**	-.082	.206	.84
D-Group in Session 1	.307			-.109		
F-Group in Session 2	-.275	-6.114	.001**	.02	-2.613	.022*
D-Group in Session 2	.218			.169		
F-Group in Session 1	-.219	.889	.391	-.082	-.824	.427
F-Group in Session2	-.275			.02		
D-Group in Session 1	.307	.827	.42	-.109	-3.503	.006**
D-Group in Session 2	.218			.169		

* $p<.05$, ** $p<.01$

By plotting the average network of the two groups in both discussions, we found that the Acupuncture concept of the F group and the other four elements (clinical operations, clinical concepts, reasoning and argumentation, TCM concepts) constitute the TCM competency model, and in these four connections in the features (see Figure 1-b and Figure 1-c). Further subtracting the two average networks, we found that although there was no statistical difference between the two networks, the concept of acupuncture and clinical concepts were more connected in the first discussion (red line in Figure 1-f), while

acupuncture Concepts, TCM concepts and clinical operations are more connected in the second discussion (blue line in Figure 1-f).

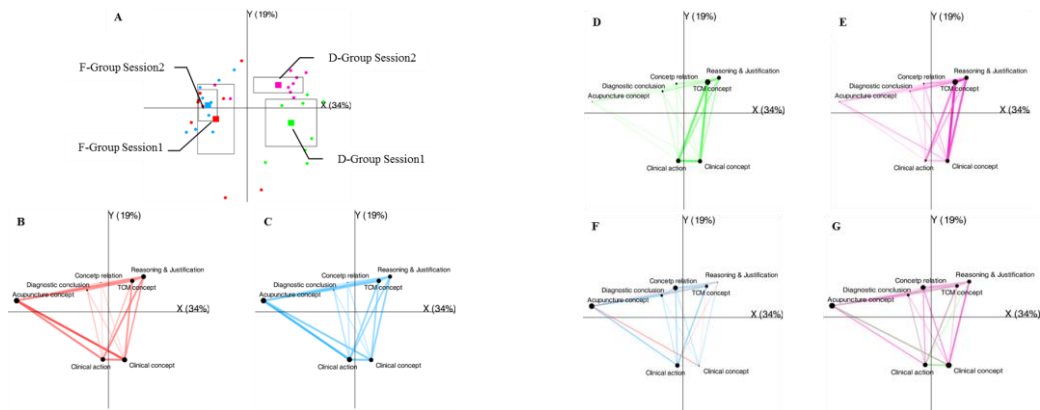


Figure 1. ENA results of two groups in two sessions. A Plotted point graph of two groups in two sessions; **B** Mean network of F-Group in the first session; **C** Mean network of F-Group in the second session; **D** Mean network of D-Group in the first session; **E** Mean network of D-Group in the second session; **F** Subtract network of F-Group between the first and the second sessions; **G** Subtract network of D-Group between the first and the second sessions.

For the D group, the network results show that the TCM ability model has more links with the four concepts of TCM concept, reasoning and argumentation, clinical concept and clinical operation, but has no obvious relationship with the concept of acupuncture (see Figure 1-d). And Figure 1-e). Subtracting the two average networks, the results show that there is more connection between clinical operations and clinical concepts in the first discussion (green line in Figure 1-g), acupuncture concepts, clinical concepts in the second discussion There is more connection between Chinese medicine concepts and significant changes (the pink line in Figure 1-g).

4.2 Guidance Modes of Two PBL Group Teachers

In order to explore the PBL regulation mode of the two teachers, ENA analysis was conducted on the data encoded according to the regulation strategy (see figure 2). The results showed that the guidance pattern of teachers in group F changed in the first and second discussions. In the first discussion, the teacher adopted specific guiding adjustment strategies, whose purpose was not only to cultivate students' understanding of TCM concepts, but also to cultivate other TCM abilities, such as reasoning and argumentation, clinical concepts and acupuncture concepts. In the second discussion, teachers adopted more diversified regulation strategies to cultivate students' TCM ability, including monitoring, positioning and explanation. However, teachers in group D adopted a similar regulatory mode in both discussions, namely, using directional strategies to promote students' understanding of TCM concepts and reasoning in clinical practice.

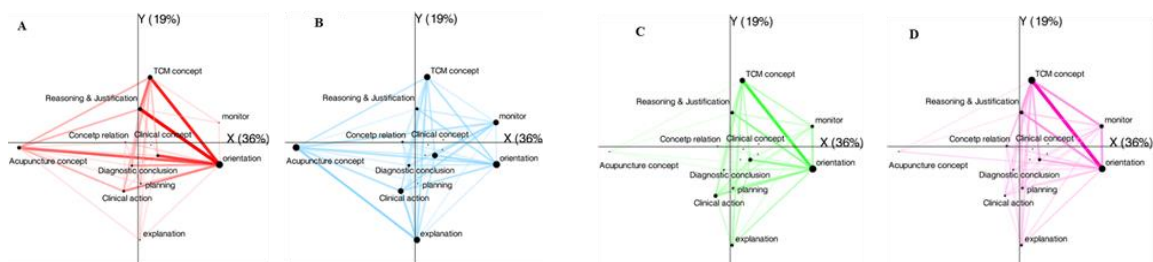


Figure 2. Two facilitators' PBL regulation networks. A Network of F-Group tutor in the first session; B Network of F-Group tutor in the second session; C Network of D-Group tutor in the first session; and D Network of D-Group tutor in the second session.

5 DISCUSSION

In order to answer the first research question about the difference in ability between the two groups, ENA results showed that group F had a more balanced knowledge structure between western problem understanding and TCM problem solving than group D. Specifically, group F focused more on acupuncture and moxibustion thinking to explain clinical manifestations. The slight change in the mean network from the first to the second time is also consistent with the objective of clinical diagnosis (TCM concepts related to clinical concepts) to clinical treatment of the tc-pbl process (acupuncture concepts related to clinical applications). On the contrary, compared with the understanding of acupuncture and moxibustion, group D paid more attention to solving clinical problems, and shifted the ability development from the understanding of western problems to the understanding of traditional Chinese medicine, while ignoring the thinking of acupuncture and moxibustion, which was far behind the second learning progress of group F.

Discourse analysis of hypothetical deductive reasoning examples further confirms the quantitative research results. We found that groups F and D showed different patterns of ability to discuss the same topic. Group F established more connections between western clinical concepts and acupuncture knowledge, considering not only diagnostic reasoning based on western clinical knowledge, clinical application and TCM knowledge, but also clinical treatment of meridian function (e.g., kunlun acupoint) and acupuncture methods (e.g., scalp acupuncture). Group D had more connections with clinical concepts and TCM concepts. But their ability to make connections between acupuncture and traditional Chinese medicine or clinical concepts is limited and extends beyond diagnostic reasoning to clinical applications and treatments. These findings are consistent with previous studies on PBL in western medicine education, suggesting that the stronger the association between different ability components, the better professional performance.

With regard to the development of TCM competence, our research shows that students randomly assigned to two TCM colleges with different teaching styles as tutors have developed distinct TCM competence patterns. Traditional Chinese medicine and acupuncture thinking, such as Yin, Yang, qi, channels and acupoints, are common images or patterns of a person's existence and behavior. They are basic patterns for detecting and synthesizing clinical information, and these patterns also create unique medical thinking processes. This study shows that in order to develop the ability of TCM, students should establish a comprehensive relationship between different elements, which should not be limited to the fields of TCM and acupuncture, but should combine western medicine and TCM/acupuncture. This is consistent with Mei's view that Ben pattern recognition and disease can provide guiding principles for the integration of TCM and western medicine.

In order to answer the second research question, we found that the mentors in group F promoted the application of various TCM competence components and adopted more flexible regulation strategies, while the mentors in group D guided the discussion of PBL in a more monotonous and directional way. This study extends previous studies on PBL counseling to further explore the relationship between different regulatory patterns and professional competence. The results of this study indicate that discussions among group F students show a more balanced TCM capacity network on two dimensions of projected space and more frequent use of acupuncture and moxibustion thinking to solve clinical

problems. The discussion among the students in group D focused more on the treatment of western medicine, shifting from the understanding of clinical problems to the superficial knowledge of traditional Chinese medicine, but not having a deeper connection with the knowledge of acupuncture and moxibustion.

By exploring the modality of the two mentors, the study provides empirical evidence that contradicts previous findings and supports mentoring. The possible reason is that TCM structures are less healthy than the introductory statistics in the background of Bude et al. For non-standard fields, the main learning objective of PBL should be to face knowledge construction rather than problem solving, while the auxiliary tutoring enables students to jointly build complex knowledge. With regard to the effective mentoring of PBL, these findings mean that mentors should limit the low-productivity start-response-feedback (IRF) sequence, but at the same time provide independent support and regulatory framework to build TCM capacity development. The f-group mentor's regulatory network represents the concept of adaptive teaching, which requires improvisation and patchwork to match existing regulatory strategies in order to dynamically promote questioning, debate, conflict and misunderstanding during PBL discussions.

6 RESULTS

For a long time, PBL has been considered to improve the ability of domain knowledge to understand and solve problems. However, providing reliable evidence based on how professional knowledge develops in the group dynamics of PBL discourse and the impact of different coaching styles on professional development remains a challenge. This study attempts to address this challenge by quantifying ethnography using ENA to simulate TCM capabilities and the PBL discourse model of the facilitative and instructional counseling groups in the PBL discourse. This study not only reflects the views of more ethnographic research in medical education, but also discusses the detailed description of the discourse data to reveal the intrinsic link of TCM competence components established in the PBL process and the diversified regulation strategies to support different Chinese medicine practitioners. Ability component.

This study adopts an innovative cognitive network analysis method to analyze the PBL conversation content modeling students' TCM's thinking ability level and development mode in the field of non-constructive problems, and explore the different guiding styles of PBL tutors for students' ability development. Impact. The research not only reveals the structural relationship between the various dimensions of the disciplinary thinking established by the students in the PBL discussion process, but also the effectiveness of the diversified learning and control strategies to support the development of different thinking abilities, and demonstrates the cognitive network analysis method of quantifying ethnography. In the context of social learning such as PBL, qualitative big data analysis is used to model and measure the great potential of students' thinking ability.

REFERENCES

- Aarnio M, Lindblom-Ylänne S, Nieminen J, et al. How do tutors intervene when conflicts on knowledge arise in tutorial groups? *Advances in Health Sciences Education*. 2014;19(3):329-345.
- Bligh, J. (1995). Problem-based learning in medicine: an introduction. *Postgraduate Medical Journal*, 71(836), 323-326.

- BudéL, van de Wiel MW, Imbos T, et al. The effect of directive tutor guidance on students' conceptual understanding of statistics in problem-based learning. *British Journal of Educational Psychology*. 2011;81(2):309-324.
- Chesler, N. C., Arastoopour, G., D'Angelo, C. M., et al. Design of professional practice simulator for educating and motivating first-year engineering students [J]. *Advances in Engineering Education*, 2013: 3 (3), 1–29.
- Chi, M. T. H., & Glaser, R. (1985). Problem-solving ability. *Human Abilities An Information Processing Approach*, 27.
- Hmelosilver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-Based Learning*, 1(1).
- Hong, N. S. . (1998). The relationship between well-structured and ill-structured problem-solving in multimedia simulation. *Educational Psychology Education*.
- Landauer TK, McNamara DS, Dennis S, et al. Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum; 2007.
- Leung WC. Why is evidence from ethnographic and discourse research needed in medical education: the case of problem-based learning. *Medical Teacher*. 2002;24(2):169-172.
- Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*. 1996;28(2):203-208.
- Nash, P., & Shaffer, D. W. Mentor modeling: The internalization of modeled professional thinking in an epistemic game [J]. *Journal of Computer Assisted Learning*, 2011: 27(2), 173–189.
- Papinczak T, Tunny T, Young L. Conducting the symphony: a qualitative study of facilitation in problem-based learning tutorials. *Medical education*. 2009;43(4):377-383.
- Shaffer DW, Collier W, Ruis AR. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*. 2016;3(3):9-45.
- Vella, F. . (2010). The tutorial process: by hs barrows, pp 63. southern illinois university school of medicine, springfield, il. 1988. \$12.95: isbn 0-931369-22-3. *Biochemistry & Molecular Biology Education*, 17(3), 161-162.
- Ward, W. C., Carlson, S. B., & Woisetschlaeger, E. (1983). Ill-structured problems as multiple-choice items. *Ets Research Report*, 1983(1), i–23.

Transitional, sequential patterns of students' knowledge inquiry and construction processes within online discussions

Fan Ouyang

Zhejiang University

fanouyang@zju.edu.cn

ABSTRACT: As collaborative learning is constituted through interactive, sustained, and dynamic dialogues over time, temporality inevitably matters for the analysis of collaborative learning. However, traditional “coding and counting” approach can only show the outcome of individual and group knowledge at a given point of time; temporal information is abandoned during the process. To show the temporal aspect of collaborative learning, this study used a combination of three learning analytics methods - content analysis, lag-sequential analysis, as well as social network visualization technique - to uncover the transitional, sequential patterns of students' knowledge advancement in online discussions. Results indicated a transitional, sequential patterns, moving from lower-level to higher-level knowledge advancement in both the individual and group levels. I hope this work serves as a trigger point for researchers to develop advanced, integrated learning analytics methods and representations in order to further demonstrate the temporal aspect of collaborative learning.

Keywords: Temporality; Learning analytics; Content analysis; Lag-sequential analysis; Social network visualization

1 INTRODUCTION

Effective collaborative learning is constituted through interactive, dynamic, and sustained dialogues over time (Chen, Resendes, Chai & Hong, 2017). The importance of time in collaborative learning has been emphasized in earlier work (e.g., Kapur, 2011; Knight, Wise, Chen, & Cheng, 2015; Reimann, 2009). To unpack the trajectory of collaborative learning, it is necessary to examine the temporal, sequential relationships between students' knowledge advancement.

As a quantified measure, content analysis has been broadly employed to examine students' collaborative learning behaviors, by analyzing data derived from learning activities according to content analysis coding schemes or frameworks (De Wever, Schellens, Valcke, & Van Keer, 2006). This type of content analysis usually takes traditional “coding and counting” approach to assess the outcome of individual and group knowledge at a given point of time. This summative way of analysis made data aggregated over time; in other words, the temporal information is abandoned during the process. Therefore, content analysis alone, without taking into account the temporality aspect, is not likely to elucidate sequential relationships of students' knowledge advancement (Chen et al., 2017). It is necessary to develop more integrated learning analytic methods to understand temporality of collaborative learning.

This study filled this gap. This study combined content analysis with lag-sequential analysis and social network visualization to uncover the transitional, sequential patterns of students' knowledge

advancement that may overshadowed by merely using summative, “coding and counting” content analysis method. The results demonstrated a transitional, sequential patterns, moving from lower-level to higher-level knowledge advancement in students’ collaborative learning process. Moreover, I hope this work can trigger further development of advanced, integrated learning analytics methods and representations in order to further demonstrate the temporal aspect of collaborative learning.

2 THE CURRENT STUDY

2.1 Research purposes and question

The research purposes were twofold: first, I aimed to understand transitional, sequential patterns of students’ collaborative learning during online discussions; second, I aimed to develop an integrated learning analytics method to better demonstrate the temporal aspect of collaborative learning. My research question was *What were the transitional, sequential patterns of students’ knowledge inquiry and knowledge construction within online discussions?*

2.2 Research context

The research context was a graduate-level semester-long course offered at a midwestern research university in the United States. This course - *Online Learning Communities*, focused on examining theories of online learning communities and practices of building online learning communities. Twenty graduate students (Female=16, Male=4) enrolled in this online course during a 14-week semester in spring 2014. Students were instructors, educators and practitioners from K-12, higher education and professional learning contexts. The course was primarily comprised of collaborative, inquiry-based discussions (see Figure 1). Students put forth ideas and perspectives, proposed and answered questions, and built on others’ ideas. Dataset for this current study was comprised of all class-level asynchronous discussions, including three instructor-designed, and three student-designed discussions. During those six discussions, students contributed to 8 initial posts, 131 initial comments, and 386 peer responses.



Figure 1: A part of a discussion thread in Ning forum

2.3 Research methods

The integrated learning analytics combined content analysis, lag-sequential analysis, as well as social network visualization technique to demonstrate transitional, sequential patterns of students' knowledge advancement in online discussions. Content analysis was first used to code students' knowledge advancement in discussion thread from a time series perspective; then lag-sequential analysis was used to examine the transitional relations between students' knowledge advancement; and finally, social network visualization technique was used to visually demonstrate the transitional relations (including strength and direction) in network formats. Together, unlike traditional "coding and counting" approach, this integrated method demonstrated temporal information of students' knowledge advancement.

2.3.1 Content Analysis

Content analysis was used to examine students' knowledge advancement in the individual and group levels. Adapting "Argumentation" and "Responsiveness" categories from the "speaking variables" coding scheme (Wise, Hausknecht, & Zhao, 2014), this study proposed a content analysis framework that includes a three-level "Knowledge Inquiry" category, and a three-level "Knowledge Construction" category (see Table 1). "Knowledge Inquiry" demonstrates individual knowledge inquiry within students' initial comments, and "Knowledge Construction" demonstrates group knowledge advancement processes within students' peer responses. Two raters coded 30% of the dataset independently and then had multiple meetings to discuss unit of analysis, resolve discrepancies, and adjust the coding scheme. After we reached an agreement of the coding scheme and unit of meaning, we separated the dataset and coded them independently. Inter-rater reliability was calculated for each level in terms of Cohen's Kappa: SKI: $k=0.945$; MKI: $k=0.910$; DKI: 0.920 ; SKC: $k=0.945$; MKC: $k=0.930$; and DKC: $k=0.905$.

Table 1: The cognitive engagement framework (adapted from Wise et al., 2014)

Category	Code	Level	Description
Knowledge Inquiry	Superficial-level Knowledge Inquiry (<i>SKI</i>)	1	A participant explores information without explicit statements of his/her own perspectives.
	Medium-level Knowledge Inquiry (<i>MKI</i>)	2	A participant presents his/her own perspectives without detailed elaborations.
	Deep-level Knowledge Inquiry (<i>DKI</i>)	3	A participant explicitly elaborates his/her own perspectives with detailed explanations, supports of resources, statistics.
Knowledge Construction	Superficial-level Knowledge Construction (<i>SKC</i>)	1	A participant simply presents (dis)agreement, without explicit statement of his/her own perspectives.
	Medium-level Knowledge Construction (<i>MKC</i>)	2	A participant extends another participant's perspectives, with detailed explanations, supports of information, statistics.

Deep-level Construction (<i>DKC</i>)	Knowledge 3	A participant extends, connects and deepens the ideas proposed by other participants, with detailed explanations.
---	-------------	---

2.3.2 Lag-sequential Analysis

Grounded upon content analysis results, lag-sequential analysis (LsA) was used to examine the transitional relations among these six code categories. LsA is a statistical method for identifying sequential contingencies of behaviors or events. Complementary to “coding and counting” measures in content analysis, LsA can examine transitional relations between different code categories and reveal temporal relations of those categories. An R package named *LagSeq* (Chen, 2015) was used to examine immediate transitions between two code categories based on three measures: *transitional frequencies*, *Yule’s Q scores*, and *adjusted residuals - Z scores*. *Transitional frequencies* among six code categories represented the number of times a code category transitioned immediately to another code category (e.g., *MKI*→*DKI*); *Yule’s Q scores*, namely the standardized measure, denoted strength of association between two code categories ranging from -1 to +1, with 0 indicating no association; *adjusted residuals - Z scores* represented the statistical significance of particular transitions (*Z scores* greater than 1.96 meant that the transitional sequence reached statistical significance $p<.05$) (see Chen et al., 2017).

2.3.3 Social network visualization

Finally, the weighted, directed social network visualization technique was used to visualize the transitional sequence networks. In the networks, the node size represented the frequency of code categories, tie strength represented relation strength and tie direction represented the transitional directions between code categories.

3 RESULTS

First, content analysis results indicated a total of 787 knowledge inquiry and knowledge construction codes, consisting of 255 *DKI* codes, 239 *MKC* codes, 158 *MKI* codes, 87 *SKC* codes, 36 *DKC* codes, and 12 *SKI* codes. Second, lag-sequential analysis results showed that, except the sequences between the same code category, the highest *transitional frequency* of sequences occurred between *MKI* to *DKI* (*transitional frequencies*=83), followed by *DKI* to *MKI* (*transitional frequencies*=58), and *DKI* to *MKC* (*transitional frequencies*=55). The highest *yule’s Q scores* occurred between *MKC* to *DKC* (*yule’s Q*=0.68), followed by *SKI* to *MKI* (*yule’s Q*=0.61), and *MKI* to *DKI* (*yule’s Q*=0.50). The highest *Z scores* occurred between *MKI* to *DKI* (*Z score*=6.14), followed by *MKC* to *DKC* (*Z score*=5.03), and *SKC* to *MKC* (*Z score*=4.68). Finally, transitional sequence networks demonstrated both the strength and direction of the relations between code categories, based on *transitional frequencies*, *yule’s Q scores* as well as *adjusted residuals - Z scores* (see Figure 2). Again, the three transitional sequence networks together visually demonstrated that there were significant transitions from *MKI* to *DKI*, *MKC* to *DKC* and *SKC* to *MKC*.

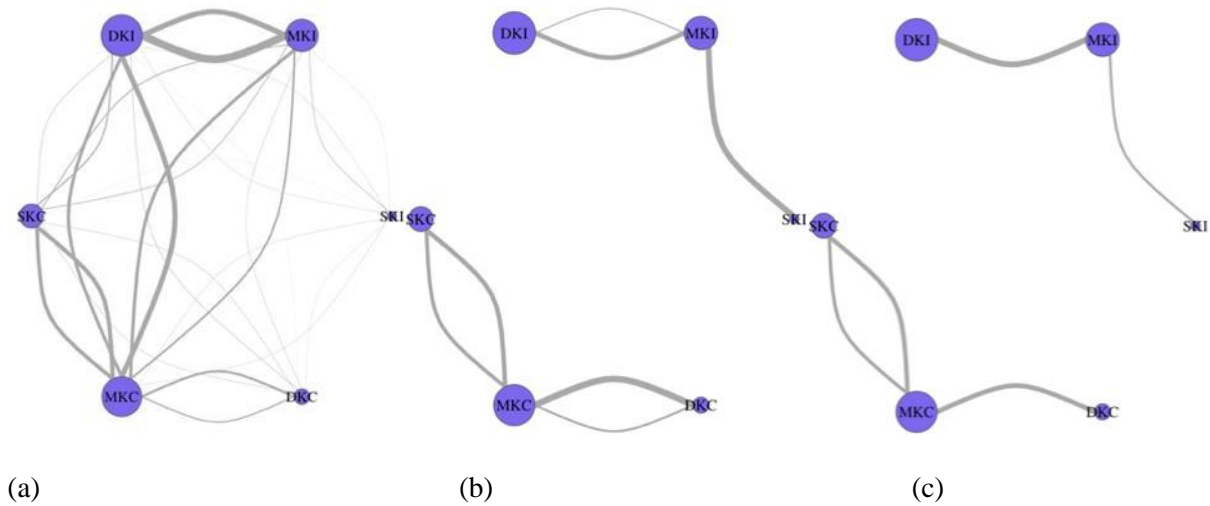


Figure 2: Transitional sequence networks based on transitional frequencies (a), yule's Q scores (b), and adjusted residuals - Z scores (c)

Note. For simplicity, the direction arrows were hidden; the directions should be interpreted clockwise.

4 DISCUSSIONS AND IMPLICATIONS

There was a progressive knowledge advancement process in both the individual and group levels, respectively. Results showed significant transitions from *MKI* to *DKI*, *MKC* to *DKC* and *SKC* to *MKC*, which indicated a transitional, sequential patterns, moving from lower-level to higher-level knowledge advancement in both the individual and group levels. Consistent with previous research results (Cress, Held, & Kimmerle, 2013; Ouyang & Chang, 2018; Zhang, Liu, Chen, Wang, & Huang, 2017), this study showed a progressive development process between individual students' knowledge inquiry and group knowledge construction. In other words, lower-level students' knowledge inquiry and construction in the individual and group level are prerequisite for deeper-level knowledge advancement.

Moreover, results also showed a significant transition from *DKI* to *MKC* across the individual and group levels, which indicated a sequential relation from deep-level individual knowledge advancement to group knowledge construction. Therefore, it was more likely for students to build up group-level knowledge when they individually contributed to deep-level knowledge advancement.

Based on the results, I provided both pedagogical and analytical implications. First, from the pedagogical perspective, instructors should not only encourage students make advanced, deeper-level contributions to individual knowledge inquiry and group knowledge construction, but also recognize students' superficial contribution in knowledge since it is more likely for students to move toward higher-level knowledge advancement based on lower-level knowledge inquiry. In addition, from an analytical perspective, this study showed an integrated way to analyze and demonstrate transitional, sequential patterns of student collaborative learning processes that may be overshadowed by merely using traditional "coding and counting" content analysis methods. Furthermore, this study encourages researchers to design and develop advanced, integrated learning analytics methods and representations in order to further demonstrate the temporal aspect of collaborative learning.

In conclusion, this work used an integrated learning analytics method - combining content analysis, lag-sequential analysis, as well as social network visualization technique - to uncover transitional, sequential relations on students' knowledge advancement in online discussions. Results indicated a

transitional, sequential patterns, moving from lower-level to higher-level knowledge advancement in both the individual and group levels. I hope this work can trigger further development of temporality-oriented learning analytics and visualized representations.

ACKNOWLEDGEMENT

I appreciate the assistance from Yu-Hui Chang, another rater of content analysis in this study.

REFERENCES

- Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology and Society*, 15(3), 3–26.
- Chen, B. (2015). LagSeq: R Implementation of Lag-sequential Analysis (version 0.0.0.9000) [R package].
Retrieved from <https://github.com/meefen/LagSeq>
- Chen, B., Knight, S., and Wise, A. F. (2018). Critical Issues in Designing and Implementing Temporal Analytics. *Journal of Learning Analytics*, 5(1), 1-9.
- Chen, B., Resendes, M., Chai, C. S., & Hong, H. Y. (2017). Two tales of time: uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interactive Learning Environments*, 25(2), 162-175.
- Cress, U., Held, C. & Kimmerle, J. (2013). The collective knowledge of social tags: Direct and indirect influences on navigation, learning, and information processing. *Computers & Education*, 60, 59– 73.
- De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, 46(1), 6– 28.
- Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning*, 6(1), 39–56.
- Knight, S., Wise, A. F., Chen, B., & Cheng, B. H. (2015). It's about time: 4th international workshop on temporal analyses of learning data. In *Proceedings of the fifth international conference on learning analytics and knowledge - LAK '15* (pp. 388–389). New York, NY: ACM.
- Ouyang, F. & Chang, Y. H. (2018). The relationship between social participatory role and cognitive engagement level in online discussions. *British Journal of Educational Technology*. [Online Version of Record]
- Reimann, P. (2009). Time is precious: Variable-and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, 4(3), 239-257.
- Wise, A. F., Hausknecht, S. N., & Zhao, Y. (2014). Attending to others' posts in asynchronous discussions: Learners' online "listening" and its relationship to speaking. *International Journal of Computer-Supported Collaborative Learning*, 9(2), 185-209.
- Zhang, S., Liu, Q., Chen, W., Wang, Q. & Huang, Z. (2017). Interactive networks and social knowledge construction behavioral patterns in primary school teachers' online collaborative learning activities. *Computers & Education*, 104, 1–17.

SASLAS19: Scalability and Sustainability of Learning Analytics Solutions.

Tom Broos
KU Leuven
tom.broos@kuleuven.be

Dragan Gašević
Monash University

Abelardo Pardo
University of South Australia

Hendrik Drachsler
Goethe University Frankfurt am Main – DIPF

Rafael Ferreira
Universidade Federal Rural de Pernambuco

Katrien Verbert
KU Leuven

Tinne De Laet
KU Leuven

ABSTRACT: Research about Learning Analytics has increasingly gained attention, as demonstrated by the geographic and substantive scope of LAK. However, as the domain of LA is maturing, the connection of research to long-term applicability is relatively underdeveloped. This may hinder further investment of policy makers and administrators. The goal of our half-day workshop is to explore and discuss the scalability and sustainability of existing and proposed solutions, and to initiate the creation of a framework of strategies available to researchers and practitioners.

Keywords: learning analytics, sustainability, scalability

1 WORKSHOP BACKGROUND

While Learning Analytics (LA) is still a relatively young discipline, it is quickly expanding, both in substantive scope and geographic interest. At each edition of the LAK conference, several promising results are being shared. However, **in many cases LA tools demonstrate difficulties in making the transition from research artefacts into scalable solutions in real-life educational contexts.** Research papers generally do not address the issues of scalability and sustainability of proposed solutions extensively, if at all, leaving practitioners with unclear guidelines to apply them in non-experimental settings. Therefore, while the domain of LA is maturing, we posit that the connection from research

to long-term applicability is relatively underdeveloped. This may hinder further investment of policy makers and administrators.

In this workshop, we want to discuss the issues and opportunities of scalability and sustainability on several dimensions, including:

1. **Generalizability:** not uncommonly, LA research takes place in favorable settings, e.g. involving a researcher-teacher with detailed knowledge of the specific course, or other highly motivated stakeholders. While an experiment-friendly context may be a valuable incubator for innovative LA solutions, it does not test or harden them for real-life applicability at scale. We would like to invite researchers to address this issue when presenting their own work, or to start from existing work to explore its reproducibility in challenging contexts.
2. **Return on investment:** several authors have raised questions about the impact of LA applications on learning (e.g. Dyckhoff et al. 2013, Dawson et al. 2017), something that may be difficult to measure. However, it has recently been argued that impact is only part of the equation when making a business case (Broos et al. 2018). As LA projects are likely to end up competing for resources with other proposals, LA researchers need to include return-on-investment (ROI) in their reasoning. LA solutions that require only limited effort can be attractive, even if the expected impact is relatively low or even uncertain. Vice versa, LA projects that would require significant investment will be challenged with higher expectations. The workshop aims at creating awareness in the LA community to this consideration.
3. **Change management:** even if issues of generalizability and ROI are addressed by LA projects, chances of sustainable and scalable implementations are limited without acceptance of learners, teachers and other stakeholders. Even the best models and feedback tools are of little use if they are not acted upon due to a lack of trust or willingness. Therefore, LA needs to address transparency, openness and understanding of user acceptance. Underestimation of the importance of institutional culture, resistance to innovation and the role of change management poses a big treat for success of LA within institutions (Macfadyen & Dawson 2012). Many lessons learned in general change management should not be ignored by the LA community and several change management frameworks are available for reuse. The ADKAR model, for instance, provides insight into five stages: awareness, desire, knowledge, ability and reinforcement (Hiatt 2006). Similarly, several maturity assessment models have been developed in management science and information systems literature. It has been argued that institutions should build their LA maturity layer by layer, starting with modest implementations (Broos 2017).

2 ORGANIZATIONAL DETAILS

2.1 Type of event

The **half-day workshop** includes consecutive phases of increasing participant activity, including jigsaw group work sessions, collaborative discussions and flipped presentations of other participants work.

2.2 Proposed schedule

- First the organizers will set the scene with a short introduction (max. 25 minutes).
- Next, participants will be grouped across submitted work, allowing respective authors to briefly explain their work with a focus on scalability and sustainability (max. 25 minutes).
- The group then delegates one person to present another's work to the entire audience using an interactive demo or mockup of the solution, critically reflecting within the theme of the workshop (max. 5 x 10 minutes).
- New (jigsaw) groups will be formed to discuss and co-develop a framework of strategies to include the scalability and sustainability aspects in LA. Previously presented work will be mapped on these strategies – identifying strengths and weaknesses (max. 2 x 25 minutes).
- In conclusion, each group will present their framework and mappings to the entire audience, while a facilitator will merge the different outcomes in a shared representation for participants to discuss and comment on (max. 5 x 5 minutes).
- An observer appointed at the beginning will summarize the workshop (max. 10 minutes).

Based on the actual number of accepted submissions and participants, the scenario of the workshop will be adopted. Several scenarios will be prepared on beforehand to guarantee the workshop's usefulness and quality in case of low or high attendance. Several co-hosts will be present to facilitate the workshop and to ask probing questions and introduce statements during group discussions

2.3 Target group and recruiting of participants

The first target group of the workshop are researchers that are actively publishing about LA solutions. The workshop invites them to assess their own or other's work from a scalability and sustainability perspective and provides them with a contribution channel to extend previous studies. For this purpose, participants are invited to submit a paper. In accordance with the workshop format, these papers will be accompanied by demo material, mockups or other artefacts to allow the flipped presentation by another participant (cf. supra). Submissions will be reviewed by a committed team of senior and junior researchers.

The second target group are policy makers, practitioners, student representatives, managers, and other stakeholders that either have hands-on experience with successful or unsuccessful implementations of LA at scale, or are exploring the opportunities. They will be invited to participate in the workshop discussions with a critical but constructive view.

Both target groups will be informed about the upcoming workshop through a dedicated website and invited using social media and the organizers' networks.

2.4 Required equipment

The workshop assumes availability of a HD-projector, a white-board and/or flipchart and large sheets of paper and markers (preferably multiple flipcharts or adhesive paper sheets).

3 OBJECTIVES AND PLANNED OUTCOMES

The intended outcome of the workshop is threefold:

1. The call-for-papers will invite participants to tackle the concern of scalability and sustainability of LA explicitly in their submissions. We welcome papers building on existing work, extending it by addressing the theme of the workshop, which was often overlooked in previous publications.
2. Participants will contribute to the ideation phase of a reusable framework of strategies to include the scalability and sustainability aspects in LA. It is intended to continue building this framework after the workshop and to present it to the community.
3. The ultimate intention of the workshop is to contribute to the domain of LA by creating awareness of long-term applicability at scale, opening the discussion, and identifying points of future work. To further concretize this, three points of immediate action will be agreed upon in conclusion of the workshop.

4 ABOUT THE ORGANIZERS

The workshop will be co-hosted by a group of researchers that have recently voiced their concerns about scalability or sustainability of LA and/or have shared concrete steps of how they accounted for the matter when applying their work at larger scale.

REFERENCES

- Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2017). Small data as a conversation starter for learning analytics: Exam results dashboard for first-year students in higher education. *Journal of Research in Innovative Teaching & Learning*, 10(2), 94-106.
- Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2018, September). Low-Investment, Realistic-Return Business Cases for Learning Analytics Dashboards: Leveraging Usage Data and Microinteractions. In *European Conference on Technology Enhanced Learning* (pp. 399-405). Springer, Cham.
- Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017, March). From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 474-478). ACM.
- Dyckhoff, A. L., Lukarov, V., Muslim, A., Chatti, M. A., & Schroeder, U. (2013, April). Supporting action research with learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 220-229). ACM.
- Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, 15(3).
- Hiatt, J. (2006). ADKAR: A model for change in business, government, and our community (1st ed.). Loveland, CO: Prosci Learning Center Publications.

Not whether, but where: Scaling-up how we think about effects and relationships in natural educational contexts

Benjamin A. Motz
Indiana University
bmotz@indiana.edu

Paulo F. Carvalho
Carnegie Mellon University
pcarvalh@andrew.cmu.edu

ABSTRACT: This paper presents a brief discussion of “effects” and “relationships” in authentic educational contexts, and endeavors to scale-up our thinking about the meaning of these constructs. To discover the mere presence of a reliable main effect relating two variables in natural educational practice is often a feeble pursuit, for any effect might be observable in variable contexts with a sufficiently narrow analysis plan or with a sufficiently large sample size. In turn, this paper argues that researchers should place less emphasis on the mere discovery of relationships, and more emphasis on the analysis of the generalizability of these relationships, the ways that the relationships under investigation may interact with educationally-relevant covariates, and the identification of authentic edge cases where an expected relationship may disappear or reverse.

Keywords: generalizability, interaction effects, meta-analysis

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!
- From *The Blind Men and the Elephant* (Saxe, 1873; p. 260)

Under just the right conditions in natural educational settings, it is possible that any variable could be associated with significant changes, in either direction, for students’ learning outcomes. For example, research into the duration of inactivity in a course site (Conjin et al., 2017), the access of assignments after the deadline (Motz et al., 2019), the order of exemplars during study (Carvalho & Goldstone, 2017), and the immersiveness of instructional examples (Day, Motz, & Goldstone, 2015), have all found opposing benefits in different contexts. Whether a researcher observes positive evidence of such an effect, fails to observe a significant effect, or observes the opposite effect, may be principally determined by the scope of the researcher’s analysis, and not by whether the effect “exists.” Like the ancient parable of blind men developing opposing theories of a single elephant (e.g., Saxe, 1873), analytical research on student learning risks a similarly-absurd dispute about the observation of effects (or lacks thereof) in isolated studies, and what these opposing observations might mean.

The goal of this essay is to recommend a shift in thinking about “effects” and “relationships” as observed in authentic educational contexts, moving past thinking of these in binary terms (they are or aren’t observed; or they do or don’t replicate), to thinking of these as existing in varying degrees in different contexts. There is no single relationship between educationally-relevant variables that would hold constant across all learners and learning environments. The question for those analyzing data from authentic educational environments should not be *whether* such relationships exist, but instead, *where* they exist, to what degree. Furthermore, educational research, including learning analytics, must exist in the context of strong theories and models of learner’s cognition that can predict and explain why these dependencies exist, toward proposals of interventions that can leverage dependencies, instead of being hampered by them.

1 ANY EFFECT MIGHT BE PRESENT IN SOME CLASSROOM

Authentic classrooms are not randomly sampled from the space of all possible educational dimensions. Curriculum and course structures are engineered by teachers, administrators, faculty committees, software designers, and textbook publishers to produce positive gains for enrolled students. Rather than being random points in the multidimensional landscape of educational contexts, classrooms are *architected learning factories*; courses are designed in just such a way so that learning activities, the behaviors of the instructor, the supporting materials, and the surrounding environment all shuttle the enrolled students in the direction of positive learning outcomes. For example, teachers who assign weekly graded practice quizzes are crafting fundamentally different systems than teachers who assign ungraded weekly practice quizzes. The differences between these classes are not limited only to this single dimension of whether the weekly quizzes are credited or not. Both could be reasonably beneficial design solutions in different contexts. Just as the same musical note can elicit different emotions in different chords, any educationally-relevant variable could be inconsequential, or could be engineered to benefit learning, in different classrooms.

When one accepts that classes are not randomly drawn instances from some grand educational roulette wheel, two corollaries follow: (1) Any naturally-occurring variable may be architected in an educational context so as to produce a larger effect on learning outcomes, β , than the same variable’s effect in a different context. And thus (2) the measurement of effect β in an authentic classroom is an interaction between the variable under analysis and the class’s other covariates, not a main effect that should be expected to generalize across contexts.

Let’s consider an example. Imagine that an intrepid team of researchers aims to examine the relationship between some variable, perhaps class attendance, and learning outcomes. They aggregate attendance records and final exam scores for a large course whose data were convenient to access. If the observed effect of attendance on exam performance is 0%, 0.1%, -1% or 10%, what might they claim in these scenarios? Surely these are not generalizable estimates of the effect that attendance *could* have on learning performance in other classes (what if students had no access to learning materials outside class? —or what if the class activities only involved review of take-home readings?) as was compellingly demonstrated by Gašević et al. (2016). That any particular main effect is observed for any limited sample is rather unremarkable, because the estimate of that effect is determined largely by the context in which it is measured. Indeed, in the context of course design,

if a *teacher* (not the research team) finds that attendance is not related to learning outcomes in the intended way, the teacher might change the relative value of attendance marks, increase active and collaborative problem solving in the classroom, or design other contextual modifications rather than simply conclude that attendance doesn't "work." The intrepid research team should also avoid the latter conclusion, which would be a severe out-of-sample overgeneralization.

The concerns discussed thus far are sometimes cast as criticisms against the broader research enterprise of mining and analyzing authentic learning data (for a discussion, see Morrison & van der Werf, 2016). But just as the intrepid research team should avoid making overgeneralizations about effects from limited samples, so too should theorists avoid making overgeneralizations about a complex domain of applied research from its youthful foibles. On the one hand, analyses of a relationship in a limited sample could be a very fruitful activity when a teacher seeks to engage in more data-driven design solutions within that precise context (Halverson et al., 2007), or when a limited sample is highly representative of a conventional instructional system that is theoretically interesting or practically relevant, perhaps because of its applicability to specific goals of education (e.g., 9th grade Algebra 1 or Introductory Chemistry recitations as gateways to STEM disciplines). But on the other hand, the broader activities of learning analytics, educational data mining, and other forms of education research utilizing big data could probably benefit from a reconsideration of how effects are analyzed and interpreted (see also Koedinger, Booth, & Klahr, 2013). Such reconsiderations may involve estimating effects separately for different kinds of courses (Motz et al., 2018c, 2019), developing new context-dependent theories of learning (Carvalho, 2018), and expanding the scope of experimental analyses to include a wide pool of independent samples (Motz et al., 2018b).

In the remaining sections, we attempt to motivate these reconsiderations by expanding on the possibility that any effect might be observed in some classroom, that thus, what may appear as main effects are more likely to be interaction effects, and then we discuss analytical tools that may scaffold a more robust and scalable perspective on effects and relationships in natural educational contexts.

2 ANY EFFECT MIGHT BE OBSERVED IN SOME CLASSROOM

When approaching a big dataset of natural behaviors, such as those increasingly available from e-learning environments, things will get messy. It might be tempting to view a theoretically-interesting effect or relationship as a needle in a haystack, but a more apt perspective might view the effect as a needle in a big stack of needles (which may also include some hay). There are no shortages of possible effects to be "discovered" during the analysis of a natural dataset, leading us to assert that in such a dataset, *any* effect might be observed (or might not be observed) in some subsample.

Consider the recent work of Silberzahn & Uhlmann (2015; et al. 2017), who recruited 29 different research teams to answer a single research question from a single dataset: Are soccer referees more likely to give penalty cards to dark-skin-toned players than light-skin-toned players? The dataset contained the full history of player/referee interactions for over 2,000 professional soccer players in four European countries, as well as the players' demographics, photos, classification of skin tone

(determined by independent raters), and a variety of additional covariates (team, position, etc.). The research analysts submitted their analytical plans (but withheld their provisional results) to a round-robin peer review and subsequently had the opportunity to revise their analyses. Nevertheless, despite this opportunity to converge on analytical approaches, final results varied widely among the participating researchers: effects ranged from 0.89 to 2.93 in odds ratio units (1.0 indicates no effect), with roughly two thirds of teams observing a significant effect, and one third finding no significant effect. The differences in outcomes resulted primarily from whether the analysis was sensitive to covariates and grouping variables present in the data.

While differences in analytical approaches will surely contribute to variability in measured effects, another factor influencing whether relationships are “discovered” is the size of the dataset. With increasing class sizes, and correspondingly increasing sample sizes, effects are more likely to fall beneath decision thresholds for statistical significance (commonly, the alpha-level), including spurious results and trivially small effects. For example, when analyzing the characteristics of digital camera auctions on eBay, Lin, Lucas, and Shmueli (2013) found that the magnitude of p -values in their analysis became meaninglessly close to zero when $n > 700$ (in a dataset containing over 300,000 observations). With a large enough sample, any scant difference is enough to claim statistical significance. In the case where an analyst might contrast two groups, A and B, Tukey (1991) observed, “The effects of A and B are always different - in some decimal place - for any A and B.” (p. 100) In this frame, whether someone detects an effect or relationship is really a question of sample size — and these days, behavioral researchers have access to some very large datasets. The observation of an effect is a fundamentally different issue from the relevance of an effect, leading many behavioral scientists toward new statistical standards concerned with effect *size* rather than effect *presence* (Serlin & Lapsley, 1985; Cumming, 2014).

The possibility of observing an “effect” is not only inflated by large samples and analytical variability; evidence for a spurious effect may also sprout in the soil of atheoretic exploratory analysis (Anderson et al., 2001). The paucity of theory in some applications of learning analytics and educational data mining yields fertile grounds for the discovery of effects and relationships that might be statistically-significant, but have no value for educational practice or for our understanding of educational systems (Wise & Shaffer, 2015). In the future, researchers will find ever-increasing opportunities to “discover” something practically meaningless as institutions continue to develop sprawling data warehouses to support as-yet-undefined future initiatives around learning analytics.

For an analyst who wonders whether an effect can be observed, the answer is surely “Yes.” In the absence of theory, in the presence of large datasets, and without clear methodological standards guiding our analytical plans, we should expect to find anything we want to find from natural educational data. In turn, researchers can benefit from a reconsideration of what is meant by the word “finding” in authentic learning contexts.

3 ANY EFFECT MIGHT BE AN INTERACTION EFFECT

Toward the goal of reconsidering what is meant by a “finding,” one useful tack might be to reimagine all *main* effects in our analyses as being *interaction* effects within educational systems. For the most part, educational research has embraced the existence of individual differences in

education. It is not controversial that different students will approach learning in a different way, and benefit differently from interventions. For example, Steyvers & Benjamin (2018) demonstrated that improvements in online brain training games interacts with the learner's age, and Kalyuga et al. (2003) demonstrated that low-knowledge students benefit more from studying worked examples than high-knowledge students.

However, this embrace of dependencies has not expanded to include the effects that different contexts (i.e., what is learned, how it is learned) have on the effectiveness of the same learning approach (Jonassen, 1982). Carvalho (2018) proposed that if we use learning theory to guide exploration of content-treatment interactions, we can not only gain a deeper understanding of the learning process, but also how it can be improved in a general and scalable way. Take, for example, the interleaving effect (see Dunlosky et al., 2013). By using an interaction design approach, Carvalho & Goldstone (2013) were able to demonstrate that the interleaved effect did not generalize to all learning materials. Moreover, and perhaps more importantly, their analyses propose a model of learning over time that can account for content-dependencies and suggests that learning does not always happen by discrimination, leading to clear predictions of when interleaved study will and will not improve learning (Carvalho & Goldstone, 2014; 2017).

Theories that embrace content-dependencies have great potential for learning analytics. If, when approaching a research question, one questions not if A “works,” but instead if A differs in context X vs Y, one can learn not only that A works, but also *why* it works. This is because interactions help us understand the mechanism by which A works — if A works in X but not in Y, what is about X that makes it work? However, it is important to note that interactions (albeit statistically less likely to be found than main effects, especially with large samples) are not always relevant. Interaction designs should be used with theory-building in mind, and not to dismiss theory by saying “it all depends,” which would be *reductio ad absurdum*. While every educational effect may depend on a contextual variable, many dependencies are generalizable and are relevant to practice and theory, which is why we advocate for a science that systematically examines *where* these effects exist.

That any effect might exist in some context, and that these effects are context-dependent may also be viewed as precipitants of Rossi's *Iron Law of Evaluation* (1987): “The expected value of any net impact assessment of any large-scale social program is zero” (p.4). If an educational intervention's relationship with learning outcomes is variable across different classes, at large scales the *aggregate* (net) benefit of an intervention will tend toward zero. Just as analysts ought to think critically about the discovery of an effect, so too should analysts be skeptical when measuring the absence of a reliable main effect at scale. Favorable conditions for an effect are unlikely to be universally-present across large samples of classrooms, and identifying the conditions for an effect's observation is an important pursuit if we are to make precise predictions about what “works.”

4 ANY EFFECT MIGHT BE SYNTHESIZED IN SOME CLASSROOM

Discussion thus far has been occupied with the discovery of effects during the observation of natural datasets, but another research method bears mentioning: the experimental manipulation of a variable to produce an effect. In laboratory studies, where the setting is artificial and the environmental regime is tightly-controlled according to experimental standards, there may be less

risk of variability in outcomes; indeed, laboratory studies are designed precisely so that the observed effects will replicate if all procedures are repeated with a new sample. But when conducting an embedded experiment in an authentic educational context (Motz et al., 2018a), the generalizability of an observed effect is much less certain.

In fairness, it should be noted that effects produced in embedded experiments have important advantages over effects found during the passive observation of natural datasets (Gordon et al., 2018). In particular, in an experiment, the context is held constant across experimental treatments, manipulating only those variables under analysis. However, the observed effect in that controlled context still may not be expected to generalize to different classes, because the size of any one measured effect is (as previously discussed) something that interacts with the structure of the class under observation.

Anecdotally, one of us recently discussed the design of an embedded experiment with another researcher, who was considering implementing the experiment in two of his sections during an upcoming semester. The researcher wanted to find a robust effect of his manipulation, so he was examining how he might structure the sections to facilitate this outcome. These considerations included: modifying the syllabus to highlight the experimental variable, emphasizing the variable with a take-home assignment, dedicating class time to a brief discussion of the variable, increasing the weight of grades more closely associated with the variable... At a certain point, we might wonder whether the observation of this effect would require an experiment in the first place! If a class can be architected to facilitate the observation of an effect, why should a researcher bother with the great effort and difficulty of demonstrating that effect?

For an effect observed in one class to be useful and generalizable, that class must be highly representative of a conventional instructional system that is theoretically interesting or practically relevant. Toward this goal, researchers should include documentation of the instructional context wherein an effect is observed. For example, in postsecondary learning environments, at minimum, authors should provide copies of class syllabi to accompany published reports from their embedded experiments, and moreover, they should highlight any course modifications made in support of the experimental contrast. But in keeping with the theme of this essay, rather than examining whether an effect is observed in a specific context, it might be more interesting to cast a wider net, examining *where* an experimental manipulation has different effects. But what might this “net” look like?

A scalable research model for evaluating experimental effects across a variety of authentic learning contexts is currently under development, called *ManyClasses* (Motz et al., 2018b). As with similar efforts in psychology (Many Labs, Klein et al., 2014; Many Babies, Frank et al., 2017), the core feature of *ManyClasses* is that researchers measure an experimental effect across many independent samples – in this case, across many classes. Rather than conducting an embedded learning study in just one educational context, a *ManyClasses* study would examine the same experimental contrast in dozens of contexts, spanning a range of courses, institutions, formats, and student populations. By inserting the same experimental manipulation across a diversity of educational implementations, and then analyzing pooled results, researchers can assess the degree to which an effect might yield benefits across a range of specific contexts. In addition to contributing to an estimation of the generalizable effect size of manipulations beyond particular

classroom implementations, a ManyClasses study will also systematically investigate how a manipulation might be more or less effective for different students in different situations.

This ManyClasses model shares common ground with a nascent analytical strategy called a *metastudy*, also used for analyzing the robustness of an empirical claim across contexts (Baribault et al., 2018). A metastudy involves the *radical randomization* of experimental design decisions; rather than fixing the study context across conditions (which might include the number of trials, properties of the stimuli, incentives for participating, etc.), these facets are randomly drawn for each observation. In turn, data obtained from a metastudy goes beyond addressing whether an effect exists, to directly estimating the contextual dependencies of the observed effect. By embracing the view that effects will vary across contexts, and directly manipulating and quantifying this variability, researchers can develop a much more complete understanding of the causal chains under analysis.

5 CONCLUSION

So, oft in theologic wars
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean;
And prate about an Elephant
Not one of them has seen!
- Final stanza from *The Blind Men and the Elephant* (Saxe, 1873)

Instructional technologists are oft to advertise new teaching and learning tools with the confident certification, “it works!” Data scientists implementing a new technique for predicting academic risk will claim, “our model works!” Psychologists examining students’ studying behaviors in a real class will conclude, “the strategy works!” In response, some skeptical and empirically-minded members of the education research community may scoff, “How do you know?” or “What is your evidence?” But all of these stances seem like non sequiturs, for any such instrument, activity, modeling approach, or strategy might “work” or might fail to “work” in different natural learning contexts. In scaling-up our perspective of these effects, perhaps learning analytics can avoid the dilemma of the blind men and the elephant, by accepting that different observations will necessarily yield different effects and relationships, and that these context-dependencies are theoretically-attractive objects of inquiry. In this paper, we hope to have motivated the view that *where* an effect exists in a real classroom, and to what degree, are much more meaningful concerns than *whether* that effect exists.

REFERENCES

- Anderson, D.R., Burnham, K.P., Gould, W.R., & Cherry, S. (2001). Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, 29(1), 311-316.
<https://www.jstor.org/stable/3784014>
- Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11), 2607-2612.
<https://doi.org/10.1073/pnas.1708285114>

- Carvalho, P.F. (2018). Understanding the dynamics of learning: The case for studying interactions. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 51-52). Cognitive Science Society.
- Carvalho, P.F., & Goldstone, R.L. (2013). How to present exemplars of several categories? Interleave during active learning and block during passive learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.
- Carvalho, P.F. & Goldstone, R.L. (2014). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, 5(936), 1-10. <https://doi.org/10.3389/fpsyg.2014.00936>
- Carvalho, P.F. & Goldstone, R.L. (2017). The most efficient sequence of study depends on the type of test. In G. Gunzelmann, A. Howes,, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp 198-203). Cognitive Science Society.
- Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle. *IEEE Transactions on Learning Technologies*, 10(1), 17-29. <https://doi.org/10.1109/TLT.2016.2616312>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7-29. <https://doi.org/10.1177/0956797613504966>
- Day, S. B., Motz, B. A., & Goldstone, R. L. (2015). The cognitive costs of context: The effects of concreteness and immersiveness in instructional examples. *Frontiers in Psychology*, 6, 1876. <https://doi.org/10.3389/fpsyg.2015.01876>
- Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- Frank, M., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory building. *Infancy*, 22, 421-435. <https://doi.org/10.1111/infa.12182>
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68-84. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- Gordon, B. R., Zettermeyer, F., Bhargava, N., & Chapsky, D. (2018). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Forthcoming at *Marketing Science*. Available at SSRN: <https://ssrn.com/abstract=3033144> or <http://dx.doi.org/10.2139/ssrn.3033144>
- Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2007). The new instructional leadership: Creating data-driven instructional systems in school. *Journal of School Leadership*, 17(2), 159.
- Jonassen, D. H. (1982). Aptitude-versus content-treatment interactions. *Journal of Instructional Development*, 5(4), 15. <https://doi.org/10.1007/BF02905228>
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23-31. https://doi.org/10.1207/S15326985EP3801_4
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142-152. <https://doi.org/10.1027/1864-9335/a000178>

- Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, 342(6161), 935-937. <https://doi.org/10.1126/science.1238056>
- Lin, M., Lucas Jr., H.C., Shmueli, G. (2013). Too big to fail: Large samples and the p -value problem. *Information Systems Research*, 24(4), 906-917. <https://doi.org/10.1287/isre.2013.0480>
- Morrison, K., & van der Werf, G. (2016). Large-scale data, “wicked problems,” and “what works” for educational policy making. *Educational Research and Evaluation*, 22(5/6), 255–259. <https://doi.org/10.1080/13803611.2016.1259789>
- Motz, B. A., Carvalho, P. F., de Leeuw, J. R., & Goldstone, R. L. (2018a). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, 5(2), 47-59. <https://doi.org/10.18608/jla.2018.52.4>
- Motz, B., de Leeuw, J., Carvalho, P., Fyfe, E., & Goldstone, R., (2018b). ManyClasses: A model for abstracting generalizable research principles from different learning contexts. *replicate.education: A Workshop on Large Scale Education Replication*. Buffalo, New York.
- Motz, B., Busey, T., Rickert, M., Landy, D. (2018c). Finding topics in enrollment data. *Proceedings of the 11th International Conference on Educational Data Mining*. Buffalo, New York.
- Motz, B., Quick, J., Schroeder, N., Zook, J., Gunkel, M. (2019). The validity and utility of activity logs as a measure of student engagement. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge*. ACM.
- Rossi, P. (1987). The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy*, 4, 3-20.
- Saxe, J.G. (1873). *The poems of John Godfrey Saxe*. Boston: James R Osgood & Company.
- Silberzahn, R. & Uhlmann, E.L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, 526(7572), 189-191. <https://doi.org/10.1038/526189a>
- Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E.C., ... Nosek, B.A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results (preprint). *PsyArXiv*. <https://doi.org/10.1177/2515245917747646>
- Serlin, R.C., & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, 40(1), 73-83. <https://doi.org/10.1037/0003-066X.40.1.73>
- Steyvers, M. & Benjamin, A.S. (2018). The joint contribution of participation and performance to learning functions: Exploring the effects of age in large-scale data sets. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-018-1128-2>
- Tukey, J. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6(1), 100–116. <https://www.jstor.org/stable/2245714>
- Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2(2), 5-13.

Assessing Institutional Needs for Learning Analytics Adoption in Latin American Higher Education

Isabel Hilliger

Pontificia Universidad Católica de Chile
ihillige@ing.puc.cl

Mar Pérez-Sanagustín

Pontificia Universidad Católica de Chile
mar.perez@uc.cl

Margarita Ortiz

Escuela Superior Politécnica del Litoral (ESPOL)
margarita.ortiz@cti.espol.edu.ec

Paola Pesántez

Universidad de Cuenca
paola.pesantezc@ucuenca.edu.ec

Eliana Scheihing

Universidad Austral de Chile
escheihi@inf.uach.cl

Yi-Shan Tsai

The University of Edinburgh
Yi-Shan.Tsai@ed.ac.uk

Pedro J. Muñoz-Merino

Universidad Carlos III de Madrid
pedmume@it.uc3m.es

Tom Broos

Katholieke Universiteit Leuven
tom.broos@kuleuven.be

ABSTRACT: In recent years, Learning Analytics (LA) has captured the attention of higher education managers who saw in this research field a means to optimize the process of teaching and learning on a large scale. So far, most studies in LA have concentrated on the development of tools to address educational challenges in the contexts of Europe, Australia, and U.S. However, tools and adoption frameworks developed in these contexts are not necessarily applicable for higher education institutions in the rest of the world. Given that there is no one-size-fits-all approach, this study aims to assess institutional needs for LA in the Latin American context by collecting and analyzing qualitative information obtained from managers, teaching staff and students at four universities (U1, U2, U3, and U4). Although most participants agreed that LA is a promising means to monitor students' academic progress at a curriculum level, findings show specific needs and considerations that differentiate each university (U1: academic support for subgroups, U2: dropout indicators,

U3: improving existing counseling tools, and U4: satisfaction indicators). Given these differences, iterative process models are required to guide LA adoption in the Latin American context.

Keywords: Learning Analytics, Learning Analytics Adoption, Stakeholder Involvement, Higher Education, Latin America

1 INTRODUCTION

Learning Analytics (LA) aims to develop different methodologies, techniques and technological tools to optimize learning processes and its environments (Siemens & Gasevic, 2012). By leveraging existing large amounts of data, LA has proved to have great potential for improving teaching, learning, and organizational efficiency and decision-making (Jones, 2015; Zilvinskis, Willis, & Borden, 2017). This explains the rapidly growing interest in LA solutions as a means to address student retention and meet other accountability demands in higher education (Macfadyen, Dawson, Pardo, & Gasevic, 2014).

So far, most studies in LA have concentrated on the development of tools and methods to support small-scale activities for a limited period of time (Ferguson et al., 2016). There is limited evidence validated by research to demonstrate the impact of these tools on informing managerial decision-making processes at an institutional level (Macfadyen et al., 2014), or teaching and learning processes at a classroom level (Ferguson et al., 2016). Moreover, the availability and deployment of LA tools does not guarantee learning benefits if its adoption is not closely integrated with learning design and decision-making across institutional and classroom levels (Gasevic, 2018). Even in regions where researchers have made more progress in the development and validation of LA solutions (i.e. North America, Europe and Australia), only a few universities have started to strategically plan for LA adoption (Colvin, Dawson, & Fisher, 2015). To implement LA at an institutional scale, higher education managers, teaching staff and students will need more guidance (Dawson et al., 2018), so more efforts have to be invested in understanding how these stakeholders could adopt LA tools and methods in their everyday practice (Ferguson et al., 2016).

Along these lines, researchers have highlighted the importance of understanding how higher education stakeholders use LA tools and methods to make successful interventions in real-life settings (Rienties et al., 2016). Researchers have begun to propose theoretical and conceptual frameworks as mechanisms to lead managers, teaching staff and students through LA adoption. Most of these frameworks are based on the idea that these stakeholders and policy makers become more involved in the design and implementation of LA solutions, this will inform stronger research that will eventually lead to a better understanding and implementation of LA (Rienties et al., 2016; Tsai, Moreno-Marcos, Tammets, & Gasevic, 2018). For example, the SHEILA project introduces a policy-development framework for LA adoption based on the perspectives of various stakeholders, including institutional managers, teaching staff, students and LA experts (Tsai et al., 2018). However, there is a paucity of research that evaluates the use of existing frameworks in real-life environments (Dawson et al., 2018). Indeed, as Ifenthaler (2017) contends:

“we need empirical research on the validity of LA frameworks and on expected benefits for learning and instruction to confirm the high hopes this promising emerging technology raises” (Ifenthaler, 2017, p. 37).

To our knowledge, there has been no formal framework based on the needs of LA in Latin America. Our study is set out to bridge this gap by addressing the following research question: **What are the needs and considerations for adopting Learning Analytics tools in Latin America?** To answer this question, we assessed the institutional needs of four Latin American universities affiliated to a large project that aims to build the local capacity to design, implement and use Learning Analytics tools in Latin American Higher Education (LALA project-<https://www.lalaproject.org/>). To date, existing LA initiatives in Latin America have been limited and isolated (Cobo & Aguerrebere, 2017), so we have chosen to carry out the study in the four Latin America universities affiliated to the LALA project to contribute to a better understanding of LA adoption in institutions that share a similar culture and political context.

As the LALA project moves forwards, its participants aim to develop a framework to facilitate LA adoption in Latin America. This framework addresses four fundamental dimensions for LA adoption: (1) the institutional dimension, which considers the institutional needs identified by contrasting the current and desired state in relation to the adoption of LA institution-wide; (2) the methodological dimension, which considers the technical needs for the design and implementation of LA tools; (3) the ethical dimension, which considers a series of guidelines to support the ethical use of the data; and (4) the community dimension, which proposes a series of guidelines to ask for support to conduct research and development in this field. In this context, this paper addresses the institutional dimension of this framework.

2 LITERATURE REVIEW

Around the globe, many higher education managers have high hopes that LA tools and methods can help them leverage large academic databases to create supportive and insightful models of teaching and learning processes - even in real time (Rienties et al., 2016). The collection and analysis of such data is a promising approach to provide personalized and scalable support for learners, besides providing information to improve teaching practices, organizational efficiency, and decision-making (Gasevic, 2018; Jones, 2015). However, the availability of analytical tools and methods does not guarantee these improvements; managers, teaching staff and students have to adopt them to make successful interventions in their own practice (Rienties et al., 2016). Considering that the limited number of experienced LA research groups already constitutes an important barrier for LA adoption in Latin America (Cobo & Aguerrebere, 2017), this section briefly reviews the literature regarding the challenges for LA adoption, as well as the models and frameworks proposed to overcome them.

2.1 Challenges of Learning Analytics Adoption

In the past few years, a growing number of publications have documented challenges that affect LA design and implementation. One challenge is the lack of case studies that empirically validate technology development on a larger scale for longer period of time (Ferguson et al., 2016; Tsai et al., 2018). Another challenge is the need for policies to address issues of privacy and ethics related to informed consent, data transparency, data ownership, and data access (Gasevic, 2018; Steiner,

Kickmeier-rust, & Albert, 2015). Other prominent challenges are related to the lack of stakeholder involvement (Macfadyen et al., 2014), LA expertise (Ifenthaler, 2017), leadership support (Tsai & Gasevic, 2017), and training opportunities (Tsai & Gasevic, 2017).

To address these challenges, higher education has made great improvements in the technical development of LA tools (Zhong, 2016), as well as in the development of policies to ensure ethical treatment of data (Steiner et al., 2015). However, a major challenge still confronts higher education institutions – stakeholder involvement (Tsai et al., 2018). On the one hand, stakeholders at different levels could have varied data-related experiences and knowledge, leading to discrepancies in the perception of LA benefits and outcomes (Tsai & Gasevic, 2017). On the other hand, some stakeholders might expect that LA per se can enable change, without realizing that their interpretation of educational data is what drives further interventions to improve learning (Zilvinskis et al., 2017).

Therefore, it is important to develop comprehensive institutional policies to encourage positive attitudes towards LA among different stakeholders (Macfadyen et al., 2014). In particular, key leadership is crucial to a clear strategy for successful LA adoption on an institutional scale (Tsai et al., 2018). Along these lines, researchers have documented success stories about stakeholder involvement in North America and Europe (Gasevic, 2018). For example, institutional leaders from Denmark, the Netherlands and Norway have begun to develop national approaches to support and enable learning analytics at a large scale (Ferguson et al., 2016). Conversely, research about LA is still considered emergent in Latin America (Cobo & Aguerrebere, 2017). The study, as part of LALA project, intends to bridge the gap by creating a community to exchange ideas, methodologies and tools to expand LA adoption in Latin American higher education (Lemos dos Santos, Cechinel, Carvalho Nunes, & Ochoa, 2017).

Given the difference in maturity of LA adoption in Latin America compared to Europe, it is necessary to develop guiding frameworks to direct the design and implement LA tools based on stakeholders' needs. To this end, our study used two data gathering techniques to explore stakeholder perceptions of the needs for LA adoption in four Latin American universities in Chile and Ecuador. The main objective is to explore the viewpoints of various stakeholders in order to assess local needs, given that there is no one-size-fits-all policy for learning analytics (Zilvinskis et al., 2017).

2.2 Existing Frameworks for Learning Analytics Adoption

To scale up and sustain LA adoption in higher education, researchers have recently developed an increasing number of frameworks as an attempt to guide the design and implementation of LA solutions at an institutional level. According to Dawson et al. (2018), these frameworks could be classified into input, output and process models. Most of them consist of input models, which define a set of dimensions or properties to assess institutional readiness for LA adoption (Dawson et al., 2018). For example, the Learning Analytics and Readiness Index (LARI) proposed by Arnold and colleagues is used to identify key factors for LA adoption readiness (Arnold, Pistilli, St, & Hall, 2014).

Another type of framework proposed to facilitate LA adoption is the one described as output or outcome-based (Dawson et al., 2018; Jones, 2015). These frameworks represent LA deployment as a linear process that unfolds over time according to different levels of organizational readiness and

maturity (Colvin, Dawson, Wade, & Gasevic, 2017). Along these lines, Dawson and others alluded to the LA sophistication model proposed by Siemens, Dawson, and Lynch (2013), which represents a five-stage process that goes from emergent data to integrated adaptive and personal learning.

Although the input and output LA frameworks provide valuable information to guide LA adoption, most of them describe conceptual dimensions or stages of LA deployment, without addressing the dynamic and unpredictable pressures that currently affect higher education (Dawson et al., 2018; Jones, 2015). In response to the dynamic contexts of higher education, process models have emerged to map alternative approaches for LA adoption regarding the evolving needs and concerns raised by higher education stakeholders (Dawson et al., 2018). Along these lines, Tsai and colleagues proposed the SHEILA policy-development framework (Tsai et al., 2018), which is based on the RAPID Outcome Mapping Approach (ROMA) (Young, J. Mendizabal, 2009). This approach consists of an iterative process to develop evidence-based policy through active engagement with relevant stakeholders.

In this study, we built upon the experience of the SHEILA framework to assess the needs of different higher education stakeholders, using a participatory action research method (see Section 3) (Creswell, 2012). This needs assessment contributed to a framework that we have developed to guide the design, implementation and use of learning analytics tools in higher education institutions in Latin America (LALA framework-<https://www.lalaproject.org/deliverables/>). Thus, this paper presents our effort to assess institutional needs for LA adoption to adapt existing process models to better suit the Latin American context.

3 METHODOLOGY

This paper addresses the following research question: **What are the needs and considerations for adopting Learning Analytics tools in Latin America?** To answer this question, we assessed the institutional needs for LA adoption in four Latin American universities that are part of the LALA project. Although the findings of the study are limited to the four chosen cases, it expands on the limited research about LA in the region by providing insights about implications for LA adoption in these and similar institutions. In the following sections, we describe the participants and samples, the data gathering techniques, and the data analysis plan used to identify the needs for LA.

3.1 Participants and Sample

Four Latin American universities participated in this study: two traditional private institutions in Chile (U1 and U2), and two public institutions in Ecuador (U3 and U4). Table 1 shows the samples used to assess the needs for LA adoption in these four universities, and Appendix 1 describes each university briefly (Appendix 1: <http://bit.ly/2OpB2va>).

Table 1: Sample of Participants per Data Gathering Technique

	U1	U2	U3	U4
LALA Canvas	5 experts	3 experts	3 experts	5 experts
Interviews with managers	7 managers	11 managers	8 managers	11 managers
FG with students	2 FG (13 students)	1 FG (5 students)	2 FG (3 students)	3 FG (24 students)
FG with teaching staff	1 FG (5 teachers)	2 FG (15 teachers)	2 FG (8 teachers)	3 FG (23 teachers)

FG: Focus groups

Interviews and focus groups were guided by the interview protocol.

3.2 Data Gathering Techniques

Two different data gathering techniques were used in this study: the LALA Canvas and a semi-structured interview protocol. The first one was used to define a general overview of the current state of LA adoption at an institutional level, while the second one was used to obtain further insights about the desired state of LA adoption and the needs to adopt LA tools at a large scale.

3.2.1 LALA Canvas

This technique consists of a template that aims to guide a group discussion about the current state of a higher education institution in terms of LA adoption (<http://bit.ly/LALACanvas>). The template was built upon the experience of the SHEILA framework (Tsai et al., 2018), with a further adaptation of the ROMA dimensions (Young, J. Mendízabal, 2009). Along these lines, the dimensions considered in the LALA Canvas were: 1) desired behaviors, 2) strategy for change, 3) internal capacities, 4) political context, 5) key stakeholders, 6) assessment and evaluation plan.

To define the current state of LA adoption, the LALA Canvas was completed in four groups of 3 to 5 experts with varied experiences in LA (e.g. education vs. computer science background, PhD students vs experienced researchers, etc.). Each group analyzed the current state of the university they were affiliated with (see Table 1). The group discussions were held in March 2017, with a moderator guiding the participants to assess their institutional context in relation to the six dimensions in the canvas. This activity lasted an hour approximately.

3.2.2 Interview Protocol

This technique consists of a semi-structured guide to interview managers, teaching staff and students, in order to explore the institutional needs for LA adoption (<http://bit.ly/2OjnwJo>). It was built upon instruments used by the SHEILA project with the objective of collecting information about the desired state of LA adoption at an institutional level. It includes questions about the expected uses of educational data and existing ethical and privacy policies.

To assess the desired state of LA adoption, the interview protocol was used to interview managers, teaching staff and students at U1, U2, U3, and U4 between January and August 2018 (see Table 1). A snowball sampling method was followed to identify suitable managers to be interviewed, while a stratified sampling method was followed to identify teaching staff and students from different

academic units (Creswell, 2012). Managers were interviewed individually in 30-minute sessions (approximately), whereas teaching staff and students were interviewed in separate focus groups, each one lasting an hour.

3.3 Data Analysis Plan

The data analysis plan consisted of three steps:

3.3.1 Defining the Current State of LA Adoption

In this step, the same experts who worked on the LALA Canvas of each university summarized elements under each dimension, aiming to reach consensus on their observations of the six dimensions in their own institutional context. All of these elements were documented in a Microsoft Word version of the LALA Canvas template.

3.3.2 Defining the Desired State of LA Adoption

In this step, one expert from each university summarized the results of interviews according to the protocol questions in an Excel spreadsheet. Then, they presented the findings in a report focusing on the desired state of LA adoption in their institution, addressing the needs for LA tools, the considerations for the design and implementation of LA methods, the ethical and privacy elements required, and the sustainability and scalability of LA initiatives in the region.

3.3.3 Assessing Needs and Considerations for LA Adoption

In this step, experts from each university identified the gaps between the current and the desired state in terms of LA adoption by contrasting the elements listed in the LALA Canvas with the results summarized from the interview protocol. Then, they used this contrast to determine how LA could be used at their universities (i.e. needs), besides anticipating issues for future design of LA tools and methods.

4 RESULTS AND DISCUSSION

This section summarizes the analysis results, focusing on the needs for LA adoption and considerations of ethical aspects in the four Latin American

4.1 Needs for Learning Analytics Adoption

Table 2 presents the needs for LA adoption that were identified in each university. All the universities in this study considered LA tools and methods as a promising means to obtain clear information about students' academic progress at a curriculum level. However, there were specific needs that differentiate each university. For example, U1 makes a specific emphasis on providing academic support for student subgroups, U2 on monitoring high failure rates and dropout risks, U3 on improving existing LA tools for counseling, and U4 on monitoring student satisfaction. Considering that needs vary according to the institutional context (Gasevic, 2018), adoption frameworks based on process models might be more suitable to guide LA adoption in Latin America (Dawson et al., 2018). This finding is consistent with our strategy of building upon the experience of the SHEILA framework to assess institutional needs in Latin American universities (Tsai et al., 2018).

Table 2: Institutional Needs for Learning Analytics Adoption

Needs for Learning Analytics Adoption	
U1	<ul style="list-style-type: none"> • Academic support for student subgroups • Timely and personalized feedback to improve the teaching and learning process. • Clear information about students' academic workload. • Clear information about students' academic progress at a curriculum level.
U2	<ul style="list-style-type: none"> • Indicators for high failure rates and dropout risks. • Timely and personalized monitoring of students' and teaching staff performance. • Clear information about students' academic workload. • Clear information about academic and psycho-socio-emotional profiles of students. • Clear information about students' academic progress at a curriculum level.
U3	<ul style="list-style-type: none"> • Improvements of existing LA tools for counseling. • Exploitation of educational data collected from both teaching staff and students. • Integrated systems to obtain information about the academic and psycho-socio-emotional profiles of the students. • Clear information about students' academic progress at a curriculum level.
U4	<ul style="list-style-type: none"> • Clear information about students' satisfaction at a course and program level. • Timely and personalized monitoring of students' and teaching staff performance. • Indicators for high failure rates and dropout risks. • Clear information about academic and psycho-socio-emotional profiles of students. • Clear information about students' academic workload. • Clear information about students' academic progress at curriculum level.

4.2 Ethical Considerations for Learning Analytics Adoption

Table 3 shows the ethical considerations for future designs of LA tools and methods. Most institutions alluded to the need for ethics-related policies to address issues concerning informed consent, data access, and data transparency, which aligns with suggestions in the LA literature (Gasevic, 2018; Steiner et al., 2015). Besides, most institutions emphasized the need for procedures to ensure data transparency, which is an important issue when adopting LA at an institutional level. However, there are certain considerations that were raised by individual cases only, such as the emphasis on informed consent at U1 and the need of training in privacy issues at U2 and U4. Thus, further work is needed to understand what considerations are generalizable for these and other similar institutions to develop privacy and data protection framework as the ones developed for European institutions (Steiner et al., 2015).

Table 3: Ethical Considerations for Learning Analytics Adoption

Ethical considerations	
U1	<ul style="list-style-type: none"> • Need for rigorous processes for informed consent. • Need for procedures for data transparency. • Policy-making to sustain ethical-related practices.
U2	<ul style="list-style-type: none"> • Importance of information security compliance. • Need for staff training in privacy issues.
U3	<ul style="list-style-type: none"> • Policies concerning data access, data transparency and informed consent.

Ethical considerations

- U4
- Need for rigorous processes for informed consent.
 - Need for procedures for data transparency.
 - Policy-making to sustain ethical-related practices.
 - Importance of information security compliance.
 - Need for staff training in privacy issues.
-

5 CONCLUSIONS AND IMPLICATIONS

This study contributes to the growing research aimed at understanding LA adoption by assessing institutional needs at four universities in Latin America. Although findings show that all stakeholders of these universities considered LA as a promising means to obtain clear information about students' progress at a curriculum level, there were specific institutional needs and ethical considerations that differentiate each university. As it has been sustained by Gasevic (2018), the "one-size-fits-all" approach does not work for data models, and it might not work for models for LA adoption either.

As needs and considerations vary according to the institutional context, there are practical implications for the development of adoption frameworks for Latin America. First, process models might be more suitable to map alternative approaches for LA adoption regarding the evolving needs and concerns raised by stakeholders, including institutional managers, teaching staff, students and LA experts. Second, these process models must be iterative, starting by assessing institutional needs and identifying ethical and privacy considerations for use of academic data. And third, considerations and other lessons learned must be discussed among LA experts in the region in order to identify generalizable knowledge to disseminate for both research and capacity building purposes.

Future work will cross-analyze the findings in more detail to extend the current research on LA adoption in Latin American universities. Findings will inform the development of an adoption framework that will be internally and externally validated as LA tools are designed and implemented in different institutions of the region.

6 ACKNOWLEDGEMENTS

Work funded by the LALA project (grant no. 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). This project has been funded with support from the European Commission. This publication reflects only the views of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- Arnold, K. E., Pistilli, M. D., St, S. G., & Hall, Y. (2014). An Exercise in Institutional Reflection: The Learning Analytics Readiness Instrument (LARI). In *LAK'14: International Conference on Learning Analytics and Knowledge*. Indianapolis, IN, USA.
- Cobo, C., & Aguerrebere, C. (2017). Building capacity for learning analytics in Latin America. In C. Ping Lim & V. L. Tinio (Eds.), *Learning Analytics for the Global South* (pp. 63–67). Quezon City, Philippines: Foundation for Information Technology Education and Development, Inc.

- Colvin, C., Dawson, S., & Fisher, J. (2015). *Student retention and learning analytics: A snapshot of Australian practices and a framework for advancement*. Sydney, Australia: Australian Government Office for Learning and Teaching.
- Colvin, C., Dawson, S., Wade, A., & Gasevic, D. (2017). Addressing the Challenges of Institutional Adoption. *Handbook of Learning Analytics*, 281–289. <https://doi.org/10.18608/hla17.024>
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. (P. E. Inc., Ed.), *Educational Research* (Fourth Edi, Vol. 4). Boston, Massachussetts. <https://doi.org/10.1017/CBO9781107415324.004>
- Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., & Gasevic, D. (2018). Rethinking learning analytics adoption through complexity leadership theory. In *LAK'18: International Conference on Learning Analytics and Knowledge*. Sydney, NSW, Australia. <https://doi.org/10.1145/3170358.3170375>
- Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., ... Vuorikari, R. (2016). Research Evidence on the Use of Learning Analytics: Implications for Education Policy. In V. R. & J. Castaño Munoz (Eds.), *Joint Research Centre Science for Policy Report* (pp. 1–150). Luxembourg: Publications Office of the European Union. <https://doi.org/10.2791/955210>
- Gasevic, D. (2018). Directions for adoption of learning analytics in the global south. In C. Ping Lim & V. L. Tinio (Eds.), *Learning Analytics for the Global South* (pp. 2–22). Quezon City, Philippines: Foundation for Information Technology Education and Development, Inc.
- Ifenthaler, D. (2017). Are Higher Education Institutions Prepared for Learning Analytics? *TechTrends*, 61(4), 366–371. <https://doi.org/10.1007/s11528-016-0154-0>
- Jones, H. (2015). The “I”s have it: Development of a framework for implementing Learning Analytics. In *Ascilite* (pp. 680–683). Perth, Western Australia.
- Lemos dos Santos, H., Cechinel, C., Carvalho Nunes, J. B., & Ochoa, X. (2017). An Initial Review of Learning Analytics in Latin America. In *Twelfth Latin American Conference on Learning Technologies (LACLO)*. La Plata, Argentina.
- Macfadyen, L. P., Dawson, S., Pardo, A., & Gasevic, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, 9(2), 17–28. <https://doi.org/10.1017/CBO9781107415324.004>
- Rienties, B., Boroowa, A., Cross, S., Kubiak, C., Mayles, K., & Murphy, S. (2016). Analytics4Action Evaluation Framework: A Review of Evidence-Based Learning Analytics Interventions at the Open University UK. *Journal of Interactive Media in Education*, 1(2), 1–11. <https://doi.org/10.5334/jime.az>
- Siemens, G., Dawson, S., & Lynch, G. (2013). *Improving the Quality and Productivity of the Higher Education Sector Policy and Strategy for Systems-Level Deployment of Learning Analytics*. Sydney, Australia: Australian Government Office for Learning and Teaching.
- Siemens, G., & Gasevic, D. (2012). Guest Editorial - Learning and Knowledge Analytics. *Education Technology & Society*, 15(3), 1–2.
- Steiner, C. M., Kickmeier-rust, M. D., & Albert, D. (2015). Let ' s Talk Ethics: Privacy and Data Protection Framework for a Learning Analytics Toolbox. In *LAK'15: International Conference on Learning Analytics and Knowledge*. Poughkeepsie, New York.
- Tsai, Y., & Gasevic, D. (2017). Learning analytics in higher education - Challenges and policies: A review of eight learning analytics policies. In *LAK'17: International Conference on Learning Analytics and Knowledge*. Vancouver, BC, Canada. <https://doi.org/10.1145/3027385.3027400>
- Tsai, Y., Moreno-Marcos, P. M., Tammets, K., & Gasevic, D. (2018). SHEILA policy framework: informing institutional strategies and policy processes of learning analytics. In *LAK 18' International Conference on Learning Analytics & Knowledge*. Sydney, Australia. <https://doi.org/10.1145/123>
- Young, J. Mendízabal, E. (2009). *Helping researchers become policy entrepreneurs: How to develop engagement strategies for evidence-based policy-making. Briefing Paper* (Vol. 53). Retrieved from <https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/1730.pdf>

- Zhong, L. (2016). A Systematic Overview of Learning Analytics in Higher Education. *Journal of Educational Technology Development & Exchange*, 8(2), 39–54. <https://doi.org/10.18785/jetde.0802.03>
- Zilvinskis, J., Willis, J. I., & Borden, V. M. H. (2017). An Overview of Learning Analytics. *New Directions for Higher Education*, 179(Fall 2017), 9–17. <https://doi.org/10.1002/he>

Development, Sustainment, and Scaling of a Learning Analytics, Prediction Modeling and Digital Student Success Initiative

Matthew L. Bernacki

University of North Carolina – Chapel Hill, School of Education
mlb@unc.edu

ABSTRACT: This project demonstrates efforts to sustain and scale a learning analytics solution that employs students' own course-specific event data in a learning management system (LMS) to predict and inform interventions to support students' academic success. An initial overview details LMS events captured, feature engineering to reflect temporal position and aggregation of like events into traces of theoretically aligned learning processes, and model building to select an algorithm that predicts course performance with the feature set. Results of intervention studies are summarized, and efforts towards sustaining partnerships and systematizing project features are discussed. Thereafter, two cases of funded initiatives examine the generalizability of the solution to additional course, university, and learning technology contexts. Case 1 is a Provost's initiative funded by internal, sustainable university resources. This case study examines how a learning analytics solution can abide naturally occurring changes – to instructional partners, digital content and assessment practices, the LMS, and key personnel – and growth to new courses with differing features. Case 1 also addresses financial sustainability and returns required to warrant ongoing university investment. Case 2 is a research initiative and examines generalizability of the learning analytics solution across multiple universities, LMSs, and data platforms.

Keywords: Learning management systems, Prediction modeling, instructional design, intervention, generalizability, sustainability, feature engineering

1 BACKGROUND

In response to issues with student performance, retention, progression, and completion many universities and educational software providers are developing “early warning systems” to identify and support students likely to obtain poor outcomes [1, 7, 8, 9, 10, 12]. These early warning systems involve the development of a data model that collects information about the characteristics of the learner – as supplied to a student's profile – by a registrar office and other repositories that store information about the individual (e.g., admissions, financial aid data). Additional data from events the learner induces can be collected when students use technologies to support learning and engagement on campus. These technologies include learning management systems, companion sites for course media like assigned textbooks, and other campus systems that provide and track use of resources (wifi access, library services, student life offices, health and dining services, etc.) [2]. Learning scientists, with the help of computer scientists and IT operations professionals in the university's employ, collect these data into a model, engineer features thought to reflect important events related to learning, and test algorithms that use these features to predict key outcomes including enrollment and performance in courses, retention across semesters, and completion of programs. Learning analytics solutions that produce timely predictions of the likelihood a student will obtain – or is at risk of not obtaining – a desired outcome provide an opportunity to intervene and support students, thus increasing the

odds of obtaining the desired outcome after receiving support. These solutions can be complicated to build, and further require effort to ensure that the key components of projects are maintained in ways that ensure (1) the reliability of data collection, (2) that event data (e.g. access of digital course content) continue to validly represent student actions and intentions, (3) that outcome variables (e.g. course exams) are held constant, and that algorithms continue to accurately predict student outcomes based on the original model. These maintenance issues are critical to the sustainability of a single learning analytics solution and its ongoing viability in one learning context. To further generalize this solution and expand its use, project components need to be systematized for broader application. Personnel must be able to rely on project documentation that guides work phases, and training programs must be developed so that new team members can be added to accommodate growth in the project and to sustain losses when team members depart.

This paper describes efforts to develop and sustain a learning analytics solution from a research project focused on three (science, math, and engineering) courses, a subsequent initiative to scale the methodology to accommodate 10 courses at the same university (Case 1), and to generalize the solution to three universities that differ in their student population, technology platforms, and resources to support student success (Case 2).

2 THE ORIGINAL LEARNING ANALYTICS SOLUTION

The original project, *Learning Theory and Analytics as Guides to Improve STEM Education* was supported by an external research award wherein the research team employed students' own course-specific event data in a learning management system (LMS) to predict course performance and to intervene to improve learning and achievement.

2.1 Learning management system events as traces of learning processes

Undergraduate students utilize a learning management system (LMS) for multiple functions, and the kinds of learning processes that can be observed are dictated by course objectives and the kinds of digital learning resources instructors provide for student learning [2]. For example, in a biology course requiring mastery of declarative knowledge about anatomical features and conceptual and procedural knowledge of physiological system functioning, students can use digital resources provided by the instructor to engage in cognitive and metacognitive learning processes as they adopt learning principles and study course topics (Table 1). Clicks on these resources produce requests to servers and events in the server logs, which can be mined and reorganized to produce records of learning events.

In order to appropriately engineer features that describe students' use of multiple pieces of course content that reflect these learning processes, learning scientists must classify content items by the kinds of learning principles the resource type affords. Based on design features of the LMS resources, patterns of student activity may further implicate how to represent data in prediction models [2, 7]. For instance, it is more appropriate to model use of a downloadable files like exam blueprints as a dichotomous event that should impact learning if it occurs once (indicating that a student has obtained the file) compared to zero times (indicating the student has not). In contrast, resources designed for repeated use online, such as practice quizzes, are best captured as count data.

Table 1. Empirically-supported learning principles and digital content that support their enactment

Learning Processes	Digital Resources in the LMS course site to support enactment
Cognitive	
[Spaced] Retrieval Practice	Chapter quizzes with item pools designed for repeated use (ungraded; provides correctness feedback, textbook reference)
Self-Explanation	Self-assessment opportunities providing prompts to self-explain & evaluate answers
Worked Examples	Annotated, diagrammatic presentation of biological systems in graphic or video form
Metacognitive	
Planning	Exam Blueprints with weighting topic & depth of knowledge
Monitoring Learning	Detailed Feedback with correctness and pointers to content areas for restudy after self-assessment completion (above)
Monitoring Progress towards Mastery	Self-assessment of Learning Goal Mastery (Editable Worksheet of Learning Goals per Chapter, Unit & level of mastery to date)
Monitoring Performance	Digital Gradebook

2.2 Data modeling and development of algorithms to predict student success

A key phase of the original project was to test different algorithms that could balance prediction accuracy obtained by the model, and coherence of model implementation where university data systems could record learning events in ways that could be programmed back into a model that provided the real time predictions of a student's likelihood of success that informed timely interventions. We examined implications of different representations of LMS resource use on the accuracy of prediction models, examine whether the most accuracy model predicts performance in subsequent samples, and whether the model can provide a basis for alerting students about their potential for poor achievement. [7]

For the biology course described above, prediction modeling involved data extracted from server logs of users' learning events in the LMS from the first four weeks of the course (i.e., prior to any exam). Early warnings could then be generated and sent in time for learners to adjust tactics or seek help a full week prior to their first unit exam (i.e., in Week 5). Events were aggregated and enriched using Splunk, a platform for search and modeling of machine data, and tables of metadata about content items. Classification of items into resource types was handled by human research programmers. Models were built and evaluated in RapidMiner. We compared models that involved different levels of aggregation of learning events (i.e., count, dichotomous representations per content items and classes of resources) and tested different temporal aggregations (i.e., the day or week of the semester). These combinations of feature classes were submitted to algorithms including forward selection logistic regressions, decision trees (J-48, J-Rip), Naïve Bayes, and K-Star. Models were cross validated using k-fold methods (usually with 10 folds) [7]. With data from an initial semester of learners (roughly 325) and with a focus on the recall metric within confusion matrices, we

settled on the use of a forward selection logistic regression that could identify 4 of 5 ($\geq 80\%$) students likely to fail to obtain the B average needed to move forward to the next course in the biology sequence for health science majors (i.e., the goal of most students enrolled).

2.3 Sustainability and generalizability of the prediction model

2.3.1 Sustainability of model accuracy over multiple semesters

We monitored drift in the course over the subsequent semesters of biology students, and refit the model once when adjustments to course materials warranted. Examination of model accuracy for students in our control group (i.e., those predicted as eligible for intervention but who were only tracked and did not receive an intervention) indicated that were able to sustain a lower bound of 75% accurate recall over the three years when the prediction model was applied in the course, as evidenced by confusion matrix reports.

2.3.2 Generalizability of the modeling approach to multiple course types

The modeling approach was later extended to a Calculus and an Engineering course. A similar level of prediction accuracy was achieved with the same feature engineering approach, despite changes in the number of weeks available to collect learning events and sparser digital content in some course sites. These models also demonstrated that, when instructional design features including digital course content and assessment practices were held constant across multiple course sections and taught by different instructors of record (i.e., a master LMS course site is developed, and instructors adopt identical course pacing, exam timings, items, and scoring), prediction models can maintain their accuracy [7, 8].

2.4 Intervention to support students predicted to struggle

Students whose LMS data informed the prediction model were classified a week before their first exams via logistic regression as likely to obtain a B or Better in the course or likely to obtain a C or worse and need to re-enroll again next semester. Because such models are diagnostic but not causal, intervention efforts were not tailored to model features. Rather, students were encouraged to consider adopting (or developing) learning strategies known to promote achievement on tasks aligned to course learning objectives. Students with prediction values indicating a grade of C or Worse was likely (i.e., > 0) were randomized into a Control group that received a message from the instructor that reminded them about the upcoming exam, or to an Intervention group whose message also recommended an advice page (Figure 1), and a training program called the *Science of Learning to Learn* (Figure 2):

Subject: A Check-in on your learning

Body: Hi [Name]!

Our first course exam is coming up in a week. {I want to check-in to make sure each student is on top of our content, learning in appropriate ways, and able to perform well. So, I'd like to direct you to **two** resources that can help you with learning the material in our course:

- A one-page summary of advice from students who have completed the course in the past.
- A set of learning modules called "The Science of Learning to Learn." These modules describe learning strategies you can use with our course materials.

Both resources can be found on the [LMS] course site under the **STEM Learning Resources** link in the left panel (and provided in this announcement, below). I hope you find that these resources help you to learn and perform well!!

Dr. [Instructor]

Note. Intervention group message received the additional text in curly brackets, (i.e., {text}).

ADVICE FROM PAST STUDENTS ON *Tackling Anatomy & Physiology*

UNLV students were asked to reflect on their experience learning in their anatomy and physiology lecture courses. Below they described the things that helped them learn and score well – and some that didn't.



Britney

Plan ahead! Seek out materials early on, make a study plan, and stick to it.

"There's definitely a learning curve in biology courses, so it's good to get advice from past students. I would recommend that students seek out any helpful materials in the beginning of the course and make a plan to use them. For courses like anatomy that require a lot of memorization, repeated practice is key. Certainly read the textbook, but then look to the learning objectives and the materials the instructor provides and plan from there – there are usually online tools like practice quizzes to help you study. Of course, studying last minute is the biggest mistake. Spacing out studying periods and making sure to never get behind is very important. Cramming is the worst thing to do and a big reason why finals can be so challenging."

I spent lots of time rehearsing my knowledge. Quizzing myself really help me learn factual information.

"I've been most successful when I spend a lot of time rehearsing my knowledge. I usually make note cards as I read through the chapters and test myself using the note cards every week. This method is it requires time and dedication, but it works for helping me remember key definitions and concepts. Using note cards and online tools for quizzing myself has been great to help learn factual information. When it comes to learning systems and processes, I also do better when I read the chapter and then follow up by watching videos on the topic to help visualize and solidify the concept. Overall, it's the constant rehearsal that's been key for me."



Alexis

Figure 1. Sample of the resource providing advice from past successful students

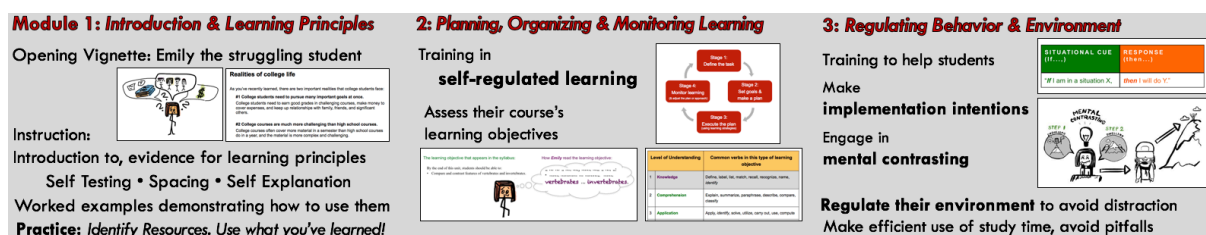


Figure 2. Visualization of Science of Learning to Learn training module design and content

Those who received the Intervention message and accessed resources significantly outperformed Control students on subsequent exams ($d_s = .2$ to $.4$ in Biology, $.6$ to $.9$ in math) [4]. These findings were sufficiently encouraging that efforts were made to scale this solution to additional courses in one institution and to test its generalizability to others.

3 GENERALIZING A LEARNING ANALYTICS RESEARCH SOLUTION TO A UNIVERSITY LEARNING ANALYTICS INITIATIVE (CASE 1)

3.1 The University Context

At the large public university where the learning analytics research solution was developed, the student body is largely comprised of first generation college students who graduated from low resource high schools (i.e., receiving supplemental federal funding via U.S. Title 1 grants) and who hail from ethnic groups that are historically under-represented in higher education and STEM fields. Issues with academic achievement and retention, progression, and completion of degree programs are pronounced challenges, and university leadership sought to leverage local learning analytics research to address a university need. Campus stakeholders were gathered by the Provost and the research team, and a partnership spanning course instructors, Information Technology and Online Education offices was assembled to scale the research solution to serve students in ten courses.

3.2 Scaling Components to Convert Research Effort to a University-wide Initiative

3.2.1 Staffing the initiative

A staff of three individuals composed the Learning Analytics Initiative team (LAI; with asterisk, in Table 2). This group worked with additional campus stakeholders to replicate

and extend a scope of work that involved instructional design efforts in each course, a common data infrastructure serving all courses, and a feature engineering, data mining, and prediction modeling process per course. Interventions specific to courses followed.

Table 2. Organizational Chart for the Learning Analytics Initiative

Position	Scope of Work	Affiliation	Effort
Learning Science Team			
Faculty Director*	Advise on digital course enrichment, feature engineering, prediction modeling; lead grant writing, academic presentations & publications	College of Education	Course release or equivalent
Learning Scientist*	Liaison with faculty and units providing learning support. Oversee initial development of prediction models and coordination of learning supports. Lead evaluation of effectiveness of learning supports (i.e., on course grade & completion; RPC metrics).	College of Education	100% FTE; postdoctoral scholar with learning sciences background
Instructional Team			
Course instructor	(Previously) develop digital course materials; maintain a master course over multiple semesters so that prediction can be conducted; message students predicted to perform poorly	Home Academic Department	None
Instructional Design Graduate Assistant*	Support Learning Scientist; Conduct individual consultations with faculty partners; lead build, maintenance and inventory of digital contents enriched for target courses	College of Education (EPHE Dept)	12-month Graduate Assistantship
System Management Team			
Data Modeler	develop and maintain data ingestion infrastructure; build, validate, and maintain a data model per course; assist in prediction modeling; lead build of infrastructure to calculate student success projections to inform learning support.	IT Operations	25% FTE; IT Specialist, Operations or Institutional research
IT Support	Support Data Scientist; conduct ongoing integrity checking of existing models; assist in building of new data models, troubleshooting, and data management.	IT Operations	Hourly, work study or wage worker

3.2.2 Course selection

Candidate courses were first selected from those with the largest impact on university retention progression and completion metrics. The Provost's Office provided a list of courses with high enrollment and high rates of grades awarded that delayed students' progress towards the major. An exemplar subset appears in Table 3. The top 20 courses

were then cross examined for their current potential to provide features predictive of achievement, using digital content items already on LMS course sites. Content items reflect current development efforts, and could be supplemented with instructional design support to digitize existing print materials, link textbook content, or design new learning objects. Thereafter, assessment format and consistency (i.e., alignment to objectives, common exam item pool across instructors and course sections) was confirmed to ensure a robust and sustainable course design so the learning analytics implementation could be achieved.

Table 3. Courses with Highest Total Enrollment and DFW Rate with Digital Content Count

Course	Title	Enrollment	Digital Content Items
ENG 101	Composition I	2049	95
MATH 124	College Algebra	1442	33
PHIL 102	Critical Thinking & Reasoning	1301	0
PSC 101	Introduction to American Politics	1100	50
SOC 101	Principles of Sociology	967	131
PSY 101	General Psychology	962	75
COM 101	Oral Communication	853	135
MATH 95/A	Elementary, Algebra	327/816	20
MATH 96/A	Intermediate Algebra	381/814	15
BIOL 189	Fundamentals of Life Science	725	105

3.2.3 *Instructional Design to sustain robust features and criterion*

When instructor groups responsible for course delivery agreed to partner with the LAI to support their students, they agreed to a collaborative evaluation of their assessment process, involving alignment of exam format and items to instructional objectives listed in the course, and the design or acquisition of digital and print materials that addressed these objectives. This is a common instructional design process, but was new and intriguing for many in disciplines where formal instructional design training was not a requirement during preparation for a faculty role. The lead course instructor and LAI instructional designer effectively produced a master course which provided a standardized digital offering across all sections of the face-to-face, large-lecture course. This master course was to be used and unchanged for the two years of the initiative, after which course revision could be undertaken and prediction models could be refit as needed. This afforded protection of instructors' academic freedom – they could teach however they wished within the course schedule during course meetings – and simply adhered to the course schedule and design established by their department and required to ensure the durability of the learning analytics infrastructure (i.e., the digital component of the course – the LMS course site).

3.2.4 *Planning to adapt to change: Iterative redesign, partnership, and redundancy*

Universities are dynamic instructional contexts. Student enrollments fluctuate, degree program requirements and course policies shift, and the instructional goals of courses must adapt to accommodate new demands of students and programs, as well as new opportunities provided by learning technologies (e.g. new content, or features provided by a publisher) or developments in the field (e.g., new content that must be covered to provide a relevant, contemporary course). The course redesign and learning analytics refitting process accommodates most of these design challenges, but additional plans needed to be made to accommodate changes in staffing within the LAI, instructional group delivering supported courses, and campus units providing data services and student support. In order

to ensure LAI activities could continue uninterrupted in the event that a member of the Initiative left to assume a new role, redundancies were built into each Initiative team (Table 2). In addition, a master documentation infrastructure was maintained using the campus' collaboration platform. This requires specific tasking of lead individuals per team to maintain documentation, and the preservation of a version history to ensure redundancy of these materials to offset periodic file loss as is common in collaborative work.

3.2.5 Funding the initiative and partnering to scale student support

Whereas research solutions rely on external support to address research questions that align to the agenda of a program, the LAI was funded entirely with existing University resources. The Provost redirected existing technology funds to employ the postdoctoral Learning Scientist who manages the project. The College of Education directed internal funds to afford course releases to free the time of the Director and assigned funds from the graduate assistantship pool for the instructional designer. The Office of Information Technology already collects learning event data from the LMS to provide trouble shooting and reporting services. This unit was thus able to consider the project as falling within their existing offerings of instructional support; current employees' time managing these existing data, was allotted accordingly. The efforts of the Intervention Team are again comprised of pre-existing campus offerings to support (struggling) students. The Office of Online Education adopted and administers the digital resources (Figures 1, 2) developed during original learning analytics solution. Home departments accepted referrals to their supplemental instruction and tutoring programs for courses where these academic resources were already provided (though historically underattended and not by those failing). The Academic Success Center accepted referrals into their Academic Coaching program. Enrollments in these programs were monitored to examine the additional load.

3.2.6 Evaluating the initiative, and return on the Provost's investment

The Learning Analytics Initiative was initially funded for a two-year period, which afforded the opportunity to develop digital content and assessments, initial data collection and prediction modeling, then semesters of master course delivery during which prediction models could be applied and interventions provided. In these semesters, course achievement metrics from periods following interventions (i.e., subsequent exam performance, course performance, re-enrollment data) are monitored to examine the effectiveness of the interventions, and inform decisions about returns on nearly costless (i.e., digital, scalable interventions, Figure 1 and 2) versus resource intensive interventions. Summatively, the collective impact of the Learning Analytics Initiative can be observed by plotting achievement and retention metrics across semesters, and examining overall impact as well as impact for target populations (e.g. first generation college students). These data can inform whether the Initiative should receive continued or expanded university support.

4 GENERALIZING A LEARNING ANALYTICS RESEARCH SOLUTION ACROSS UNIVERSITY AND TECHNOLOGY CONTEXTS

4.1 Evaluating the generalizability of a learning analytics solution to new university contexts

The focus of this multi-university generalizability project was described to the funder as aiming to support, retain, and increase the achievement of undergraduates who traditionally do not persist in STEM majors (e.g., underrepresented minority groups, first-

generation college students) by (1) applying an existing data-driven solution to identify struggling STEM learners before they begin to fail, (2) developing targeted, effective achievement and retention interventions combining the expertise of two universities who are leading sources of empirically supported approaches for STEM success, and (3) demonstrating the applicability of these solutions to a variety higher education contexts.

The primary benefits to this generalizability effort are to provide higher education institutions with a proof of concept that learning analytics solutions developed at one institution can be adapted and employed at another, and to provide exemplar cases such that leadership of any future institution can select an exemplar that most closely mirrors their own institutional features and design their learning analytics solutions accordingly. This model further allows each institution to conduct a self-study of the resources available for collecting data on student learning (i.e., their LMS and other technologies for student support), their access to these data (Table 4), and their existing intervention resources that can be efficiently directed to support their students' success when prediction models identify students whose learning event data suggest a need for learning support.

The primary challenges to scaling such a learning analytics solution are three-fold. The first is to establish an appropriate collaborative stakeholder group similar to the one needed to scale the research solution in Case 1. The latter challenge is to map the data collection and prediction modeling solution to new learning management systems and data infrastructures that capture and can afford prediction of achievement using students' learning events. Whereas the first challenge is covered sufficiently in Case 1, this Case 2 requires description of the variability in campus infrastructure for data collection and modeling, and the various software contracts that dictate the LMSs that universities. Further, different universities serve different student populations and are staffed by instructors with different levels of commitment to the university (i.e., tenure-track and teaching-track faculty, adjunct instructors, etc.). Attention must thus be paid to partnerships with instructional units, so that they can design, sustain, and thoughtfully iterate through an established instructional design plan that guides course objectives, content, and assessment required to initiate and continue employing a learning analytics solution. A final challenge is to systematically examine the universe of interventions available, and to determine which are most likely to successfully meet the needs of struggling students.

4.2 Considering campus infrastructure for data collection and modeling

Across the three universities that serve as research sites for this generalizability project, learning management systems varied, and included Blackboard, Canvas, and Sakai. Further, the capacity of each institution's infrastructure, personnel, expertise, and budget for information technology platforms and personnel within operations management units varied considerably. These institutional uniquenesses led to variations in project staffing, as well as to the design of data collection and modeling platforms, even when two institutions supported the same LMS. To illustrate how these differences impacted the generalizability of the project, the parallel methods of tracing the same learning events are summarized in Table 4. The first column illustrates the nature of an event obtained from logs of servers hosting the Blackboard Learn LMS utilized in the original learning analytics solution. The methods required to collect the same events on the Sakai LMS appear in the second column. The third and fourth column demonstrate data collection methods for gathering

Table 4. Learning events obtained from Blackboard, Sakai, and Canvas Data and Canvas Live Events LMS Data Infrastructures.

Key values	Learning Process	Where found in Blackboard Learn server logs	Sakai (log)	Canvas Live Events (log)	Canvas Data (API tables)
Timestamp	temporally ordered event	in each event in log (date, HH:MM:SS)	same	same	available in request table
Student identifier	student-initiated event	LMS specific "duid"; can be matched via lookup to student ID	same	user_id in Body (metadata) JSON format	available in request table
Course, Section identifier	context-specific event	available in each server event as a registrar-provided value (lookup table)	same	same	available in request table
Navigation (links, folders)	context-specific event	GET request for specific "content_id" (metadata human coded)	same	same	available in request table
Downloading syllabus	planning course engagement	GET request, content_id	syllabus.read event in log	appears as asset.accessed	available in request table
Download of study guide	planning future study	GET request, content_id	webcontent.read event	appears as asset.accessed	available in request table
Use of Ungraded Self-assessment quizzes	rehearsal, monitoring learning	GET request "assessment_id" value; Note: this is a limitation of the server log approach, as assessment_id events collapse attempts at items and review of feedback on items into a single type of logged event.	assessment.event Metadata: assessment table (duration, score)	appears as asset.accessed event with dedicated identifier for course	event in request table; metadata in Quiz_submission_fact
Checking gradebook	monitoring performance	Identify GET request with specific, hard-coded tool_id for Gradebook (i.e., same identifier is applied in all courses, making this event simpler to capture than course-specific content)	gradebook.StudentView event	appears as asset.accessed event with dedicated identifier for course	event in request table; metadata in grading_period_score_fact

the same learning events from cloud-hosted instances of the Canvas LMS, but accessed through different methods. The Canvas Data API was chosen by the 4-year institution where the Blackboard model was first built. This university changed LMS providers after the project, and required a new data collection solution to sustain two additional research projects and to scale the original learning analytics solution into a service provision to meet university demand for data-driven student support (i.e., Case 1). The final partner on the generalizability project also utilized a cloud-hosted instance of the Canvas LMS, but had a pre-existing need to capture learning event data via a different data modeling tool – Canvas Live Events – that informed a software for accessibility and learner accommodations. Unlike Canvas Data, the Live Events API infrastructure captures only a subset of user activity.

5 CONCLUSION & SPECIFIC RECOMMENDATIONS

These cases overview scalability considerations in one institutional context and the way variants in infrastructure across contexts pose challenges to generalizability of a single solution. Hundreds of learning analytics solutions are in place in higher education, but few are evaluated for their effectiveness [12]. More exemplar cases need to be evaluated in order to identify principles that can inform design, and so that learning analytics solutions are worth scaling. From our development process and early efforts to scale and replicate a solution, we share some specific lessons we learned that can serve others aiming to adopt and scale similar learning analytics solutions in additional learning contexts, to customize their solution, and to maintain their fidelity once established (Table 5).

Table 5. Lessons learned from prediction modeling, intervention, and replication studies

#	Lesson
1	Start with instructional design and work closely with instructors. Learning about instructors' goals and designing content and assessments accordingly increases course quality and prediction accuracy. Well-designed master courses can be easily replicated to ensure consistency, minimizing variability that undermines predictions. Maintaining an active partnership limits drift in course design and implementation.
2	Satisfice when choosing a prediction algorithm. A model needs to produce accurate predictions and to also be programmable in order to produce them in real time. Simpler algorithms ease implementation in data models and enable timely action.
3	During model selection, select criteria for judging accuracy pragmatically. The purpose of the prediction model dictates how accuracy should be appraised. For example, identifying those likely to fail was critical and our intervention was not so costly that overapplication is a problem. Choosing a model with high recall was important to identify those needing support, and the risk was small: lower precision meant only that we suggested learning resources to some who did not need them.

6 ACKNOWLEDGEMENTS

This paper refers to data and analysis from project DRL 1420491 and design from DUE 1821594 and 1821601 from the National Science Foundation. The views expressed in this paper are the authors' and do not necessarily represent the views of the National Science Foundation. Additional support comes from the Office of the Provost and Offices of Information Technology and Online Education at the University of Nevada Las Vegas.

REFERENCES

- [1] Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM.
- [2] Bernacki, M.L. (2018). Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: trace data and their analysis. In D.H. Schunk & J.A. Greene (eds.) *Handbook of Self-Regulated Learning and Performance*. New York: Routledge.
- [3] Bernacki, M.L., Vosicka, L. & Utz, J. (April, 2016). *Can brief, web-delivered training help STEM undergraduates "learn to learn" and improve their achievement?* Paper presented to American Educational Research Association Annual Meeting, Washington, DC.
- [4] Broos T., Verbert K., Langie G., Van Soom C., De Laet T. (2018) Low-Investment, Realistic-Return Business Cases for Learning Analytics Dashboards: Leveraging Usage Data and Microinteractions. In Pammer-Schindler V., Pérez-Sanagustín M., Drachsler H., Elferink R., Scheffel M. (eds.) *Lifelong Technology-Enhanced Learning*. EC-TEL 2018. *Lecture Notes in Computer Science*, 11082. Springer.
- [5] Canvas Data. [Computer software]. Canvas Data Services. Retrieved from <https://community.canvaslms.com/community/answers/data>
- [6] Canvas Live Events. [Computer software]. Canvas LMS - REST API and Extensions Documentation. Retrieved from https://canvas.instructure.com/doc/api/file.live_events.html
- [7] Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 474-478). ACM.
- [8] Dominguez, M., Bernacki, M. L., & Uesbeck, P. M. (2016, July). Using learning management system data to predict STEM achievement: implications for early warning systems. In T.Barnes, M. Chi and M. Feng (eds.) *Proceedings of the 9th International Conference on Educational Data Mining*. Educational Data Mining Society. Retrieved from <http://educationaldatamining.org/EDM2016/>
- [9] Ferguson, R., Clow, D., Macfadyen, L., Essa, A., Dawson, S., & Alexander, S. (2014, March). Setting learning analytics in context: overcoming the barriers to large-scale adoption. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 251-253). ACM.
- [10] Hong, W. & Bernacki, M. L. (2017, June) A prediction and early alert model using learning management system data and grounded in learning science theory. In X. Hu, T. Barnes, A. HersHKovitz and L. Paquette (eds.) *Proceedings of the 10th International Conference on Educational Data Mining*, (pp. 358-359). Educational Data Mining Society. Retrieved from <http://educationaldatamining.org/EDM2017>
- [11] Sakai. (2018). Sakai [Computer software]. Retrieved from <https://confluence.sakaiproject.org/display/UDAT/Sakai+11.x+Event+Table+Descriptions>
- [12] Sønderlund, A. L. Hughes, E. & Smith, J.(2018).The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*. doi:10.1111/bjet.127208.

Predicting admission test success using SPOC interactions

Pedro Manuel Moreno-Marcos

Universidad Carlos III de Madrid
pemoreno@it.uc3m.es

Tinne De Laet

Katholieke Universiteit Leuven
Tinne.DeLaet@kuleuven.be

Pedro J. Muñoz-Merino

Universidad Carlos III de Madrid
pedmume@it.uc3m.es

Carolien Van Soom

Katholieke Universiteit Leuven
carolien.vansoom@kuleuven.be

Tom Broos

Katholieke Universiteit Leuven
Tom.Broos@kuleuven.be

Katrien Verbert

Katholieke Universiteit Leuven
katrien.verbert@cs.kuleuven.be

Carlos Delgado Kloos

Universidad Carlos III de Madrid
cdk@it.uc3m.es

ABSTRACT: In order to start Medical or Dentistry studies in Flemish universities, prospective students have to pass a central admission test to guarantee they have the proper level of proficiency. To support those learners, a blended program with a SPOC (Small Private Online Course) was designed on Edge edX. The logs from the platform provide a great opportunity to delve into the behavior of learners and to try to predict their success in the test based on students' interactions with the SPOC. This article has the following objectives: (1) analyze the differences of user interactions between learners based on their background, (2) develop and analyze predictive models to forecast who will pass the admission test, (3) discover which variables have more effect on success in this test, and (4) discuss about the generalizability of the solution. The results show that the SPOC learning behavior differs significantly between students with different background; it is not possible to predict success the admission test until the last months; and the average grade using only first attempts stands out as the best predictor.

Keywords: SPOCs, prediction, learners' success, learning analytics, indicators, generalizability

1 INTRODUCTION

In most countries, entry into medical schools is restricted by a high-stake admission test. In Flanders, this test consists of a scientific part with questions on chemistry, physics, mathematics and biology, and an information processing part. The passing rate fluctuates around 20%. The most influential success predictor is the prior educational track, giving students with a science and/or mathematics background (hereafter called “traditional students”) an advantage (Roggemans & Spruyt, 2014). As a result, students train intensively for the admission test to be optimally prepared.

In the digital era, technologies have enabled new ways to provide learning that can support those future students. With the popularity of online learning (and particularly with MOOCs, Massive Open Online Courses) because of its flexibility (Orlando & Howard, 2018), new kind of courses have appeared that use new learning facilities such as quizzes and video interactions. SPOCs (Small Private Online Courses) (Fox, 2013) have emerged as a way to use MOOC technology for specific on-campus training (e.g., for students enrolled in a course). Moreover, note that all these digital platforms not only serve as a repository to upload teaching materials, but they can also get comprehensive traces about learners' interactions, which can be very useful to detect patterns about students' behaviors and to predict trends on advance (e.g., who will pass the course) (Moreno-Marcos, Muñoz-Merino, Alario-Hoyos, Estévez-Ayres, & Delgado Kloos, 2018)

Prediction in education has a special relevance because stakeholders (e.g., teachers and students) can anticipate what will happen in the course, so they can adapt their teaching/learning behavior to improve. Furthermore, predictions can be presented through dashboards to aid sensemaking (Ali, Hatala, Gašević, & Jovanović, 2012), e.g., presenting information about students' success or students at risk (Park & Jo, 2015) to make students self-reflect on their learning. At this point, stakeholder engagement is very important and course builders and instructors should be involved in the design of visualizations, predictions, etc. (without neglecting students). However, although many people are involved, and accurate and meaningful predictions are obtained, a prominent issue is how to make the results generalizable because the course context can considerably affect the results.

Particularly, the course context and course design have special relevance in online or blended courses where learners are more at risk to procrastinate and need good self-regulation skills for success (You, 2016). That is the case for the SPOC KU Leuven developed to support last year high-school students to prepare for the chemistry component of the admission test. In that course, any student can enroll to access videos, theoretical background and exercises to prepare for the admission test. In particular for students from non-traditional study programs, the SPOC format would allow them to study at their own pace. However, it is not clear how the learning behavior in a SPOC to prepare for a high-stake admission test can influence success of the student and how results of the SPOC can be generalizable. In this context, this work aims to address the following objectives:

- Analyze the difference on grades and platform behavior between learners depending on their secondary school background
- Analyze the moment in which we can anticipate accurately if students will pass the admission test

- Identify the variables that have more influence on the predictive models to forecast success in the admission test
- Discuss about how to achieve the generalizability of the results presented in the previous objectives

2 RELATED WORK

In literature, there is an increasing interest in developing predictive models in education. Some of the most typical cases are related to forecasting dropout (e.g., Aguiar, Chawla, Brockman, Ambrose, & Goodrich, 2014) and student success (e.g., Ashenafi, Riccardi, & Ronchetti, 2015). Particularly in MOOCs, which have similar format to SPOCs although their contexts and characteristics of learners are different, Moreno-Marcos, Alario-Hoyos, Muñoz-Merino, and Delgado Kloos (2018) carried out a literature review on prediction. They found that dropout is the most-used outcome variable (e.g., Jian & Li, 2018), followed by final or assignment scores (e.g., Brinton & Chiang, 2018) and certificate earners (e.g., Ruipérez-Valiente, Muñoz-Merino, and Delgado Kloos, 2018). They also stated that there are many possible prediction features (although those related to platform use stand out) and indicated that new ones could be introduced (e.g., self-regulated learning variables, as used by Maldonado-Mahauad et al., 2018, to forecast success).

Among the most prominent variables to predict are test scores. For example, Okubo, Yamashita, Shimada, and Ogata (2017) used a Recurrent Neural Network (RNN) to predict the grade (between A-F) in a university course and compared the predictive power in the 15 weeks of the course. Fewer contributions focus on SPOCs. Yu (2018) used combined linear regression and deep neural network (DNN) to predict the final score of a computer science course. Moreover, Ruipérez-Valiente et al. (2018) predicted learning gains in a 0-course for freshmen students. This article presents a similar kind of study, although the logs and context (e.g. course duration and objective, pedagogy, etc.) are different. Finally, regarding state exams, Feng, Heffernan, and Koedinger (2006) developed a regression model to forecast grades in the exam based on interactions with an Intelligent Tutoring System (ITS). More recently, Fancsali et al. (2018) also predicted a math state exam from logs of their ITS (MATHia), such as solving time, knowledge components (KC) mastered, etc.

This paper presents a study that analyzes how admission test success can be predicted from the learning behavior in a SPOC and which variables affect the prediction. That contributes to the analysis of learning behavior in SPOCs and how it relates to student success. One of the differences with previous research is the identification of learning factors that are important in relation to the educational background (i.e. between learners whose background is appropriate or not for a certain bachelor) and success. Moreover, we innovate with new variables (e.g., variables related to the run of consecutive actions, pauses in videos, whether a student asks for the answer). In addition, the context is different (e.g., sequence of activities, pedagogy, etc.), there are reflections about the best moment to predict and models are not developed only at the end. Finally, we also include reflections about the generalizability of the solution, which are often neglected in articles about prediction.

3 METHODOLOGY

3.1 Case study and data collection

The study was carried out in a SPOC about chemistry, which was developed in Edge edX as a joint project of the Faculty of Science and the Faculty of Medicine at KU Leuven. The SPOC consists of 11 modules including 66 videos and 121 exercises, which cover the required contents for the chemistry component of the medicine admission test in Flanders. This entrance exam contains several tests, although this SPOC was focused only on chemistry. The SPOC was part of a blended learning support program: online modules were released gradually every fortnight (from September to May) and alternated with three face-to-face interactive sessions that used a flipped classroom approach, with the intention to stimulate SPOC learners to spread their learning activities over the year. Nevertheless, in practice many students enrolled late and they studied at their own pace. The target users were students in the last year of secondary school (in the academic year 2016-2017) who wanted to enter Medicine in any university in Flanders and paid a registration fee for the blended learning program. A total of 1,062 students accessed the course, although only 680 completed at least one exercise, and only 750 had interactions with videos.

For the analysis of data, two main sources were used. The first one includes the tracking logs from Edge edX (edX, 2018). Particularly, the following events have been considered: (1) *problem_check*, (2) *problem_show*, (3) *play_video*, (4) *pause_video*, (5) *seek_video* and (6) *stop_video*. The second source consists of the information about the self-reported results of 133 students of the science part of the admission test (which contains chemistry, physics, mathematics, and biology). The limited number of students completing the survey is a clear limitation of the study. As the sample is limited, all learners who have at least one access to the platform and completed the survey are included in the study.

3.2 Variables and techniques

Once the events from the tracking logs are filtered, high-level variables are derived to be used in the prediction models. Particularly, indicators are classified depending on their relationship with accesses to the platform, videos, and exercises. The list of considered features is shown in Table 1. The dependent variable is the binary result of passing/failing the test.

Predictive models have been created using the library *caret*¹ of R, and four of the most common algorithms have been considered: Random Forest (RF), Generalized Linear Model (GLM), Support Vector Machines (SVM) and Decision Trees (DT). With these models, results are obtained using 10-fold cross-validation and 10 repetitions. AUC (Area Under the Curve) is used to evaluate the quality of the prediction as this metric is widely used, generally appropriate for student behavior classification problems (Pelánek, 2015), and avoids some problems that other metrics face (e.g. accuracy) in imbalanced datasets (Jeni, Cohn, & De La Torre, 2013).

¹ <http://topepo.github.io/caret/index.html>

Table 1: Features used in the study.

ID	Variable	Description
Variables related to accesses to the platform		
1	streak_acc	Longest consecutive run of accesses to the platform
2	ndays	Number of days the student has accessed to the platform
3	avg_con	Average number of consecutive days that the student accesses the platform
4	per_pc	Percentage of accesses from a PC (and not from a mobile, tablet, etc.)
5	per_wk	Percentage of accesses during weekend
6	per_night	Percentage of accesses during evening/night
Variables related to interactions with videos		
7	per_vtotal	Viewed percentage of total video time
8	per_compl	Percentage of completed videos
9	per_open	Percentage of opened videos
10	avg_rep	Average number of repetitions per video
11	avg_pause	Average number of pauses per video
Variables related to interactions with exercises		
12	per_attempt	Percentage of attempted exercises over the total
13	avg_grade	Average grade of formative exercises (only using the first attempts)
14	avg_attempt	Average number of attempts in the exercises attempted
15	per_correct	Percent of correct exercises over attempted exercises (using all attempts)
16	CFA	Number of 100% Correct exercises in the First Attempt
17	streak_ex	Longest consecutive run of correct exercises
18	nshow	Number of times the user asks for the solution of an exercise (without submitting an answer)

4 RESULTS

This section is divided into four parts, which address each of the first four objectives that were introduced in Section 1.

4.1 Differences between learners based on secondary school background

In this section, we analyze the differences of students based on their educational background. The medicine admission test can be taken by any student finishing secondary school, but students from educational tracks with sciences and math (traditional students, TR) are better prepared for the test compared to students who do not have this background (non-traditional students, NTR). The aim of this section is to analyze if there are significant differences in the learning behavior depending on the educational background. To do that, data was separated in four groups: (1) traditional students who pass (TP, $n=92$), (2) traditional students who fail (TF, $n=22$), (3) non-traditional students who pass (NTP, $n=6$), and (4) non-traditional students who fail (NTF, $n=10$). In addition, we measured the difference of students with respect to all the variables of Table 1. As not many learners completed the survey, the number of cases of some groups is limited. Therefore, we used the Mann-Whitney test to compare the groups. Table 2 shows the results when comparing different groups and the mean of each variable for each group.

Table 2: Statistical comparison between traditional and non-traditional students.**P1: p-value TP-TF, P2: p-value TR-NTR, P3: p-value TP-NTP, P4: p-value TF-NTF**

Variable	TR	NTR	TP	NTP	TF	NTF	P1	P2	P3	P4
streak_acc	1.87	3.31	2.04	3.17	1.14	3.40	0.01	<10 ⁻²	0.04	0.01
ndays	12.18	18.44	13.76	22.17	5.55	16.20	<10 ⁻⁴	0.02	0.04	<10 ⁻²
avg_con	0.49	0.76	0.51	0.70	0.37	0.80	0.04	0.05	0.59	0.03
perc_pc	0.88	0.94	0.91	0.97	0.76	0.93	0.14	0.75	0.66	0.26
perc_wk	0.31	0.29	0.32	0.34	0.25	0.26	0.01	0.84	0.81	0.13
perc_night	0.07	0.05	0.07	0.01	0.09	0.07	0.69	0.50	0.43	0.92
perc_vtotal	0.57	0.78	0.62	0.88	0.34	0.71	<10 ⁻²	0.01	0.05	0.01
perc_compl	0.45	0.64	0.50	0.71	0.28	0.59	<10 ⁻²	0.02	0.07	0.02
perc_open	0.61	0.79	0.67	0.89	0.37	0.74	<10 ⁻²	0.04	0.12	0.02
avg_rep	1.24	1.52	1.39	1.83	0.62	1.33	<10 ⁻³	0.03	0.04	0.01
avg_pause	5.94	11.21	6.37	12.60	4.16	10.38	<10 ⁻²	0.01	0.05	0.01
perc_attempt	0.49	0.72	0.54	0.80	0.27	0.67	<10 ⁻³	<10 ⁻²	0.01	0.01
avg_grade	0.48	0.50	0.54	0.53	0.24	0.47	<10 ⁻⁶	0.61	0.78	0.03
avg_attempt	1.42	1.81	1.57	1.66	0.78	1.90	<10 ⁻³	0.03	0.97	<10 ⁻²
perc_correct	0.72	0.85	0.80	0.83	0.37	0.86	<10 ⁻⁴	0.18	0.73	<10 ⁻²
CFA	31.29	43.63	35.14	52.33	15.18	38.4	<10 ⁻⁴	0.04	0.05	0.01
streak_ex	5.38	7.25	6.10	7.67	2.36	7.00	<10 ⁻⁵	0.03	0.21	<10 ⁻²
nshow	40.17	69.19	43.22	79.33	27.41	63.10	<10 ⁻²	<10 ⁻²	0.01	0.01

* p-values under 0.05 (confidence level) are colored in blue

Results show that there is statistical significant difference in most of the variables (excepting *perc_pc* and *perc_night*) between TP and TF (no comparison has been done between NTP and NTF because of the few number of cases), which suggests that the learning behavior in the SPOC can influence success. Similar results are obtained when comparing all TR and NTR, with the exception of the variables related to user habits too (*perc_pc*, *perc_wk* and *perc_night*). Note that no statistical difference in some variables related to exercises achievement (e.g., average grade) were found for students who pass (TP vs. NTP). This entails that if the performance in the SPOC is similar, both groups can manage to pass. Nevertheless, NTP are more active on the SPOC as they access more often and watch more videos. Indeed, the SPOC format has the advantage that NTR, who have less background knowledge, can study at their own pace. Regarding the students who fail, there is statistical difference in most of the variables. NTR put more effort on the SPOC, and in some cases, they work harder than TP, as they access and watch more videos on average than TP. Their background seems to be a strong disadvantage however given the low number of NTP. To sum up, there are many differences in the behavior between TR and NTR and these groups should be treated separately to avoid bias in the models.

4.2 Anticipation of grades and results of predictive models

This section is focused on how success in the sciences part of the admission test can be predicted and more importantly, how early it can be anticipated. For that purpose, seven dates were selected (T_i) corresponding to crucial deadlines in the blended learning program (specific dates are in Table 3). T_1 , T_2 , and T_4 correspond to the face-to-face interactive sessions that were organized to discuss

problems on specific topics of the SPOC. At T3, traditional lectures were organized on topics that were not part of the SPOC, but that were crucial for the exam. The first session of the admission test was organized at T5, and the second at T6 (there were two sessions of the test to give a second chance to students who failed the exam). T7 includes all the interactions in the SPOC. With these dates, predictive models were trained for TR (NTR are excluded because they are very different from TR, and there are few students to develop models with representative samples, although it will be interesting to develop them if more NTR students appear in future editions) from the beginning of the course (September 7th) to each T_i . Table 3 shows the results of the models.

Table 3: Results of the predictive models (in AUC).

Period	T1	T2	T3	T4	T5	T6	T7
finish	22/10	14/01	07/04	06/05	05/07	30/08	
RF	0.46	0.45	0.70	0.78	0.84	0.87	0.87
GLM	0.59	0.71	0.72	0.73	0.74	0.77	0.77
SVM	0.55	0.51	0.72	0.73	0.84	0.85	0.85
DT	0.50	0.50	0.70	0.71	0.78	0.80	0.80

Results show that at the beginning of the course, the predictive power is poor. With an AUC threshold of 0.8 (as used by Moreno-Marcos et al., 2018), the predictive power of the model is only considered good from T5, the first session of the exam. A possible reason is the low activity at the beginning of the SPOC (57.45% of interactions occur after T4). If medium predictive performance (AUC=0.7) is acceptable (there is always a trade-off between anticipation and predictive power), the prediction from T3 can be considered. In that case, at least 31.03% of interactions are included, which is much more than the 13.43% in T2, which is not enough to predict. The low level of activity may also indicate that the SPOC does not really work in the synchronous way it was planned. That may affect the prediction because the activity is not uniform among students during the course. This can be important to reflect about the methodology. If face-to-face sessions with flipped classroom are organized, it would be advisable to enhance its relevance to ensure more people attends and are engaged from early stages.

In terms of the algorithms, the best model from T3 onwards is RF, which achieves an AUC of 0.87 at the end of the course. While differences are not big in some periods, this algorithm seems to be more consistent in this scenario. However, if the continuous grade was predicted and the RMSE (Root Mean Square Error) was used, SVM would be better (0.110 vs. 0.119), although both SVM and RF also perform better than the others.

4.3 Influence of variables on predictive models

After evaluating the predictive power of the models, the next challenge is to determine the variables that contribute most to the prediction, as this identifies the activities that are important for success. From the best model (RF in T7), the importance of the variables has been evaluated using the *Mean Decrease Gini*, which is often used to evaluate importance in RF (Louppe, Wehenkel, Suter, & Geurts, 2013).

The results in Table 4 indicate that the average grade of exercises using only the first attempt (*avg_grade*) is the most important variable. This is reasonable as correct answers at first attempt

indicate successful processing of the learning material, and after several attempts, the correctness of the answer can be affected by chance. Next, the number of days the user accesses (*ndays*) and the number of times the user asks for the solution (*nshow*) stand out. The last variable represents that students who demand and read the explanation of answers are more likely to pass. Regarding the variables about streaks, results show that long consecutive runs of correct exercises (*streak_ex*) have strong effect on success, unlike long consecutive runs of accesses to the platform (*streak_acc*). Finally, regarding video interactions, the variables that have more effect on success are the percentage of videos opened (*per_open*) and the number of times learners repeat the videos (*avg_rep*).

Table 4: Variable importance (VI) and correlation of variables of all students (CA) and traditional students (CT).

Variable	VI	CA	CT	Variable	VI	CA	CT
streak_acc	0.35	0.01	0.14	avg_rep	1.55	0.21	0.26
ndays	2.67	0.18	0.26	avg_pause	1.41	-0.05	0.07
avg_con	0.64	-0.06	0.04	per_attempt	1.33	0.20	0.30
per_pc	1.08	0.09	0.12	avg_grade	8.42	0.41	0.44
per_wk	1.13	0.15	0.14	avg_attempt	1.41	0.24	0.38
per_night	1.96	0.06	0.08	per_correct	2.14	0.35	0.47
per_vtotal	1.11	0.14	0.21	CFA	1.15	0.29	0.35
per_compl	0.62	0.12	0.19	streak_ex	2.12	0.32	0.40
per_open	2.10	0.17	0.24	nshow	2.34	0.10	0.21

5 DISCUSSION ABOUT THE GENERALIZABILITY

In Section 4, results of the prediction analysis in a SPOC were presented. Nevertheless, one important question is how results from this research can be generalized and extrapolated to different courses. Although results are valid for the analyzed SPOC, it is difficult to export the models because of the importance of the course context, which needs to be considered for the predictive models (Gašević, Dawson, Rogers & Gasevic, 2016). It may be possible to generalize the results in a very similar course (blended course with similar thematic), but results might change even in another run of the same SPOC if the context changes. For example, if more/less face-to-face sessions were organized, students might behave different and thus results may change. Similarly, if materials were all released at the beginning of the course or if students were required to do certain activities to continue after some deadlines, behaviors would also change and the interpretations of the results as well. Ocumpaugh, Baker and Gowda (2014) already experienced this problem when they developed EDM (Educational Data Mining) models to detect affective states with different populations and they analyzed whether their models were valid when changing the group of students.

Because of that, we believe there is no one-size-fits-all model to be used for all scenarios. Instead, existing models need to be taken and adapted to the specific context. This means that the scalability of the solution is about reuse and adaptation. For example, if we have different courses from the same source (e.g., several courses from edX), it is possible to use the same or similar algorithms to collect the indicators and train the models, but specific data of the course should be used, and the interpretations of the results should be done based on the methodology and pedagogical

background of the course. If the course has some specific features, perhaps some new indicators could be included as part of the adaptation. This way, each model would be specific for each context. Moreover, this approach also opens the door to a possible framework to guide learning analytics researchers and developers in the process of adaption of the models. While the context is different, there are several common steps in the adaptation, such as the reutilization of indicators, and future work should be focused on analyzing this process.

Related to this, there is a question about the validity of the research results. Even if the results can vary depending on the context, results provide insight in effective learning behaviors and when combining results from different scenarios, it may be possible to reach global conclusions about how students learn and what behaviors have relevant effect on their success. If we consider the case study presented in this paper, one finding has been the differences in the behavior based on the background. While it is possible to find another course where educational background may not be so important, e.g., a possible introduction course to something where all learners start from scratch, this conclusion raises the importance of the background, and particularly in admission tests (same context), and suggests considering it when adapting the models to other contexts.

6 CONCLUSIONS

In this paper, an analysis of SPOC data, including predictions for success on a high-stake admission test, has been done. One interesting finding was that there are strong differences in the behavior of students depending on their background. Moreover, prediction models only behaved reasonably well in the last three months, which were also the months with more than half of the activity. Among the variables, the average grade using first attempts stands out, although other behaviors such as accessing to the platform regularly, asking for the solution of exercises and repeating videos had also a positive relationship with success. The discussion of the generalizability also points out that the course context is very important and that makes models need to be adapted to be reused in each scenario. This also opens the door to the definition of a framework to guide people involved in learning analytics in how to adapt and reuse their models.

With regard to the limitations of the study, it is noteworthy that the dataset was limited due to the lack of information about the admission test results. Moreover, that information was self-reported data and, although it appeared reliable, it could only partially be verified (62% of the cases). In future work, it would be interesting to include data about more cohorts to improve the dataset (and be able to develop models for non-traditional students). Furthermore, it would be interesting to design and evaluate some visualizations based on the prediction results to provide SPOC learners with useful interventions based on their interactions. Finally, it would be also interesting to develop a framework about how generalizability can be achieved and reflect about other factors, such as the stakeholder involvement, which are also important to guarantee the sustainability of the learning analytics solution.

ACKNOWLEDGEMENTS

This work has been co-funded by the Erasmus+ Programme of the European Union, through the project LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), by the Madrid Regional Government, through the eMadrid Excellence Network (P2018/TCS-4307), and by the Spanish Ministry of Science,

Innovation and Universities with the project Smartlet (TIN2017-85179-C3-1-R). The latter is financed by the State Research Agency in Spain (AEI) and the European Regional Development Fund (FEDER). It has also been supported by the Spanish Ministry of Science, Innovation and Universities, under an FPU fellowship (FPU016/00526). This publication reflects the views only of the authors, and funders cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014, March). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the 4th International Conference on Learning Analytics And Knowledge* (pp. 103-112). ACM.
- Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education*, 58(1), 470-489.
- Ashenafi, M. M., Riccardi, G., & Ronchetti, M. (2015, October). Predicting students' final exam scores from their course activities. In *Proceedings of the 45th IEEE Frontiers in Education Conference* (pp. 1-9). IEEE.
- Brinton, C. G., & Chiang, M. (2015, April). MOOC performance prediction via clickstream data and social learning networks. In *Proceedings of the 34th IEEE International Conference on Computer Communications* (pp. 2299-2307). IEEE.
- edX. (2018). *EdX Research Guide*. Retrieved from <https://media.readthedocs.org/pdf/devdata/latest/devdata.pdf>
- Fancsali, S. E., Zheng, G., Tan, Y., Ritter, S., Berman, S. R., & Galyardt, A. (2018, March). Using embedded formative assessment to predict state summative test scores. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 161-170). ACM.
- Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006, June). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 31-40). Springer, Berlin, Heidelberg.
- Fox, A. (2013). From MOOCs to SPOCs. *Communications of the ACM*, 56(12), 38-40.
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68-84.
- Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction* (pp. 245-251). IEEE.
- Jiang, F., & Li, W. (2017). Who Will Be the Next to Drop Out? Anticipating Dropouts in MOOCs with Multi-View Features. *International Journal of Performability Engineering*, 13(2), 201-210.
- Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 431-439).
- Maldonado-Mahauad, J., Pérez-Sanagustín, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Delgado Kloos, C. (2018, September). Predicting Learners' Success in a Self-

- paced MOOC Through Sequence Patterns of Self-regulated Learning. In *Proceedings of the 13th European Conference on Technology Enhanced Learning* (pp. 355-369). Springer, Cham.
- Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies* (In Press).
- Moreno-Marcos, P. M., Muñoz-Merino, P. J., Alario-Hoyos, C., Estévez-Ayres, I., & Delgado Kloos, C. (2018). Analysing the predictive power for anticipating assignment grades in a massive open online course. *Behaviour & Information Technology*, 37(10-11), 1021-1036.
- Ocuppaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487-501.
- Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017, March). A neural network approach for students' performance prediction. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference* (pp. 598-599). ACM.
- Orlando, M., & Howard, L. (2018). Setting the Stage for Success in an Online Learning Environment. In *Emerging Self-Directed Learning Strategies in the Digital Age* (pp. 1-9). IGI Global.
- Park, Y., & Jo, I. H. (2015). Development of the Learning Analytics Dashboard to Support Students' Learning Performance. *Journal of Universal Computer Science*, 21(1), 110-133.
- Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2), 1-19.
- Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2018). Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. *Expert Systems*, e12298.
- Roggemans, L., & Spruyt, B. (2014). Toelatingsproef (tand) arts: een sociografische schets van de deelnemers en geslaagden. Brussel: Onderzoeksgroep TOR, Vakgroep Sociologie, Vrije Universiteit Brussel (140 blz.)-TOR, 29.
- Xu, B., & Yang, D. (2016). Motivation classification and grade prediction for MOOCs learners. *Computational intelligence and neuroscience*, 2016, 4.
- You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, 29, 23-30.
- Yu, C. (2018). SPOC-MFLP: A Multi-feature Learning Prediction Model for SPOC Students Using Machine Learning. *Journal of Applied Science and Engineering* 21(2), 279-290.

Adopting Learning Analytics at Universidad Austral de Chile

Julio Guerra, Ignacio Huichal, Eliana Scheihing, Valeria Henríquez

Universidad Austral de Chile

[jguerra, ihuichal, escheihi, vhenriquez]@inf.uach.cl

ABSTRACT: Higher education in Chile is going through systemic changes to improve quality due to several contextual issues imposing a strong pressure over high education institutions to innovate in their academic processes. In this context, learning analytics offers means to monitor, support and evaluate these innovations. However, adopting Learning Analytics is not a straightforward endeavor, because a real adoption of LA supposes appropriation that is scalable and sustainable. In this paper we summarize an ongoing process of adopting two analytics tools for different users and context in Universidad Austral de Chile. The process involved active participation of users and stakeholders, strategies to direct concrete and focused discussions with a diversity of actors, and the work of adaptation of tools comprising usability and usefulness validations.

Keywords: adoption of learning analytics, latin america, higher education

1 INTRODUCTION

Higher education (HE) in Chile is going through systemic changes to improve quality due to several contextual issues. In the one hand, access to HE, a highly selective system, has massified during the last decades without solving the deep gap of quality between private and public secondary education offer. In the case of smaller and regional universities such as Universidad Austral de Chile, this phenomena implies that more students come to the university lacking the needed background knowledge and skills. As a result, dropping rates and academic failure become an important problem in this type of institutions. In the other hand, and as a systematic effort to improve the HE quality, Chile has advanced in creating a system of quality assurance of educational institutions centered in the National Commission of Accreditation. These efforts aim to increase rates of academic success (lowering dropout rates) through a continuous process designed to verify the accomplishment of several quality criteria defined by law. As part of this process, the institutions (Universities and other HE institutions) are requested not just to improve their academic indicators, but to track and provide concrete evidence of their current academic and improvement processes.

At Universidad Austral de Chile, these requirements are being addressed by a series of curricula innovation and complementary learning support efforts that permeates the whole institutional structure, fostering a rich discussion across levels, from institutional direction to practitioners and students. Here lays a strong opportunity for Learning Analytics (LA). LA is defined by The Society for Learning Analytics Research (SoLAR, <https://solaresearch.org/>) as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." LA brings in knowledge, frameworks,

techniques, and experiences that could be applied, for example to bring an objective perspective (showing academic data) into the discussion of course sequences while introducing innovation into a career program; or to provide at glance academic status of a student that files a special request to register courses, or who needs orientation on which courses to take.

However, adopting LA is not as straightforward as we may think: the process of real adoption supposes appropriation that is scalable and sustainable. This is a complex process in a complex institution such as a university because involves considerable discussion, agreements, and new policies. This work presents an ongoing process of adopting two LA tools at Universidad Austral de Chile, in which we have applied an approach that consider the involvement of all level stakeholders and top-down and bottom-up strategies to align needs and views from high level institutional directions, to practitioner and students. The effort is being performed within the activities of the Erasmus+ project "LALA: Building Capacity to Use Learning Analytics to Improve Higher Education in Latin America" in which Universidad Austral de Chile participates.

In the next section, we give details about the context and several aspects that we have identified as potential keys for the effective adoption of LA in UACH. 'Potential' because this is an ongoing process and we moderate our conclusions accordingly. Following section describes the process and the resulting prototypes of the tools, focusing in the design rationale behind the adaptation efforts. Then we provide our insights to motivate discussion towards the adoption of LA.

2 CONTEXT

2.1 Higher Education in Chile

The higher education (HE) in Chile is organized by the reform of 1981 in a mixtured system of public and private institutions, including universities and technical centers. This reform also reduced the economical involvement of the Chilean state in the cost of HE, leaving the most of the education costs to be paid by the students and their families with the support of educational loans. The reform of 1981 produced a massification of the offer and the access to HE: the system went from 8 universities to more than 70 institutions of higher education, with an increase of total enrollment in the universities of more than 600% (100 thousand students in the year 1980 to 650 thousand in the year 2017) (Lemaitre, 2018).

The HE system is selective, based in a national entry test. Applicants compete to access to the best HE institutions. Since the primary and secondary education also implements a public/private approach, the economically upper section of the chilean population access to the best private schools that prepares them better to score in the national HE entry test, thus could access to the higher end Universities. Lower scoring applicants can still access to other non-selective institutions that are equally or more expensive, but that are usually poor quality.

The privatization of HE offer and the massification of the access have failed to positive impact in the quality and the pertinence of the system, and have failed bridging the gap of inequality of the Chilean society. The situation motivated large and sustained social movements starting from Revolución Pingüina in 2006 at the secondary education, and the later university student movement

in 2011 (Bellei, Cabalin & Orellana, 2014). From the student movement of 2011, the Chilean society began to demand greater equity in the entrance to the universities and improvement of the quality of the HE system overall. These demands were gradually accepted with partial gratuity policies and a boost in quality assurance mechanisms by the Chilean state (Labraña, 2018). As a consequence of the first, access to higher education has been extended to 40% of high school students including young people who are the first generation in their family to attend university, and that come from semi-rural areas. With regard to the quality, the main measures used by the State are the quality assurance systems, some state financing mechanisms moderated by the performance of institutions, and university rankings (Reyes, 2016). This framework generates in higher education institutions the need to systematize and quantify in detail all their educational processes.

2.2 Universidad Austral de Chile

The Universidad Austral de Chile (UACH), founded in 1954, is one of the twenty-seven traditional universities of Chile. It offers 55 undergraduate degrees, 36 master's degree, 11 doctoral programs and 7 technical courses. UACH counts with 14,202 undergraduate students and 928 graduate students, 1364 full-time teachers who generated 542 WOS publications, 158 Scielo publications and 572 Scopus publications in the year 2016. It has 295 agreements with educational institutions in 37 countries, and 8 programs accredited internationally. Despite of being considered a small university, UACH is well ranked among Chilean universities occupying the sixth place at the national level according to the latest (2018) Times Higher Education ranking. Being located in the south of Chile, which is essentially a rural zone with an agrarian economy, UACH has become a preponderant social actor in the development of the southern Chile.

The scenario of HE in Chile has had a strong impact in Universidad Austral de Chile. As explained before, the massification of access in the selective Chilean HE system, has lowered the academic entry level of the students applying to regional and smaller institutions: more students come to the university lacking the needed background knowledge and skills. This leaves the institution in a complex scenario to reach quality indicators, such as the approval rates and time-to-graduation, which are now increasingly demanded by the quality assurance policies implemented by the State. This situation has motivated deep efforts such as adopting a competency-based model to re-design career program structures, continuously course curricula innovation, and incorporating new evaluation methods (Lemaitre, 2018).

2.3 Opportunity for Learning Analytics

The configuration of the educational context imposes a series of requirements for the universities in Chile, and to UACH in particular, to develop mechanisms of continuous improvement in all their academic and administrative processes. Such mechanisms imply the need of continuous measurements, and feedback processes that involve actively collect and analyze academic data at all levels, from micro-curricular data such as intra-course traces and grades to assess pedagogical innovations and learning activities, to macro-curricular data such academic paths to assess study programs innovations. The good news is that universities have data already, systematically at the

macro level: traces of students academic trajectories such as course registrations, grades, entry tests, students' demographics, among others. The scenario presented is then a fertile ground to adopt Learning Analytics in a broader sense.

However, adopting LA in a complex institution is not a straightforward endeavour. The adoption has to be aligned with the high level institutional requirements and policies, and at the same time, has to be useful and pertinent to provide support for practitioners and students. These considerations motivates the project Erasmus+ KA2 "*LALA: Building Capacity to Use Learning Analytics to Improve Higher Education in Latin America*" (LALA from now on) in which UACH participates.

3 PROCESS

The process of adopting Learning Analytics in the institution starts by following the LALA Framework. This framework is inspired in the approach developed in the SHEILA project (Tsai, Y. et al, 2018) and later adapted to the Latin-American context by the LALA Project. The framework involves diagnostic, socialization, community building around LA, and the exchange of know-how and experiences. The goal is to advance in the development and adaptation of LA tools keeping a high level view of the institutional context, requirements, policies and needs.

3.1 Diagnosis phase

As a first step, the LALA Framework deploys diagnostic processes to assess the needs, preparation and expectations of using LA in a broader sense and considering all stakeholders: decision makers, administrators, teachers, students. During this diagnosis stage of the LALA Framework, we performed:

- 11 semi-structured interviews that last from 30 minutes to 1 hour and involved different stakeholders, including three directors of undergraduate programs, the director of the undergraduate school, the director of the institutional analyses office, the director of learning support unit, the director of TI department, the Dean and pro-dean of the Engineering School, a staff of accreditation office, and the director of the student services unit
- two focus groups involving 15 teachers from different departments
- one focus groups with 5 students of one of the engineering programs

The diagnostic phase allowed us to identify needs that find alignment at different levels of the institution, and at the same time, it fostered initiate discussion and generated expectations. At this point we recognize the important backing up role of participating in the LALA project to open doors and get attentive ears. Overall, reception to LA is highly positive and stakeholders value the idea of supporting their work with LA tools. Needs detected included the systematization of counselling process such as academic and personal advice to students, visualizing academic trajectories for curricula innovation evaluation, analyzing factors influencing dropout, visualizing courses' historic academic data to reflect on the evolution of the teaching (teachers), being able to compare academically to peers (students), and feedbacking students with data collected by the institution

using surveys. To move forward, detected needs were analyzed considering the potential impact of the solutions and a reasonable effort to concretize, before selecting needs to address.

3.2 Analysis of needs vs reasonable effort

To select needs of LA to address we considered several aspects that we think are important to be on track for LA adoption:

Alignment of institutional and practitioner levels. Needs that resonate at different levels of the institution are more likely to receive attention and produce commitment of all relevant actors. Different level stakeholders involvement is important since it makes possible to have richer discussions about how potential solutions could address the needs broadly, considering aspects that otherwise could be neglected, such as security concerns for technical access to data and privacy issues. This is the case of the need of supporting the counselling work that program directors at UACH perform especially at the beginning of each semester due to the high rate of students in 'special' academic situations and who file special course registration or cancellation.

Data available right now! Research in LA has growth strongly using large collections of fine-grained traces of learners, specially while using online learning environments. While this is true, we put our feet on the ground and "limit" the needs to be addressed and ideas of solutions to the data that is currently available. This allows to focus discussions to short term reachable and concrete solutions. For example, stakeholders could envision very useful artifacts that could keep track of students academic situation considering formative and summative assessment and grades. However, partial grades of the students are not systematically being collected by the information systems which makes it difficult for LA to try to face this need. The data that is actually available consider courses' final grades, thus we 'limit' expectation to the exploitation of this data.

Existence of a tool to start with. Bringing a concrete example in as a ground-zero artifact makes the process more effective and efficient. It allows different stakeholders to concretize ideas, concerns, comments and suggestions, and serves as a canvas to reasonable new ideas and adaptation features. In our case, a pool of existing tools were provided by the partners of the LALA Project. Partners did not provide only the tools, but the experience behind adopting and developing the tool in their institutions.

3.3 Selection of existing tools

Considering the criteria exposed in the previous sections, we move forward with two LA ideas to address two needs that also finds existing tools to start the adaptation process.

Firstly, the visualization of academic records of course trajectories to support monitoring of curricular progress. This idea finds support at different levels, from students that want to be able to see their progress in comparison to others, to teachers, that want to see their courses and their academic indicators, to directors that want to see overall curricular picture. The existing tool that we

selected to start the adaptation is LISSA (Charleer et al, 2018). LISSA presents curricular trajectories of individual students and it is used to help the face to face counseling process.

Secondly, UACH applies every year a questionnaire of learning skills and self-concept to the freshmen students with the main goal of reporting back to program directors about the aggregated profile of the incoming cohort. Currently the results of these surveys are not presented back to students. According to the Learning Support Unit at UACH (the unit that applies the questionnaires), there is an opportunity of presenting individual feedback to students to foster self-reflection and help promoting the work of the learning support unit to reach student population. The existing tool to be adapted is LASSI (Broos et al, 2017). LASSI was designed to feedback students with their responses of an homonymous survey about learning skills. The tool helps students to reflect on the importance of learning skills in their learning process and the relation between study efficiency and learning skills.

3.4 Co-design phase

Taking the existing tools as a starting point for the development of adapted solutions, we implemented a co-design process with the different stakeholders. Regarding the need of monitoring academic trajectories, the LISSA was originally presented to program directors and academic administrators to collect initial impressions. The process moved forward with an iterative development process involving a series of requirements and validation meetings with one program director. The result of this process was a prototype that was later evaluated with other three program directors.

Regarding the need of presenting the results of surveys to students, the tool LASSI was initially shown to staff of the Learning Support Unit who were enthusiastic on the idea of adapting it. An initial prototype with the survey information in spanish was presented as mockups to the same unit and later evaluated in a user study with seven students in a group session that lasted 50 minutes. A functional prototype was then developed and validated through periodic meetings with the staff of the Learning Support Unit and finally tested in two sessions involving a total of five students.

4 RESULTING TOOLS

As mentioned before, two main needs were selected to move forward in adapting and deploying solutions based on LA. Respectively, the work focused in two tools that are were named TrAC and VERA.

4.1 TrAC for directors of programs

TrAC (from the spanish Trayectoria Académica) visualizes all the relevant information about the academic trajectory of the student over a layout representing the structure of semi-flexible study programs, and give a overall view based on concrete data (Figure 1). Directors of programs can use TrAC to support decisions regarding special request for students, such as dropping courses, allow to registration of courses with unfilled course requisites, among others.

To design TrAC we used an interactive approach of redesign with program directors, having weekly meeting and using LISSA Dashboard as a starting point of discussion. Since the tool is mainly a visualization tool, the focus of the discussion was how can we visualize each of the functional requirement. Therefore, the tool was built in an iterative process using semi-functional prototypes. From an initial broad view of the need, the tool was quickly focused in supporting registration applications, for which some of the characteristics of the program structure and academic information popped out as relevant to be shown: course grades and comparison with averages of the course. While the original tool, LISSA, supported these aspects, there was an fundamental aspect not covered: since the curricular structure of the programs is mainly fixed at UACH, the academic progress of a student gain meaning if it is overlaid in top of the (fixed) program structure. This mean that when we talk about a semestre, we could referer to the semestre of the program structure (the I, II, III, IV, V... semestre of the program), or to the relative semestre of the student in the program (e.g. the fifth semestre of a student in the University). All students start with the same set of courses pre-registered in the semestre I, then depending on their individual progress, students deviate from the "ideal" pathway. Visualizing the trajectory of specific students on top the program structure, allow to identify for example, un-balanced pathways, or deep delays. It also generated the requirement to see states of progress at different student's relative semesters.

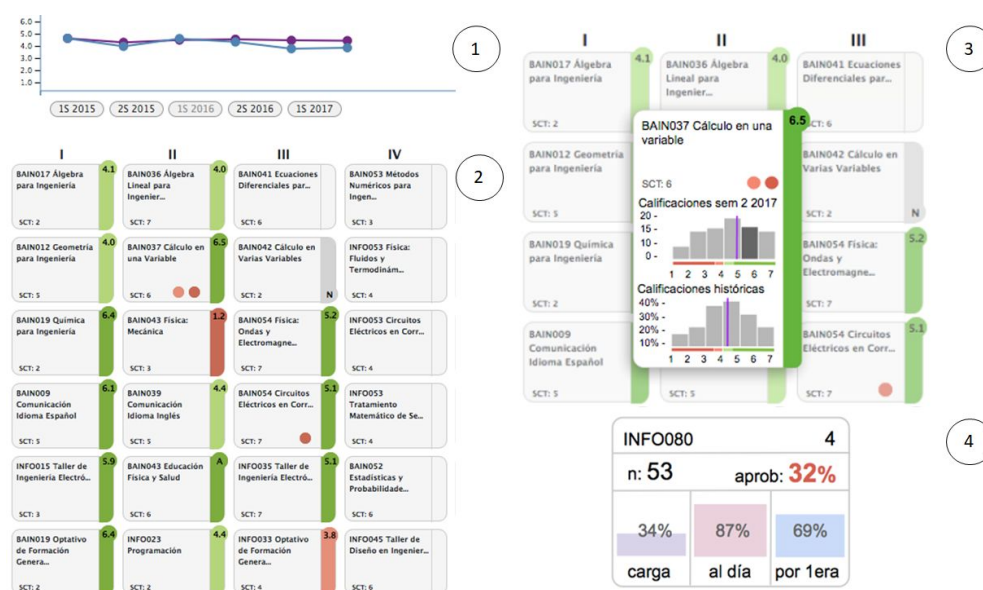


Figure 1. TrAC dashboard elements.

Figure 1 shows the different visual elements of TrAC dashboard. (1) Shows the average grade of the student in each term. Users can see the historical data of the student clicking in each term, as snapshots of the situation of the student's specific semesters of his/her academic life. Figure 1 (2) shows the student program structure, the trajectory and the performance of a particular student on it. Each box represents a course with the name of the course, number of credits (SCT), and final grade. Different visual elements are used, such as the color in the right bar that represents the grade in the last time the student had taken the course. Colors represent grades ranging 1.0 – 3.5 (red), 3.5 – 4.0 (pale red), 4.0 – 4.5 (pale green) and 4.5 – 7.0 (green), of a 1-7 score grading system which

passing grade is generally the middle point (4.0). Small circles represent previous tries of the course allowing to see course repetitions. Figure 1 (3) shows more detail about the course and includes two histograms. The first (top) histogram shows the distribution of grades of the class, in darker grey where the student is located. The line represents the median. The second histogram shows the historical distribution of grades (considering previous versions of the course) and presents a baseline to judge the first histogram. This is specially relevant when courses have been modified or taught by other instructors. Figure 1 (4) shows a different visual element to represent a course, and correspond to a design variation of the dashboard to be used for teachers and it is described in the following section.

4.2 TrAC for teachers

TrAC helps teachers providing information to develop a well informed reflections about the results of their courses. TrAC give a overall view of indicators of each course and the evolution of them during the time (Figure 1 (4)). The design of TrAC for teachers followed a co-design methodology implemented through a face to face working activity teachers of different schools. The main goal of the activity was to reach a common view of a useful tool, starting from scratch in a blank piece of paper, but with a very strong restriction that was the available data to develop the tools: academic records and program structures. Participants were told explicitly that no other information (inside course grades, attendance rates, demographics of students, etc) was available. The activity was separated in four working blocks and participants join one of four work groups. In the first block, each group performed collaborative work to generate ideas of visualization to support common goals expressed as questions "which data and statistics will allow you to i) summarize the evolution of your course, improve your course and prepare the next version? The second block was the sharing phase in which each team showed their own work and explained to the rest of participants. In the the third block participants joined different groups and collaboratively worked in refining design result. The fourth block was the closing phase, making a plenary discussion about the outcomes and the (expected) usefulness of the tools on their daily activities.

Form the codesign experience we differentiate four views of the aggregated academic information that were relevant. The main view is the distribution of grades on the course and its evolution through the versions of the course (past versions). Teachers use to relate performance in their courses with different aspects that were elaborated in other 3 views: the parallel workload of the students taking the course, the relative delay of the students (are they taking the courses too late?), and if the students are taking the course by the first time, the second time or even third or fourth time. Figure 1 (4) shows a box representing course in which the described information is summarized along with the identificator of the course, number of credits (SCT), number of student. The designed box includes the percentage of passing students (passing rate), percentage of students taking the course with the expected workload (expected workload according to the program structure), percentage of students taking the course in time (according to the program structure), and percentage of students taking the course for the first time. We are currently working on an extended view of this data to unfold details when clicking in the course-box.

4.3 VERA for resenting freshmen survey results

Every year, new UACH freshmen students answer several surveys applied to collect psycho-educational information about self-concept (García & Musitu, 1999) and learning strategies adapted to chilean context (Truffello & Pérez, 1988). This information is later processed, aggregated and combined with socio-demographic information to inform each of the schools about the characteristics of the incoming cohort. Currently, the survey responses are not informed to students, and the Learning Support Unit recognizes this as a need. VERA (from spanish "Visualiza Encuestas para Reflexión Académica") addresses this issue showing each students her responses in the surveys and complementing the information with same cohort surveys' answers and past cohorts performance (in aggregated level) to promote self-reflection and help-seeking. As mentioned before, VERA is an adaptation of the tool LASSI (Broos et al, 2017).

Figure 2 shows a screenshot of VERA in which four sections has been numbered to facilitate explanations. (1) the tabs represent the different questionnaires answered by the students; in this case self-concept and learning strategies. (2) The tabs represent the dimensions evaluated in the questionnaires and Figure 2 shows the emotional dimension of the self-concept survey. The results are divided in three figures (bad, medium and good result from left to right) and each dot represent a student of the same cohort. Figure 2 (3) shows historical data of students relating their survey responses with their academic performance on the first year. The colors means the level of success; a green square means 1% of students passing all the courses, yellow if fail one course, and red if fail two courses or more. Figure 2 (4) shows "what to do next," comprising recommendations, resources available and contact information of the Learning Support Unit.

VERA was designed by the LA team using LASSI as starting point and it didn't need many adaptation because of the similarities to the original use context and the current UACH need and surveys applied. VERA was exposed to end users in several low scale user studies aimed to validate that the information is understood by students without additional support, their perception of utility and their perception of seeing their own sensible data.

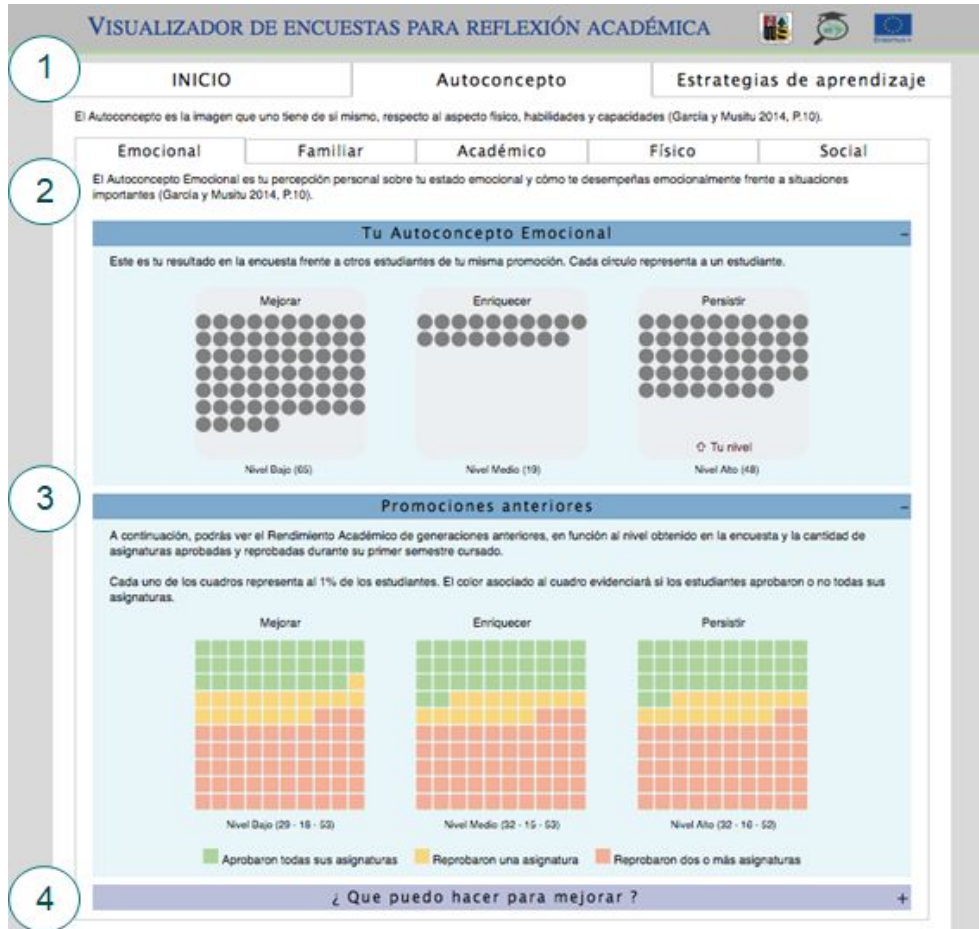


Figure 2. Screenshot of VERA based on LASSI tool.

5 DISCUSSION

As far as we know there is no one-size-fits-all approach for adopting Learning Analytics in a sustainable and scalable manner. To ensure continuous and effective use of the new tools, the process has a broad dimension that involves not only the adoption of technological artifacts, but the adoption of policies, practices and knowledge and supposes the involvement of different levels of the institution stakeholders and users. Following initial steps in the adoption of LA in UCh, we extract several ideas that we found relevant to carry out this process.

- Sense the needs and expectations for LA at the institution. The LALA project provides an umbrella to perform guided diagnosis and bring ideas of LA, opening discussions and setting up expectations.
- Engaging diverse stakeholders is not easy. They are busy people, of course. Also, they probably have a different level of understanding of the problems, and interact with these problems in different context. This can lead to a richer discussion (addressing a problem from different angles), but has the risk of produce a deaf discussion. This is why we recognise the need to present them a discussion around a concrete idea. For example, in our

institution we found an overall need to monitor, at different levels, the academic information related to the program curricula. Then a very fruitful meeting resulted of sitting together high level decision makers and technicians where an example tool was shown. While directors could see how this tool represents high level indicators needed to analyse bottlenecks in curricula structures, technicians could reason about implications of getting and processing the data needed. Surprisingly, privacy concerns arose from the high standards of data security of the technical department, which motivates further considerations regarding access to academic records by different types of users.

- Provide means to focus discussions and work in concrete and realizable ideas. We used two strategies to foster concrete discussions: i) bringing an example tool as a baseline foster discussions and provides a canvas to see similarities and differences; and ii) focusing in ideas that could use only data that is currently available. We recognize this as an effective strategy, specially to open the institution to adopt LA. We acknowledge that there should be spaces for planning for more complex tools that needs collecting other data.
- Adoption needs adaptation, anyway. Adoption of the tools is possible if these tools have value and are usable for the users of the institution. Tools will need some level of adaptation that could depend in many factors. We addressed tool adaption using artifacts and methodologies borrowed from agile software development and usability testing, because they count among our team expertise.

We value these insights and keep a positive expectation. But we want to acknowledge that is is an ongoing process. Currently, we are working on a piloting stage in which the described tools will be tested within the institution in real usage scenario.

Acknowledgment

Work funded by the LALA project (grant no. 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). This project has been funded with support from the European Commission. This publication reflects only the views of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- Bellei, C., Cabalin, C., & Orellana, V. (2014). The 2011 Chilean student movement against neoliberal educational policies. *Studies in Higher Education*, 39(3), 426-440.
- Broos, T., Peeters, L., Verbert, K., Soom, C. V., Langie, G., & Laet, T. D. (2017). Dashboard for Actionable Feedback on Learning Skills: Scalability and Usefulness. *Learning and Collaboration Technologies. Technology in Education Lecture Notes in Computer Science*, 229-241. doi:10.1007/978-3-319-58515-4_18
- Charleer, S., Moere, A. V., Klerkx, J., Verbert, K., & Laet, T. D. (2018). Learning Analytics Dashboards to Support Adviser-Student Dialogue. *IEEE Transactions on Learning Technologies*, 11(3), 389-399. doi:10.1109/tlt.2017.2720670
- Labraña, J. (2018). La primavera chilena: Ni conservadora ni revolucionaria. Una explicación sociológica del significado histórico del movimiento universitario chileno del año 2011.

- Calidad En La Educación*, (48), 251. doi:10.31619/caledu.n48.476
- Lemaitre, M. J. (2018). Mecanismos de aseguramiento de la calidad: Respuestas a los desafíos del cambio en la educación superior. *Calidad En La Educación*, (21), 87. doi:10.31619/caledu.n21.323
- Reyes, C. (2016). Medición de la calidad universitaria en Chile: La influencia de los rankings. *Calidad En La Educación*, (44), 158-196. doi:10.4067/s0718-45652016000100007
- Tsai, Y., Moreno-Marcos, P. M., Tammets, K., Kollom, K., & Gašević, D. (2018). SHEILA policy framework. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18*. doi:10.1145/3170358.3170367

DIY: learning analytics dashboards

Martijn Millecamp

Dept. of Computer Science, KU Leuven
martijn.millecamp@cs.kuleuven.be

Tom Broos

Dept. of Computer Science, KU Leuven
tom.broos@cs.kuleuven.be

Tinne De Laet

Tutorial services, faculty of Engineering Science, KU Leuven
tinne.delaet@cs.kuleuven.be

Katrien Verbert

Dept. of Computer Science, KU Leuven
katrien.verbert@cs.kuleuven.be

ABSTRACT: The number of data generated by educational technologies is increasing every day. To make sense of this overload of data, learning dashboards are becoming more and more popular inside the learning analytics field. However, most of these dashboards are implemented in a very specific context and are not easily scalable to other contexts. To use these dashboards in other contexts, there is a need for guidelines to adapt and create learning dashboards. To address this need, we developed a guideline to identify the context of the learning analytics dashboard as a first step in the process of adapting and creating learning dashboards. To test our guideline, we held a workshop with 12 participants at in Riobamba, Ecuador that resulted in a modified version of the guideline. This final guideline states that to identify the context of a learning dashboard, at least the objective, the stakeholders, the interactions, and the key moments have to be identified.

Keywords: Learning Dashboards, Scalable, Latin America, Learning Analytics

1 INTRODUCTION

According to American business magazine Forbes, we are producing over 2,5 quintillion bytes of data each day¹. This incredible amount of data is mostly due to social media, communication and the IoT. However, a vast amount of this data is created by educational technologies, such as learning management systems (LMS), virtual learning environments (VLE), and massive online open courses (MOOC). This increase of data led to the opportunity to use these online traces of learners to improve their learning and to the birth of the field of Learning Analytics (LA).

¹<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>

One of the possibilities to understand the meaning of these large amounts of data is by showing them in some kind of visual display, also called a learning analytics dashboard (LAD), because our human brain is able to process lots of data as long as it is presented in a meaningful way [7].

In the past decade, the field of LA has been growing and this also resulted in an increasing number of learning dashboards. However, this growth took mostly place in European, Australian, and American higher education institutes. Other regions such as great parts of Latin-America lacked the local capacity to gather, manage, process, or visualize the data and to enable the growth of LA in Latin-America.

Due to the past years modernization, there are opportunities to close this gap within the LA field. The previous lack of capacity does not need to be a disadvantage as this enables Latin-American higher education institutes (HEI) to lean on the shoulders of the existing LA research to become one of the leading regions of LA [6]. However, to enable HEIs to build on top of existing work, there is a need for guidelines to adapt or to create good LADs.

In this paper, we will first give a definition of a LAD and position LADs in the field of LA by looking at existing literature. Next, we go deeper into the problem of scalable LADs and propose different steps to follow when adapting/creating a LAD. We then go deeper into the first step of this process and present the results of a workshop to finetune the identification of the context. Concluding remarks as well as future work end this paper.

2 RELATED WORK

2.1 Learning dashboards

During the past years, learning dashboards have been widely used and researched in the domain of LA [1]. In this research, there are different synonyms in use such as ‘educational dashboard’, ‘learning analytics dashboard’, etc. Unfortunately, there is not only a variety of synonyms, but there are also a variety of definitions of a learning dashboard. In this paper, we decided to use the following definition, mostly based on Schwendimann et al. [7]:

A learning dashboard is an interface that aggregates multiple visualizations to create a holistic view about learner (s), learning process(es) and/or learning context(s).

In Figure 1, the whole field of LA is visualized as the biggest dark grey circle. In this field, there are multiple subdomains, visualized as light grey circles. In the center, the subdomain of learning dashboard is shown in blue. As this figure illustrates, LAD's are not considered as an independent subdomain as LAD's can be used to visualize data coming from a variety of different LA subdomains with the different sub-domain which are illustrated by overlapping the blue LAD's circle.

Despite the increased popularity of LADs in recent years, most of the dashboards are still used in either a scientific, and thus small and unscalable, setting or in a commercial, large scale setting where there is no proof about the impact or the perception of the dashboard². Only in recent years, various

² <http://blog.associatie.kuleuven.be/tinnedelaet/category/learning-dashboards/>

attempts have been made to deploy learning dashboards in a scientific, large-scale setting such as the dashboards of Millecamp et al. [5] and Broos et al. [2].

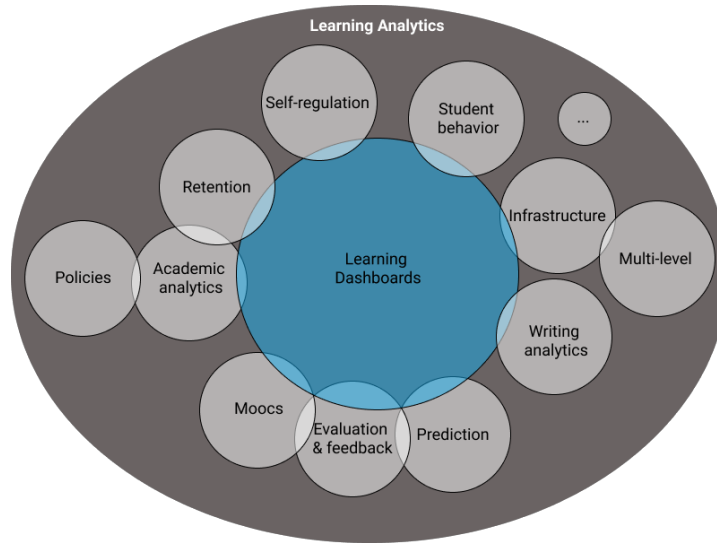


Figure 1: The field of LA (dark grey biggest circle) with different subdomains (smaller grey circles). The subdomain of learning dashboards (blue central circle) overlaps with a lot of different other subdomains of which it can visualize the data.

2.2 The learning analytics framework

To analyze the field of LA in general, Drachsler and Greller [3] created a learning analytics framework with six critical dimensions. All of these dimensions need to be covered to ensure an appropriate exploitation of LA in an educationally beneficial way. The six dimensions are: stakeholder, internal limitations, external constraints, instruments, data and objectives, but for LADs we focus mostly on stakeholders and objectives. Similar to the other dimensions, these two dimensions are subdivided in sub-dimensions or instantiations. For stakeholders, these sub-dimensions are: institutions, learners, teachers and others. The dimension objective is divided only in two sub-dimensions: reflective and predictive.

3 SCALABLE LEARNING DASHBOARDS

In the LALA project, a European project to build local capacity to implement learning analytics in Latin-America, the initial idea was to transfer the existing learning dashboards implemented in several European HEIs to HEIs in Chile and Ecuador.

During this project, we learned that one of the problems of LA and learning dashboards in particular is that most dashboards are very useful in a small, scientific context, but that it is very difficult to scale these tools to reach a broader audience.

To address this scalability problem, we identified several problems you can encounter when implementing learning dashboards at scale: actionability, unrealistic expectations, privacy, data availability, different infrastructure and difference in context. In this paper, we focus on the difference in context.

As discussed before, most LADs are implemented and researched in a very specific context and have proven their value in that specific context. However, when implemented in a different context, even if that difference is small, this dashboard can turn out to be totally useless. In this regard, we think that it is not possible to see a LAD as an individual and independent entity, but that the LAD should be considered in a holistic view. Namely, not only the dashboard with the visualizations itself, but also the context in which this dashboard is implemented. As a consequence, we question the scalability and especially generalizability of static, one-size-fits-all dashboards.

Instead of these one-size-fits-all dashboards, we propose to adapt or create a LAD based on previous work in a similar context. To do so, we propose to first identify the context in which the dashboard will be deployed and the context of other, existing LADs. Once these contexts are identified, it is possible to compare LADs based on their context and to find dashboards that are deployed in a similar context. Once the similar LADs are identified, these LADs can be adopted to the own context or elements of these LADs can be used to create a new LAD. The different steps of this process are shown in Figure 2.

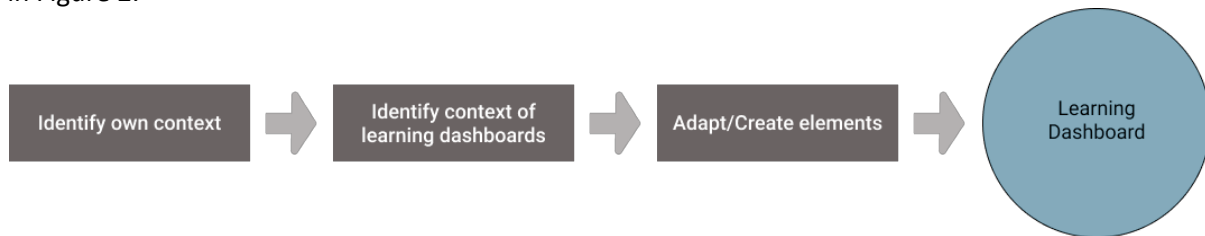


Figure 2: Steps to adapt/create a learning dashboard: 1) Identify the context of the new LAD. 2) Identify the context of existing LAD's to find similar LAD's 3) Adapt these existing LAD's to the new context.

4 IDENTIFICATION OF THE CONTEXT

4.1 First proposal

4.1.1 Theoretical background

In a first attempt to identify the context, we focused on identifying the objective and the stakeholders of the LAD. Both the objective and the stakeholders are part of the six dimensions of the learning analytics framework of Drachsler & Greller [2].

One of the key dimensions of this learning analytics framework is the objective of the learning analytics tool. In the framework, this dimension is divided into two sub-dimensions: reflection or prediction, but in the case of LADs, we propose to extend this dimension to not only defining what the objective of the LAD is, but also which underlying problem it is trying to solve. As designing and implementing a LAD requires quite some time and effort, it is interesting to analyze why the dashboard is needed and in which way it will solve this problem.

The second dimension we want to focus on are the stakeholders. In the learning analytics framework, this dimension is divided into four different stakeholders: learners, teachers, institutions and others. These four sub-dimensions are also sufficient to describe the different stakeholders of LADs, but a LAD is not limited to a single stakeholder, but it can also have multiple stakeholders [3].

4.1.2 Experiment

To test the identification of the own context, we held a workshop at the TIC.EC conference in Riobamba³, Ecuador. Participants were a mix of professors and educational technologists from both inside and outside the academic field. In total, 12 participants attended the workshop, including five professors, four people working in the educational technology (two in industry, two in a university), one person working in the administration of the university, one student and one data analyst.

During the workshop of one hour, we gave a short presentation about the learning dashboards at our university in which we explained the objective of the tool and the stakeholders that are involved. After this presentation, we asked the participants to list the problems of education in Ecuador that they would solve with a learning dashboard and to categorize them based on the stakeholders of this dashboard.

4.1.3 Results

In total, this workshop resulted in seven possible dashboards for learners, seven possible dashboards for teachers, six dashboards with the institution as stakeholders, and five with other stakeholders, as shown in Table 1.

Table 1: Results of the workshop

Learners	Teachers	Institution	Other
Decision support	Detect students at risk	Optimize physical space and resources	Opening up data for community
Tracing improvements	Monitor use of technology	Decision support based on failure rates	Ranking universities
Make learners more responsible	Time management	Making data more transparent	Regulatory compliance
Showing different courses	Detect need for own knowledge upgrade	Graphical instrument for research	Helping communities
Detect difficulties inside course	Detect difficulties inside course	Moving people across areas	Demographic information
Coaching/feedback from teachers	Optimize teaching strategy	Measuring results vs money	
Detect learning deficiencies	Cluster students for feedback		

As explained in the previous section, these are the results of the first step in the adaptation process we proposed in Section 3. In a second step, we propose to construct a similar table with existing dashboards that identifies the context in which these existing dashboards have been used. The third step then consists of selecting and adapting elements of a LAD deployed in a similar context as illustrated in Figure 3.

³ <https://ticec.cedia.edu.ec/es/>

4.1.4 Further steps

In this section, we illustrate the outlined adaptation process based on existing dashboards created by our research group. We took a context from Table 1 and mapped these to dashboards with a similar context at our university (LASSI, LISSA, REX, and POS).

We started with “detect learning deficiencies” as the problem to address for a learner. In the set of our dashboards, the REX dashboard [2] addresses the same problem for learners and seems to be a good to address the needs as identified by Latin-American researchers and practitioners. In addition, our LISSA dashboard [8] addresses decision support for learners as to whether which courses to retake or drop, and whether to continue with a study program. However, from our experience in the LALA project, we have learned that the Flemish (Dutch- speaking part of Belgium) context is different from Latin-American context. During a research visit in Chile, we tried the exercise to tailor our dashboards to address these needs to the institutional context of a University in Chile. However, we noticed that the identification of stakeholders and objectives is insufficient in the selection of dashboards. One of the issues we noticed it that other key contextual elements need to be captured to select suitable dashboards. At our institution, learners receive advice to support decision making at several key moments in the academic year – including a positioning test before the start of the academic year, first-semester exams, and second-semester exams. In Chile, these key moments, however, do not exist, and advice is grounded on very different progress rules. So even if the dashboards have the same objective and the same stakeholder, they cannot be easily adapted to this context. As such, we concluded that we need to modify our way of identifying the context to better reflect the reality. We discuss this context capturing and adaptation process in the next section.

4.2 New proposal

From our experience with transferring LADs to the context of HEI in Ecuador, we learned that even if a LAD has the same stakeholder and the same objective this still can be a different context. We experienced a need to include additional elements to take into account when talking about the context of a LAD. We propose to add two additional elements: interaction and key moments, as illustrated in Figure 3.

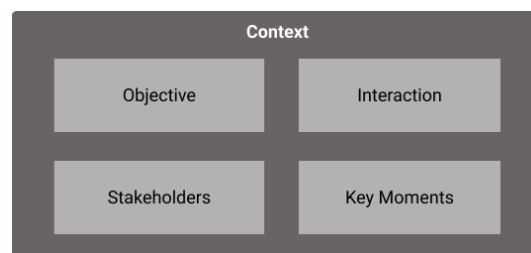


Figure 3: Context of a learning dashboard

The first additional element we propose to add to the context of a LAD is the **interaction** of the stakeholders with the LAD or interactions between the stakeholders. We propose this element

because we believe it is essential to not only take into account the stakeholders, but also the way these stakeholders will use the LAD. In the case the dashboard creates indirect interactions, the LAD functions as a catalysator for interactions between the stakeholders, for example the LISSA dashboard that works as a catalysator for the conversation between a student and a student-advisor and that also triggers reflections by the student [8].

A second element we propose to add are the **key moments** in which the stakeholders will be using the LAD. This element is crucial in the context of a LAD, as two dashboards with the same context, but with the intention to use at different key moments, will be completely different: we observed that the key moments (positioning test, first-semester and second-semester exams, etc.) we use in LISSA for instance do not exist in other contexts, and that the dashboard needs to be tailored to different key moments used at other institutions. More generally, the difference between dashboards that are used intensively during a program versus only once or twice needs to be encoded, as the use will be very different.

This new definition of identifying the context is not yet tested, but we plan to test this in future research.

5 CONCLUSIONS

Due to the increase of data in an educational context, the popularity of learning analytics and more specifically learning dashboard has also been increasing. Despite this increase of popularity, most learning dashboards are not scalable because they are mostly deployed in a small, scientific context. To include areas where learning dashboards are less common, there is a need for guidelines to adapt and create learning dashboard.

In this paper, we propose a guideline consisting of four different steps: (1) identifying the context in which the LAD will be used, (2) the context in which other dashboards are used, (3) the adaptation or creation of different elements based on similar LAD, and (4) the creation of a LAD.

In the second part of this paper, we focused on the first two steps where it is needed to identify the context in which the learning dashboard is deployed and the context of which existing dashboards are deployed. First, we proposed to analyze the context of a learning dashboard by identifying the objective and the stakeholders of the dashboard. To test this process of identifying the context, we held a workshop at a conference with 12 participants. From the results of this workshop, we learned during a research visit that identifying only those two elements is not enough to proceed with the process of analyzing, adapting or creating a learning dashboard.

To overcome this problem, we proposed a new definition of context with two additional elements: interactions and key moments. As this new definition is not yet tested, we plan to test this definition in future research. We hope that this new definition helps to identify the context of the learning dashboards which is an essential part of the guideline we proposed in the beginning of this paper.

6 ACKNOWLEDGMENT

Part of this work is funded by the LALA project (grant no. 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- [1] Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, 10(4), 405-418.
- [2] Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2017). Small data as a conversation starter for learning analytics: Exam results dashboard for first-year students in higher education. *Journal of Research in Innovative Teaching & Learning*, 10(2), 94-106.
- [3] Drachsler, H., & Greller, W. (2012, April). The pulse of learning analytics understandings and expectations from the stakeholders. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 120-129). ACM.
- [4] Duval, E. (2012). Learning Analytics and Educational Data Mining. *Online only*. Retrieved from: <https://erikduval.wordpress.com/2012/01/30/learning-analytics-and-educational-datamining>.
- [5] Millecamp, M., Gutiérrez, F., Charleer, S., Verbert, K., & De Laet, T. (2018, March). A qualitative evaluation of a learning dashboard to support advisor-student dialogues. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 56-60). ACM.
- [6] Ochoa, X. (2018). Learning analytics in Latin America present an opportunity not to be missed. *Nature Human Behaviour*, 1.
- [7] Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1), 30-41.
- [8] Charleer, S., Moere, A.V., Klerkx, J., Verbert, K. and De Laet, T., 2018. Learning analytics dashboards to support adviser-student dialogue. *IEEE Transactions on Learning Technologies*, 11(3), pp.389-399.

Expanding Adaptive Algorithms in New Ways: Echo-Adapt Software-As-A-Service

Bingnan Jiang, Michelle Barrett

ACT, Inc.

{bingnan.jiang, michelle.barrett }@act.org

ABSTRACT: High-quality assessment platforms rely on appropriate content and reliable statistical models to estimate the examinee's ability efficiently. The assessment content must represent the knowledge, skills, and abilities of interest. Statistical models need to accommodate differences between items, e.g., some are more difficult than others. Further, it is important to understand the probability of a response given both the examinee's ability and characteristics of the item. Practically, additional requirements include: 1) Low latency for user experience, 2) Interoperability with test drivers, 3) Simple ways to apply statistical models to new content (aka "field testing"), 4) Ways to control the exposure of content, and 5) Levels of adaptivity. A well-known approach within psychometrics is the shadow-test approach (van der Linden, 2005), which allows for the simultaneous management of many of these aspects. In this session, we will provide background on this approach and include a demonstration of RSCAT, an open-source R package solution available to researchers using the approach. We will also briefly demonstrate Echo-Adapt®, software-as-a-service built to deliver adaptive assessment at scale and a reference implementation for the IMS QTI standard for adaptive testing.

Keywords: Adaptive assessment; item response theory; shadow-test approach; interoperability standards

1 INTRODUCTION

Assessment for learning is an approach to identify students' learning needs and help teachers to plan learning programs. Accurate assessment of students' skills and knowledge closes the gap between their current situations and learning goals. Conventional large-scale testing is designed with items from a wide range of difficulty because it assumes a broad ability range of examinees. It is inefficient to assess skills and knowledge of examinees with high or low abilities. On one hand, highly proficient examinees will waste time on easy items that contain little information to distinguish them from less proficient examinees. On the other hand, less proficient examinees can be frustrated by answering difficult items (Wainer et al., 2001). Computerized adaptive testing (CAT) saves time and improves efficiency by administering the best items to measure the ability of an individual examinee. Since the principle of adaptive testing was first implemented in Binet's IQ test in 1905 (Binet & Simon, 1905), theories and technologies for CAT have been significantly developed to personalize testing with reduced testing time, improved accuracy, increased security, and reliable delivery. To date, well-designed adaptive algorithms have been shown to produce a reasonably stable estimate of an examinee's ability within about 10 items (van der Linden & Pashley, 2010). Most big assessment companies are actively conducting research on CAT or have their own adaptive testing engines. ETS has studied the effectiveness of item response theory (IRT) proficiency estimators under adaptive multistage testing (Kim, Moses, & Yoo, 2015). McGraw-Hill Education has launched a CAT for its Acuity assessment based on the shadow-test approach ("McGraw-Hill Education's Acuity Launches Adaptive Assessment Solution," 2015). A well-recognized approach for CAT is the shadow-test approach (van

der Linden, 2005). It assembles a complete test form at each adaptive stage based on the current estimate of examinee's ability. Besides scientific models and algorithms for CAT, some computational and implementation issues must be addressed before applying CAT to real-world, high-stakes assessments. This paper briefly discusses some of the key CAT algorithms and implementation technologies in Echo-Adapt, high-performance and reliable software-as-a-service for adaptive testing, and RSCAT, an open-source R package available for CAT research. Demonstrations on Echo-Adapt and RSCAT will also be provided at the workshop.

2 SHADOW-TEST APPROACH TO ADAPTIVE TESTING

A fundamental dilemma in adaptive testing is to administer optimal items sequentially and to meet content specifications simultaneously. First, administered items should be statistically optimal with respect to the examinee's ability estimation. Second, all content specifications, from the fixed-form predecessors, must be met throughout the testing. If optimal items are always administered early in a test, some content constraints may have to be violated in the middle or at the end of the test. The shadow-test approach (van der Linden, 2005) solves the dilemma by assembling a sequence of simultaneous fixed forms, each of which is dynamically updated based on the examinee's ability at a stage. Shadow-test CAT has many advantages, including full coverage of the test blueprint, separation of test specifications from CAT algorithm for easily modifiable configurations, and supporting flexible and reliable delivery options (e.g., linear on-the-fly, multi-stage, fully adaptive). The shadow-test approach can also be easily integrated with statistical models for item exposure rate control, field testing, test speediness, etc.

2.1 Shadow-Test Assembly

The shadow-test approach sequentially assembles test forms based on real-time updates of the examinee's ability estimate. Each test form assembly is modeled as a mixed integer programming (MIP) problem. A MIP model optimizes (either in the minimization or maximization sense) a function of variables (the objective) by selecting the best possible set of decisions (Smith & Taşkın, 2007). A standard shadow-test assembly MIP selects a subset of items from an item pool to maximize the test information:

$$\begin{aligned} &\text{Maximize} && \sum_{i_j \in S} I_{i_j}(\hat{\theta}) x_{i_j} \\ &\text{Subject to} && \text{content specification constraints} \end{aligned} \quad (1)$$

where S is the set of items in the item pool, $I_{i_j}(\hat{\theta})$ is the Fisher information of item i associated with passage j at the examinee's ability estimate $\hat{\theta}$, and x_{i_j} is the binary decision variable for the selection of item i_j in the shadow test. $x_{i_j} = 1$ if item i_j is selected in the shadow test, otherwise $x_{i_j} = 0$. Content specification constraints can be at either the item, passage, or test level, including but not limited to: 1) test length, 2) number of items and passages with specific attributes, 3) enemy items, and 4) passage positions in the test.

At each adaptive stage (after each real-time ability update), the shadow-test approach administers optimal items in two steps as shown in Figure 1. The first step is to construct the shadow test by solving the shadow-test assembly MIP. It selects a set of optimal items from the item pool based on the

examinee's ability estimate while conforming to all constraints. A shadow test consists of two parts, a set of items that have already been administered and a set of items that are unseen to the examinee. The second step is to administer the optimal item from the set of unseen items with rules including maximizing the item information and ensuring the correct passage order and item order in a passage. When a shadow test is assembled for the next adaptive stage, all previously administered items are constrained to be selected in the MIP model.

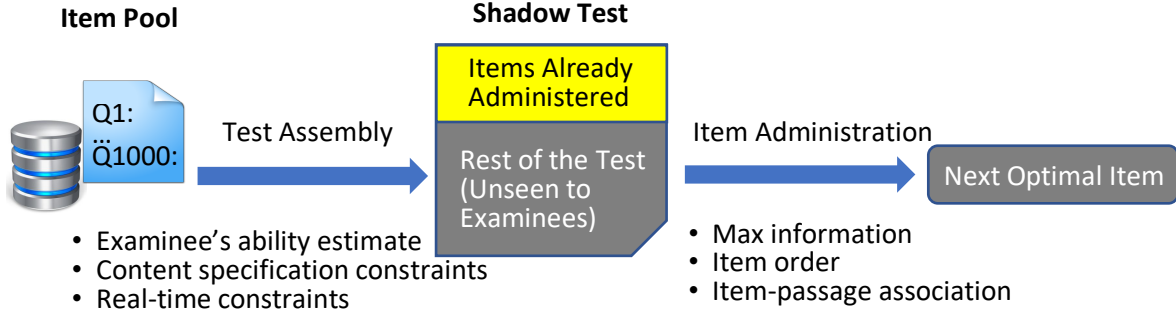


Figure 1: Item administration through the shadow-test approach

2.2 Exposure Rate Control

It is necessary to control the exposure rates of items/passages in adaptive testing for test security and to prevent over/under usage of items in the item pool. Existing exposure control methods can be easily integrated into the shadow-test approach, including the alpha stratification method (Chang & Linden, 2003), the Sympton-Hetter method (Sympton & Hetter, 1985), and the ineligible constraint method (Linden & Veldkamp, 2007). The main idea is to convert the exposure control rules or probability experiment results to constraints and/or objectives in the shadow-test MIP model.

2.2.1 Ineligible Constraint Method

The ineligible constraint method describes item/passage administration eligibilities in an $I \times K$ probability matrix, where I is the number of items/passages in the pool and K is the number of contiguous intervals across the theta continuum (from $-\infty$ to $+\infty$). An individual probability $\hat{P}^{(j+1)}(E_i|\theta_k)$ is calculated to decide if an item/passage i is eligible for administration to an examinee with ability in the theta range k :

$$\hat{P}^{(j+1)}(E_i|\theta_k) = \min \left\{ \frac{r_{\max} \varepsilon_{ijk}}{\alpha_{ijk}}, 1 \right\}, \text{ for } \alpha_{ijk} > 0 \quad (2)$$

where r_{\max} is an exposure goal rate, α_{ijk} is the number of examinees through examinee j who visited theta range k and took item/passage i , and ε_{ijk} is the number of examinees through examinee j who visited theta range k when item/passage i was eligible. The eligibility probabilities are then used to conduct $I \times K$ binomial experiments:

$$X_{ik} \sim B(1, p) \quad (3)$$

where $p = \hat{P}(E_i|\theta_k)$. If $X_{ik} = 0$ then item/passage i is ineligible at theta interval k ; otherwise the item/passage is eligible.

2.2.2 Big M Method

Theoretically, a hard constraint $x_i = 0$ can be added to the shadow-test MIP to avoid selecting an ineligible item i . Practically, however, this may cause an infeasibility issue (no solution) when shadow test is being assembled, especially when the item pool size is small and most items have already been frequently exposed. The big M method is proposed to address the infeasibility issue. For each ineligible item, a penalty term M is subtracted from the shadow-test MIP objective function:

$$\text{Maximize} \quad \sum_{i_j \in S} I_{i_j}(\hat{\theta})x_{i_j} - M \sum_{i_j \in V} x_{i_j} \quad (4)$$

where V is the set of ineligible items due to the exposure rate control. The penalty M is selected as a value greater than the maximum item information value of the items in the pool at the current ability estimate. In this way, the big M penalties serve as soft constraints to avoid selecting ineligible items if feasible shadow tests still exist after excluding them, because the selection of infeasible items will reduce the MIP objective value that is to be maximized. But ineligible items are still allowed for selection in some scenarios to prevent infeasibility and test interruption.

2.3 Optimal Field Testing

An adaptive testing program selects optimal items from an operational item pool to maximize the efficiency of estimating examinee's ability. It is common that some items in an item pool need calibration with response data from examinees in (oftentimes separate) field testing. The shadow-test approach enables optimal field testing to be embedded in adaptive testing. Advantages of embedded field-testing include the adaptive selection of field-test items based on real-time updates of the examinees' ability, consistent motivations of operational testing and field testing, and cost savings when separate calibration studies are not necessary.

2.3.1 Optimal Design Criteria

The goal of adaptive testing and field testing is to maximize the information about the latent attributes, i.e., ability of examinees and response model parameters of field-test items. Optimal designs are usually used to achieve this goal with respect to some statistical criteria. Among them, a favorable criterion is the Bayesian D-optimality (Holling & Schwabe, 2018), the choice of a minimum determinant of the covariance matrix of the estimators of the intentional parameters given the nuisance parameters. Minimizing the determinant of the covariance matrix for maximum-likelihood estimators is asymptotically equivalent to maximizing their Fisher information matrix. To fit the continuous process of item selection in testing, the D-Optimal criterion is designed to select item with the maximum marginal profit for the determinant; that is, the contribution by a test taker across all field-test items to the determinant of the information matrix relative to its current value. Its posterior expected value is calculated as:

$$D_f \equiv S^{-1} \sum_{s=1}^S \left[\det \left(C^{-1}(\boldsymbol{\eta}_f) + I(\boldsymbol{\eta}_f^s; \theta^s) \right) - \det \left(C^{-1}(\boldsymbol{\eta}_f) \right) \right] \quad (5)$$

where $C^{-1}(\boldsymbol{\eta}_f)$ is the inverse of the covariance matrix of field-test item parameters calculated from the last posterior draws for the field-test item $\boldsymbol{\eta}_f$. $I(\boldsymbol{\eta}_f^s; \theta^s)$ is the expected item information matrix

at the current field-test item parameter sample η_f^S and the examinee's ability sample θ^S . S is the sample size.

2.3.2 Shadow-Test Approach to Optimal Field Testing

For optimal field testing embedded in the shadow-test approach to adaptive testing, new constraints are added in the shadow-test assembly MIP to ensure the required numbers of operational and field-test items are selected in a shadow test. Additional auxiliary constraints are added to manage field-test item positions in the test and resolve their conflicts with passage delivery. The MIP objective function is also adjusted to include the optimal design criteria for field-test item selection. If content specifications are specified separately for operational items and field test items, the MIP objective function can be the sum of Fisher information (of selected operational items) and D-optimality criterion values (of selected field-test items). Otherwise, the MIP objective function switches between maximizing the sum of Fisher information and maximizing the sum of D-optimality criterion at an operational stage and a field-test stage, respectively.

3 INTEROPERABILITY AND PERFORMANCE

Interoperability and performance are critical implementation issues that need to be addressed in any assessment platforms. Establishing good interoperability between adaptive testing engines and test delivery platforms eliminates the requirement for costly proprietary integrations. An examinee also typically cannot wait more than 1-2 seconds without negative impact, so delivery latency is an important non-functional requirement.

3.1 IMS Global QTI and CAT APIs

Decoupling adaptive testing engines and algorithms from the platform which delivers items to examinees has advantages. For example, in a recent demonstration of interoperable adaptive testing engines and test delivery platforms from three different organizations (Aarnink, Barrett, & Molenaar, 2018), the same adaptive engine was used to deliver items in two different test delivery platforms, one for formative assessment and the other within a game for pre-school age children learning phonics. In addition, the same test delivery platform was demonstrated to use two different adaptive engines, one for formative assessment and one for high-stakes large scale assessment. Interestingly, this allows rigorous computer adaptive testing to be used in a number of learning and measurement applications.

To ensure a high degree of interoperability with test delivery platforms, adaptive testing engines may be designed and built to conform to the IMS Global Question & Test Interoperability (QTI) standards. The IMS Global Learning Consortium is a non-profit collaborative with a goal to “enable a plug and play architecture and ecosystem that provides a foundation on which innovative products can be rapidly deployed and work together seamlessly” (IMS Global, 2018). IMS Global includes interoperability standards for integrated assessment (QTI, APIP, CAT), learning data and analytics (OneRoster, Caliper Analytics), learning platforms, apps, and tools (Learning Tools Interoperability), digital curriculum (Common Cartridge), and digital credentials and pathways (CASE, Comprehensive Learner Record, Open Badges).

QTI standards for CAT define a format for the exchange of which test question(s) to deliver, scoring information for individual questions, the examinee's interim and final ability estimates, precision of the ability estimates, etc. Multiple options exist for architecture of the test delivery platform and the adaptive testing engine; an adaptive engine delivered as software-as-a-service is typically implemented as a service (not a library) and usually accessed by the test delivery platform using the HTTPS protocol. QTI compliant CAT engine APIs define actions including creating test session, verifying items, submitting results, ending a test for an individual examinee, and ending test session. Their implementations follow the OpenAPI specification (formally known as Swagger Specification). An example of HTTP request from a test delivery platform to submit results for the first item is shown in Figure 2:

POST **BASE_URL**/sections/**632**/sessions/**04a87995-1090-433d-9bbb-a5f9378b3dee**/results

```

1. {
2.   "assessmentResult":{
3.     "itemResult":[
4.       {
5.         "identifier":"I105_32418", ← Item Administered
6.         "outcomeVariables":[
7.           {
8.             "identifier":"SCORE",
9.             "value":[
10.              {
11.                "baseType":"integer",
12.                "value":1 ← Item Score
13.              }
14.            ]
15.          }
16.        ]
17.      }
18.    ]
19.  },
20.  "sessionState":"eyJhZGFwdG12ZXN0Ywd1IjoxLCJwcmV2aW91c2x5U2V1bk10ZW1zIjpbXSwibmV4dE10
    ZW1zIjpbIkkxMDVfMzI0MTgiLCJMTA1XzMyNDk0Iiwic2VhbnV8YnJlLCJzaGFkb3dUZXN0IjpbIkkxMDVf
    fMzI0MTgiLCJMTA1XzMyNDk0Iiwic2VhbnV8YnJlLCJpdGVtU2NvcnVzIjpbXSwiZWNUaGV0YVVBvaW50cy
    I6bnVsbCwiZW50bGlnaWJpbG10eUluZGljYXRvcnMiOm51bGwsImVjVHlwZSI6MSwiYmVzdE51bWJlciI6M
    iwidGVzdFRha2VyU2Vzc2lvcmlkIjoieMDR0dC50TUtMTA5MC00MzNkLTliYmItYTVMOTM3OGIzZGV1In0="
21. }

```

Figure 2: Request from a test delivery platform to submit results for the first item

IMS Global member organizations can retrieve detailed specifications and reference implementations from the IMS Global website (IMS Global, 2018).

3.2 Performance Optimization

A high-quality CAT engine must provide consistently low latency for item delivery in all applications of as it relies on real-time information from an examinee. Thus, two issues need to be solved: 1) optimizing CAT algorithm runtime performance; and 2) scaling for test delivery with concurrent examinees.

Because MIP is NP-hard, the shadow-test assembly usually consumes the majority of computation time in a CAT cycle, especially when the item pool size is large and the test blueprint is heavily constrained. However, the MIP solving time can be significantly reduced, without sacrificing the solution quality, if appropriate approaches and techniques are applied. First, the fundamental approach is to use a commercial MIP solver, e.g., FICO Xpress and IBM CPLEX, which typically solves

MIPs 10~20 times faster than open-source MIP solvers. The runtime performance of different solvers on benchmark optimization problems is shown in Figure 3 (Meindl & Templ, 2012).

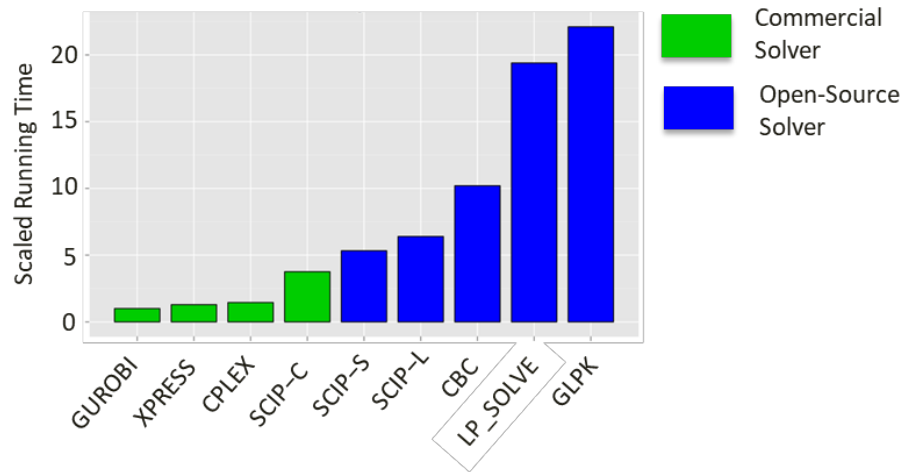


Figure 3: Scaled running time of several solvers applied to benchmark test data

Second, the shadow-test MIP formulation can be preprocessed using techniques such as applying tighter bounds, scaling the coefficient matrix, and fixing variables. For example, adding a constraint to restrict the number of passages to tight upper and lower bounds can greatly reduce the MIP solving time for a complex CAT configuration based on a large item pool, although the constraint may not be explicitly required by the test blueprint.

Last but not least, a warm start technique can be used to preload the shadow test solutions from the previous adaptive stage, since shadow tests at two adjacent stages are usually very similar. Thus, the MIP algorithm, e.g., the branch and bound algorithm, speeds up searching for the optimal solution by starting with a solution of high quality. Figure 4 shows the shadow test assembly performance improvement from applying the warm start technique to a CAT configuration based on an item pool of 720 items. The primary vertical axis represents the MIP solving time taken at one adaptive stage while the secondary vertical axis represents the MIP solving time reduction percentage after applying the warm start technique. The horizontal axis represents the ability estimate change from the previous adaptive stage to the measured stage. The warm start technique can reduce the solver time by as much as 77%. As the theta change increases, the solving time reduction decreases. The exact improvement that can be achieved also depends on the CAT configuration on a case-by-case basis.

For large-scale assessment, it is not cost effective to give each examinee their own CAT engines. Existing cloud computing algorithms and techniques, e.g., load balancing and auto-scaling, can efficiently forward concurrent test service requests to optimally sized CAT engines to meet the latency requirement at different scales of assessment.

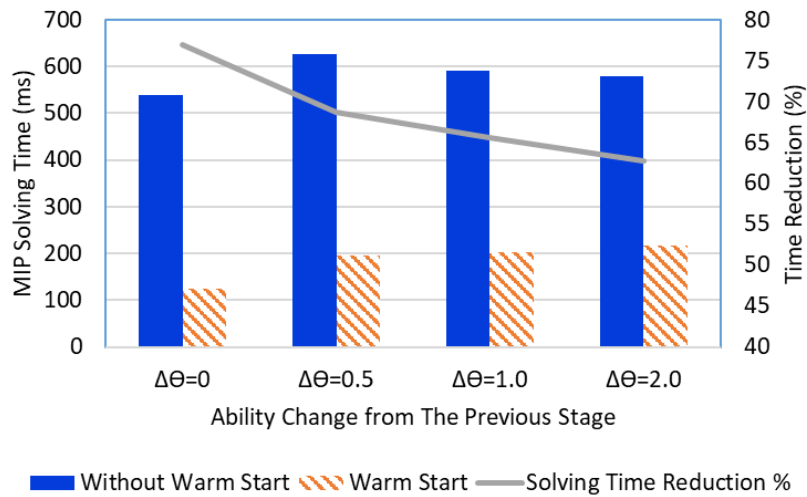


Figure 4: Shadow test assembly solving time improvement from using the warm start technique

4 CAT APPLICATIONS

4.1 Echo-Adapt

Echo-Adapt (version 1.34) is software-as-a-service built to deliver adaptive assessment at scale and provide personalized testing to minimize testing time while maximizing certainty about what an examinee knows. The Echo-Adapt high-level design diagram is shown in Figure 5. Echo-Adapt is based on the 3-parameter logistic item response theory (3PL IRT) model and the shadow-test approach, where the expected a posteriori (EAP) method and the Markov chain Monte Carlo (MCMC) method are available to score the examinees' ability. The optimal field testing functionality in Echo-Adapt allows users to embed field testing in adaptive testing, i.e., delivering field-test items at specific or random stages and calibrating their parameters with response data. The Echo-Adapt CAT engine is implemented in Java while the shadow-test assembly MIP is modeled in the Mosel scripting language. Echo-Adapt is deployed on Amazon Web Services (AWS) and deeply optimized to achieve required performance at large-scale assessment, i.e., less than 500ms latency for an item administration given 40,000 concurrent examinees. It uses a commercial MIP solver to solve the shadow-test assembly MIP in a real-time and reliable manner. Echo-Adapt conforms to the IMS Global QTI standard so that it can interoperate with other QTI compliant platforms and has been integrated with two different test delivery platforms to date. To enhance user experience and save time on configuring a CAT, Echo-Adapt also provides a web user interface for the intuitive configuration of content specification constraints, algorithms, and simulations.

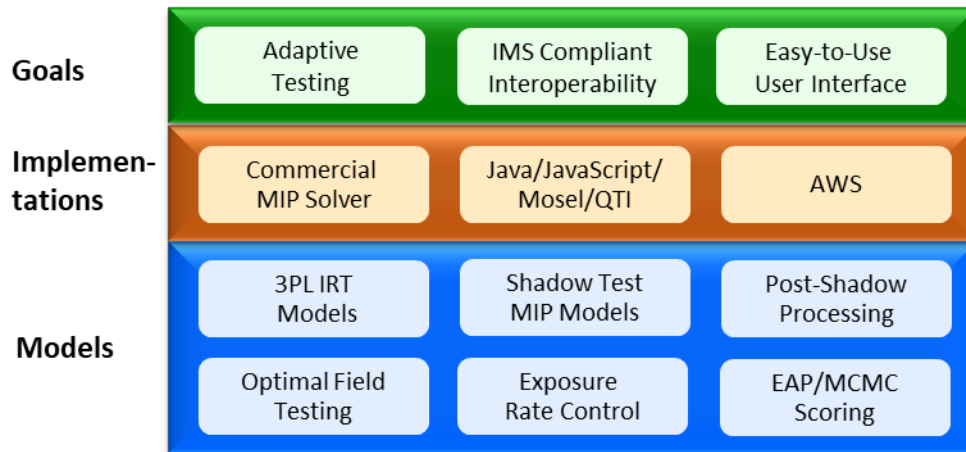


Figure 5: Echo-Adapt high-level design diagram

4.2 RSCAT

An R package for shadow-test approach to computerized adaptive testing (RSCAT) is to be released as an open-source package that is free to use for CAT research. RSCAT is different from other CAT packages in three aspects. First, the CAT engine and algorithms in RSCAT are implemented in Java for high efficiency and good runtime performance. They are then encapsulated in R APIs that can be called by other R programs. Second, a Shiny user interface is implemented to assist users with CAT and simulation configurations. Third, RSCAT is compatible with most existing MIP solvers. It allows users to bring their own commercial MIP solvers, e.g., CPLEX and Xpress, to maximize the efficiency of solving shadow-test assembly MIPs in a large-scale simulation or simply use open-source solvers for quick research validation. As a lightweight version of Echo-Adapt, RSCAT has some limitations with respect to scoring methods, content constraint types, and scaling. Currently, EAP is the only available scoring method in RSCAT. Some complex constraints like the passage position constraint are also not supported. Since RSCAT runs on a user's local machine, it is not scaled for large-scale simulations. In addition, the field testing functionality is unavailable in the package. Despite these limitations, users can still take benefits of the shadow-test approach by using RSCAT in their research.

5 CONCLUSION

The shadow-test approach has been recognized as an elegant solution to implement adaptive testing, where content specification constraints are strictly met. It also effectively integrates exposure rate control, field-test item calibration, and other statistical models for assessment. Some implementation and computational issues within CAT platforms are discussed, with solutions to enhance the interoperability and optimize the run-time performance. As an application of the shadow test approach, Echo-Adapt is built as software-as-a-service with good interoperability and desirable performance. RSCAT is also to be released as an open-source package for CAT research using shadow testing.

REFERENCES

Aarnink, A., Barrett, M.D., & Molenaar, M. (2018, September). CAT in a box: Implementing a standard on computer adaptive testing. Paper presentation at e-ATP, Athens, Greece.

- Binet, A., & Simon, T. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191–244.
- Chang, H.-H., & Linden, W. J. van der. (2003). Optimal Stratification of Item Pools in α -Stratified Computerized Adaptive Testing. *Applied Psychological Measurement*, 27(4), 262–274. <https://doi.org/10.1177/0146621603027004002>
- Echo-Adapt (Version 1.34). [Computer Software]. Iowa City, IA: ACT.
- Holling, H., & Schwabe, R. (2018). Statistical optimal design theory. In *Handbook of item response theory Volume 2: Statistical tools*. Boca Raton, FL: Chapman & Hall/CRC.
- Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of Item Response Theory (IRT) Proficiency Estimation Methods Under Adaptive Multistage Testing. *ETS Research Report Series*, 2015(1), 1–19. <https://doi.org/10.1002/ets2.12057>
- Linden, W. J. van der, & Veldkamp, B. P. (2007). Conditional Item-Exposure Control in Adaptive Testing Using Item-Ineligibility Probabilities. *Journal of Educational and Behavioral Statistics*, 32(4), 398–418. <https://doi.org/10.3102/1076998606298044>
- McGraw-Hill Education's Acuity Launches Adaptive Assessment Solution. (2015, April 28). Retrieved December 6, 2018, from <https://www.mheducation.com/news-media/press-releases/mcgraw-hill-educations-acuity-launches-adaptive-assessment-solution.html>
- Meindl, B., & Templ, M. (2012). Analysis of commercial and free and open source solvers for linear optimization problems.
- Smith, J. C., & Taşkın, Z. C. (2007). A Tutorial Guide to Mixed-Integer Programming Models and Solution Techniques. Retrieved December 6, 2018 from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.464.6182&rep=rep1&type=pdf>
- Simpson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego CA: Navy Personnel Research and Development Center.
- van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York, NY: Springer.
- van der Linden, W. J., & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3–30). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-85461-8_1
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2001). *Computerized adaptive testing: A primer* (second edition). New York, NY: Routledge.

Predicting student knowledge at scale at Duolingo

Klinton Bicknell, Burr Settles

Duolingo

{klinton,burr}@duolingo.com

ABSTRACT: One of the central promises of computerized instruction is the personalization of learning. A prerequisite to fully realizing this promise is making high quality inferences about a learner’s current knowledge state, in order to optimize what material is most useful to present next. Here, we describe recent progress in solving this problem at scale at Duolingo, a language learning platform with over 300 million users worldwide spread across over 80 separate language courses. We focus on two instantiations of this general problem: (1) how to rapidly assess the overall knowledge level of a new user who is starting Duolingo already having substantial experience with the language and (2) how to estimate the details of what an existing Duolingo user knows from their history of interactions with the platform. In both of these case studies, solving the problem at scale adds additional complexity and limitations on possible solutions. However, working at scale also enables the possibility of leveraging a large amount of learning data from other users to improve inferences. We present techniques taking advantage of this data by combining machine learning with classical models from psychometrics and cognitive science, to yield state-of-the-art inferences about user knowledge state.

Keywords: machine learning, cognitive science, computerized adaptive testing (CAT), item response theory (IRT), memory

1 INTRODUCTION

Achieving effective personalization of instruction requires obtaining high quality predictions about a learner’s current state of knowledge. This general problem of predicting learner knowledge commonly occurs in two particular situations. The first of these situations is when a learner is new to a learning platform but already has substantial prior experience with the material being taught. Here, prior knowledge needs to be efficiently and rapidly estimated for the purpose of determining where in a course the learner should start. The second common situation is when a learner has been using a learning platform for some time, such that there is substantial data from interactions with this learner with the platform’s material to use in determining what they do and do not know. Here, we present techniques we have used to effectively solve each of these problems at scale in Duolingo, a language learning app with over 300 million users, by adapting classical techniques from cognitive science and psychometrics to large datasets with machine learning. We start by discussing the problem of estimating new user knowledge and then move on to discuss the problem of estimating the knowledge of existing users.

2 NEW USER KNOWLEDGE MODELING

Each language course in Duolingo is structured as a sequence of *skills* to be learned. These skills are then grouped into ordered *rows* in the app, such that learners must complete all the skills in one row before they have access to any of the skills in the next row. When a learner is new to a language

course, they are asked whether they have any prior knowledge of the language taught in that course. Learners who have no prior knowledge will start the course at the beginning, i.e., no skills will be initialized as completed. For learners who say they do have prior knowledge of the language, it is the role of new user knowledge modeling to determine where in the course to place them, i.e., to determine how many rows of the course will be shown as already completed.

Most of the users who are new to a course are also new to the Duolingo app, and so are eager to start learning new material to evaluate the app. For this reason, it is impractical to administer a long placement test, which would require a substantial upfront time investment from new users before they even know if they would like to use Duolingo going forward. Instead, this situation requires estimating new user knowledge very efficiently by administering a very short placement test. We achieve this goal of a very short efficient placement test using a computerized adaptive testing (CAT) framework (for an introduction, see Segall, 2005) backed by a generalization of item response theory (IRT) models of user knowledge (for an introduction, see Embretson & Reise, 2013).

Complicating this situation are effects of exercise type. That is, skills in each language course are defined in terms of material to be taught (words, grammar, etc.), but there are different ways that a language learner could interact with this material, each of which has its own associated difficulty. To make this concrete, consider a speaker of English learning Spanish. One of the things an early skill might teach them is that *gato* is the Spanish word for *cat*. They may be asked to translate a sentence in Spanish with *gato* back into English (a relatively easy task), or to generate the Spanish translation *gato* from the English *cat* (a harder task), or to recognize a sentence in Spanish containing *gato* that they hear auditorily (also hard). Thus, a learner who can reliably recognize that *gato* means *cat* when they see it won't necessarily be able to generate the form *gato* from *cat*. However, if a learner can generate *gato* from *cat*, they can probably also recognize what *gato* means (i.e., *cat*) when they see it. Because exercise type combines with the depth (row number) of a skill in the course to determine difficulty, estimating user knowledge effectively requires building models that incorporate both of these sources of information.

We solve this problem by using a parametric generalization of classical item response theory, in which the item response theory parameters (difficulty, discrimination, etc.) are additive functions of both depth and exercise type. We efficiently fit these parameters to very large datasets, marginalizing over the skill levels of different learners, by using variational Bayesian inference. Once fit, we use the parameters inside an adaptive computerized adaptive testing (CAT) framework, in which we select exercises that will maximize information about the learner's placement, given the answers they have provided so far in the test. Implementing this test in Duolingo yielded significant improvements in user retention metrics.

3 EXISTING USER KNOWLEDGE MODELING

The second setting for knowledge modeling at Duolingo is for existing users. For these learners, we already have substantial data about exercises they have completed correctly and incorrectly, the mistakes they have made and when they made them. The goal of existing user knowledge modeling is to use all of this data to predict how likely a given learner would be to respond accurately to a given exercise on a particular topic.

If each learner’s knowledge were static, this problem would be relatively easy. However, this problem is difficult because a learner’s knowledge and skills change over time. In principle, there are two ways in which knowledge will change over time: a learner could learn something (going from a state of not knowing to a state of knowing) or a learner could forget something (going from a state of knowing to a state of not knowing). In many applications of knowledge modeling, such as Bayesian Knowledge Tracing (Corbett & Anderson, 1994), the focus is on learning rather than forgetting. However, Duolingo is structured so that in the lesson where a new knowledge element (e.g., a new word) is first presented, it is presented multiple times and with easy enough exercise types that most learners will get these exercises correct by the end of the lesson, i.e., will have learned it. The main challenge, then, is predicting forgetting.

There is a long history in cognitive science of modeling forgetting, at least since Ebbinghaus (1885). In all this work, the probability that an item in memory will be able to be correctly recalled decays over time, but different work adopts different functional forms. Much work has used exponential decay (e.g., Ebbinghaus, 1885; Pimsleur, 1967; Leitner, 1972) and power law decay (e.g., Wixted & Ebbesen, 1997; Cepeda et al., 2006). Here, we use exponential decay. That is, *immediately* after learning an item (or after being reminded of a previously forgotten item), a learner will almost certainly be able to recall it ($p \approx 1$), but after a delay of that item’s half-life h , the probability will be 0.5, after $2h$, the probability will be 0.25, etc. Under this model of forgetting, then, knowledge modeling reduces to modeling half-lives: the goal is to predict the half-life of a given item after a given presentation to a given learner.

These half-lives are functions of a range of factors. Perhaps most obviously, an item’s half-life in a given instance will depend on how many times a learner has seen that item previously, and how many of those times they responded to it accurately versus inaccurately. There are also large differences between items in terms of half-life, holding exposure constant. For example, the Spanish translation of the English word *bar* is the Spanish word *bar*, spelled exactly the same but pronounced a bit differently. An English speaker learning Spanish might never forget this word in Spanish even after only seeing it once. By contrast, after first seeing that the Spanish translation of the English word *pregnant* is *embarazada*, a learner might be very likely to forget that (and maybe think that *embarazada* means *embarrassed*) by the next day. Similarly, words in many languages can appear in multiple grammatically inflected forms, which may also differ from each other in terms of half-life. For example, it may be relatively hard to forget that *es* is the Spanish form for the singular third-person present tense of *to be* since it is so similar to the English equivalent *is*; but it may be much easier to forget the past tense version of that same verb, which is *fue* in Spanish, since that is more different than the English equivalent *was*.

To capture all these effects of exposure history, item effects, and effects of item properties like inflection, all within a single model of half-lives, we developed a novel machine learning technique called half-life regression (Settles & Meeder, 2016). This model generalizes many prior models of forgetting in cognitive science, while also allowing for effects of arbitrary item properties like words and inflected forms. At the same time, the model is designed so that its parameters can be efficiently estimated from our large dataset of learner interactions using stochastic gradient ascent. We show that this model reduces error in predicting which exercises users will answer correctly and

which incorrectly by nearly half. Additionally, using this model to prioritize material for practice in the Duolingo app significantly improved user retention.

4 CONCLUSION

In this paper, we described two methods for estimating student knowledge at scale, in the context of the Duolingo app. The first one generalizes the classical psychometric technique of item response theory with parametric probabilistic models estimated with variational Bayesian inference. The second one generalizes classical cognitive science models of forgetting by incorporating parametric effects of item properties estimated with stochastic gradient descent. Both of these methods share the common theme of augmenting classical approaches with modern machine learning to leverage large datasets and improve predictions. We expect this type of approach to continue to produce substantial improvements in knowledge modeling over the coming years.

REFERENCES

- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354.
- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4), 253-278.
- Ebbinghaus, H. (1885). Über das Gedchtnis. Untersuchungen zur experimentellen Psychologie. Leipzig: Duncker & Humblot. Translated into English as Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology*. New York: Teachers College, Columbia University.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Leitner, S. (1972). *So lernt man lernen. Angewandte Lernpsychologie – ein Weg zum Erfolg*. Verlag Herder, Freiburg im Breisgau, Germany.
- Pimsleur, P. (1967). A memory schedule. *Modern Language Journal* 51, 73–75.
- Segall, D. O. (2005). Computerized adaptive testing. *Encyclopedia of Social Measurement*, 1, 429–438.
- Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1848–1858.
- Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, 25(5), 731-739.

TutorGen SCALE® - Student Centered Adaptive Learning Engine

Ted Carmichael, PhD

TutorGen, Inc.
ted@tutorgen.com

Mary Jean Blink

TutorGen, Inc.
mjblink@tutorgen.com

John Stamper, PhD

Carnegie Mellon University, TutorGen, Inc.
john@stamper.org

ABSTRACT: TutorGen’s Student Centered Adaptive Learning Engine (SCALE®) represents a breakthrough in developing adaptive educational systems by using big data collected from new or existing educational software systems to automatically generate intelligent tutoring capabilities. This work aims to make adaptive learning widely available; to give students real-time, useful feedback; and to provide tools to teachers for assessing student performance. SCALE does this by collecting student data from new or existing digital learning systems and then automatically generating adaptive capabilities based on this data. In this way SCALE is able to efficiently turn any edtech product into an intelligent tutoring system, with very little need for new software customization or expensive and time-intensive manual input. Adaptive learning has long been proven to decrease time to learn and increase retention and understanding for students, but has not been widely adopted due to the high cost of implementation. SCALE solves this challenge with a human-centered, data-driven approach by using Artificial Intelligence and machine learning techniques to generate adaptability in a way that is content and system agnostic. Here we report on our approach for creating and implementing SCALE, and the refinements created to bring this technology from the research lab into the classroom.

Keywords: ITS, Intelligent Tutoring System, EDM, Educational Data Mining, Artificial Intelligence, Big Data, Adaptive Learning, Teacher dashboard, edtech.

1 INTRODUCTION

TutorGen’s Student Centered Adaptive Learning Engine (SCALE®) represents a breakthrough in developing adaptive educational systems by using big data collected from new or existing educational software systems to automatically generate intelligent tutoring capabilities. SCALE collects data from existing computer/web based training software, and uses educational data mining and artificial intelligence techniques to automatically generate student models. SCALE improves these models over time as more data is collected, and tracks student progress on specific concepts or skills (knowledge tracing). This allows for easy assessment at any point in time. The system also dynamically selects the next best problem to maximize student learning and minimize time needed to master a set of skills (problem selection). For complex multi-step problems, SCALE can provide context specific, just-in-time hints. SCALE also provides data adapters so edtech developers can

easily hook into the SCALE system using Web APIs. Finally, a main differentiator of our system is our transparent process of data curation and the related visual tools that expose the workings of the problem and student-model generation process. **By building on award-winning learning science and data mining research, we have designed and developed a system that makes adding adaptive capabilities to existing systems easy and affordable.**

SCALE grew out of a recognized need in the marketplace, for automatic adaptability that can be added to any edtech product, without having to custom design and implement a non-generalizable solution. We found that many companies are designing new edtech delivery and content products, and new platforms, but that adaptability is often missing, or only done in a cursory way. Yet methods for using Artificial Intelligence (AI) and machine learning to generate adaptability in a way that is content and system agnostic have already been proven in university research labs, including our own work [e.g., 1-4]. This work has been supported by multiple grants from the NSF and the Commonwealth of Kentucky, to get the SCALE technology out of the lab and into the hands of students, teachers, administrators, and developers of edtech products.

2 BUILDING THE FOUNDATION FOR SCALE

Following extensive discussions with potential customers, we identified the features necessary to make the TutorGen SCALE system a success. The key innovation of SCALE is the automatic generation of adaptive learning capabilities. These include: 1) Problem selection 2) Skill tracking (and skill modeling) 3) Hints and feedback on multi-step problems 4) Assessment of student learning.

In addition to the ability to automatically generate these features, SCALE also includes a “feedback loop” to continue the improvement of the features over time as more data is collected. Several potential customers have noted that existing providers of adaptive student learning software do not have any way to explain why the system behaves as it does. In contrast, SCALE provides tools that let the instructors and developers explore the data using meaningful visualizations that will provide insights into student learning. In order to achieve this vision we identified the following technical objectives to be completed during the NSF SBIR Phase I and Phase II projects for SCALE:

1. Research and implement overall data and system architecture
2. Research, design, and implement knowledge tracing algorithms and handler
3. Research, design, and implement the hint and feedback mechanism
4. Design and implement student assessment and support tools for teachers
5. Design and implement integration of support tools for computer-based training systems and content designers/developers
6. Pilot test SCALE in the classroom
7. Perform Full Integration and Load Testing

2.1 Data and System Architecture

The high level design of the system can be seen in Figure 1, and the main product is designed around the SCALE Engine, which implements the model discovery and creates the mechanisms for automated and expert feedback. All information collected and generated is stored in a database, which can be accessed by developers and educators through a set of tools. These tools allow for exploring models, improving models, and assessing student performance. A set of universal data

connectors provides a simple API for connecting software to the system for data collection through web services. This documented communication method provides an open architecture for existing computer-based training system providers to connect to SCALE in a seamless fashion.

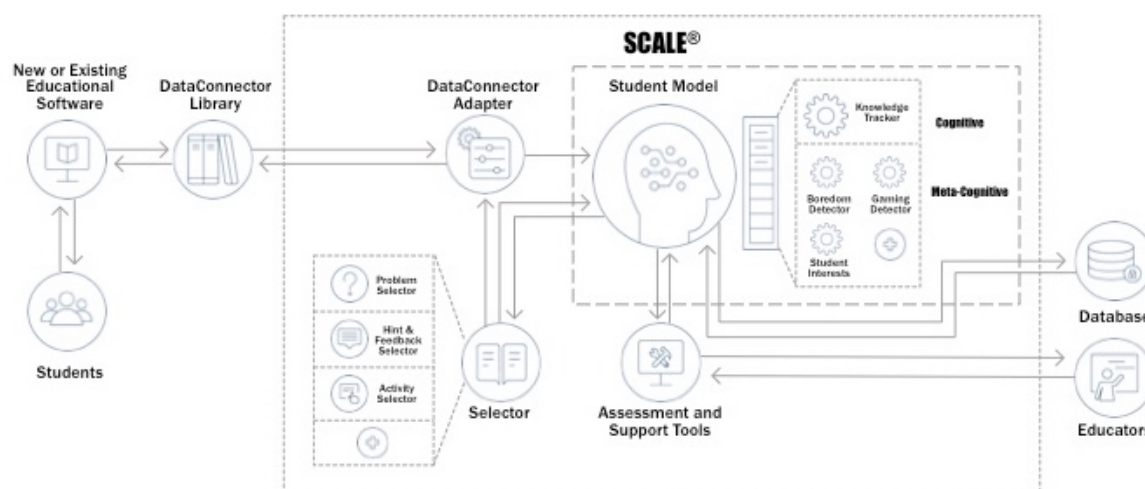


Fig 1. Design of TutorGen SCALE showing how new or existing educational software can use our data connector libraries to connect to the SCALE system which includes three main components: the knowledge tracker with student models, the hint and feedback handler, and assessment and support tools.

The database has been designed to be flexible enough to manage data for students, knowledge tracing, hint generation, student models, custom attributes, and transaction log. These capabilities require data to be managed at a very granular level while still providing exceptional performance for real-time processing of tutor data and analytics. This led to the design of a multidimensional data model, allowing for the highest performance. The design of the knowledge storage has extended the common format used in the PSLC DataShop repository (<http://pslcdatashop.org>) and centers on storing student knowledge in the form of Knowledge Component (KC) models. In these models, a KC represents a piece of trackable knowledge: skills or concepts. There is also a hierarchical component to expert models that is captured, which can be influenced by both the type of knowledge being taught and instruction used. The database design takes into account these hierarchical relationships, and also has the ability to compare models against various metrics.

2.2 Knowledge Tracing Algorithms and Handler

Student models have traditionally been developed by domain experts applying manual analysis of course content. Further refinements of these models have utilized Cognitive Task Analysis (CTA). A number of studies have demonstrated how detailed CTA can result in dramatically better instruction [5,6]. CTA methods are useful for creating student models, but they have several limitations. First, CTA is more of an art than a science. Structured interviews, think aloud protocols, and rational analysis are all highly subjective and different analysts may produce very different results. Second, the most successful CTA approaches are heavy in human effort. Structured interviews, think-alouds, or student model simulation all require high level of psychological and subject-area expertise in addition to significant time investments. This is both time-consuming and costly.

Our research into finding expert models is founded on our belief that the expert models are discovered, not created. This means that the correct model exists for a specific domain and set of instruction, and it must be identified in the search space of all possible models. A correct expert model is one that is consistent with actual student behavior. It predicts task difficulty, as well as transfer between instruction and test. Finding the exact model for a set of instruction is difficult because some students may be relying on a flawed model that they have created internally [3]. With this in mind, our system is able to discover the expert model that best describes the data and leads to the most robust learning.

To discover better expert models, we apply human and machine learning techniques and evaluate the resulting model using statistical analyses. This forms the basis of our “data-driven, human-centered” approach. Our core statistical model is based on the Additive Factors Model (AFM), which has been previously implemented in DataShop. AFM is a generalization of the log-linear test model (LLTM) [7]. It is a specific instance of logistic regression, with student-success (0 or 1) as the dependent variable and with independent variable terms for the student, the KC, and the KC-by-opportunity interaction. By clustering student knowledge areas, we have augmented the traditional AFM model, adding additional parameters that weight learning rates for these clusters, which, as more data is collected, makes an impact into optimizing an individual’s learning.

We have confirmed that our new technique for model discovery will implement the best features of the studied techniques and result in a human readable model. We start with human generated models, if they exist, and from these existing models we will split or merge KCs in the search of alternative models. Because the resulting models will have a hierarchy from existing labeled KCs, it is much easier for humans to evaluate the results (with the help of our visualization tools). These models can then be evaluated using the AFM statistical model and a variety of metrics to score the models including variants of Akaike information criterion, Bayesian information criterion, and cross validation [8].

2.3 Hint and Feedback Handler

The hint and feedback handler is based on the Hint Factory, our novel method of automatically generating context specific, just-in-time (JIT) hints for multi-step problems [10]. The method is designed to be specific, on-demand, and to provide the right help at the right time. In order to deliver hints and feedback, the Hint Factory first constructs a graph of states and actions that represents all previous student approaches to a particular problem. The state-action graph is transformed into a Markov decision process (MDP). A MDP is defined by its state set S , action set A , transition probabilities T , and a reward function R [11]. The goal of using an MDP is to determine the best policy (i.e., the best path through this graph) that corresponds to solving the given problem. This is achieved by calculating a “value,” the expected discounted sum of the rewards to be earned by following an optimal policy from state s , calculated recursively using value iteration [12]. When the hint button is pressed, the hint provider searches for the current state in the MDP and checks if that a successor state exists. If it does, the successor state with the highest value is used to generate a hint sequence. Once a student performs a correct step, the hint sequence is reset.

Barnes and Stamper (2008) demonstrated the feasibility of this approach on historical data, showing that extracted MDPs with the proposed hint-generating functions could provide correct next-step

hints towards the problem solution over 80% of the time [10,12,15]. A pilot study showed that students were able to solve more problems when hints were included [1]. Since the Hint Factory is data-driven, the system can be bootstrapped with expert solutions [15]. And it can evolve, providing some automatically-generated hints initially, and improving as additional expert information and student attempts are added to the model.

2.4 Student Assessment and Support Tools

It is important for educators to have excellent visualization tools for student assessment. Further, research has shown that interaction with intelligent tutors could possibly be a better predictor of a student's knowledge than standard tests [17]. Such models are the basis for the kind of student-customized adaptive instruction that intelligent tutoring systems can provide [18].

On the knowledge tracing side, the use of visualizations to assess student performance using cognitive models [2] will be used to give educators the ability to assess student knowledge at a given point in time. Croy, Barnes, and Stamper applied a technique to visualize student proof approaches to allow teachers to identify problem areas for students [19]. The goal of implementing these assessment and support tools is to provide educators and administrators the view of student learning and how the system works to support their initiatives and improve learning. The idea that educators can identify areas in SCALE that seem to contradict what they see in the actual classroom will allow for the refinement of the student models in ways that make sense to the educator.

2.5 Support Tools for Content Designers and Developers

A key differentiator of SCALE is the “data-driven, human-centered” approach that achieves superior results to existing systems. The human-centered portion allows developers and educators to explore and improve the models discovered. To implement this, we have designed tools that are based on existing tools in the EDM community (www.educationaldatamining.org). These include:

1. Learning Curve Analysis tool
2. Performance Profiler tool
3. Model Exploration and Tagging tool
4. The Replay Tutor student simulator

The Learning Curve Analysis tool allows researchers to identify smooth learning curves. We expect that the learning curve for each well defined KC will be reasonably smooth. When the learning curve of a purported KC is noisy, with upward or downward “blips,” the student model is suspect. This can often mean a KC needs to be split, since we know that when two KCs are represented as one, the learning curve will not be smooth [3]. If the student model is accurate, we expect the error rate to decline over the number of opportunities a student has to both learn and apply a KC. Thus, a flat learning curve is another indication of a potentially flawed student model. This is especially true if data supports the idea that students are learning a specific KC but the model does not reflect this.

The Performance Profiler tool allows researchers to view error rates by problem or KC and also view the predicted value of one or more models. Problems will be displayed in a column by error rate with shading representing the error. We also will show the predicted values of two proposed models with the lines and points. This will show that one model does a better job predicting some

problems than others. This is useful both to identify problems that may have an excessively high error rate, and to identify problems where the models have a difficult time correctly predicting the student's answer.

The Model Exploration and Tagging tool allows educators and researchers to explore log data in a graph. This tool allows researchers a way to identify places in the instruction where KCs are tagged incorrectly as well as see areas where improvements can be made.

The Replay Tutor allows experts and educators to validate new models using existing log data. The simulator predicts what students might do by using a Bayesian knowledge tracing model [20]. This tool allows researchers to better understand how changes in a model might affect student performance without having to test the model in an actual classroom environment.

In addition to the ability to automatically generate these features, SCALE also includes a “feedback loop” to continue the improvement of the features over time as more data is collected. Several potential customers have noted that existing providers of adaptive student learning software do not have any way to explain why the system behaves as it does. In contrast, SCALE provides tools that let the instructors and developers explore the data using meaningful visualizations that will provide insights into student learning. Often this means identifying areas where the existing models contradict the data collected. Built around the concept of data curation, these tools can also be used to prompt the developers, educators, and users of the educational software for more human input in order to improve the underlying models that the system generates.

The pilot for SCALE has been conducted in multiple courses with our partners HarvardX and the Open Learning Initiative (OLI), and has been reported in citations [21-22]. The full integration and load testing for SCALE, including the dashboard and the problem selection, is discussed in the next section.

3 CONTINUED DEVELOPMENT AND ENHANCEMENTS

After completing pilot testing on multiple platforms and in classroom and online courses, we instituted our iterative feature driven design (FDD) process to evaluate feedback from content and platform developers. This feedback, and SCALE-generated quantitative data, was analyzed and used to determine key enhancements to realize a viable, market-ready version of the research findings on which SCALE is based. Here we discuss four primary feature areas:

1. SCALE Dashboard Visualizations and Functionality
2. Learner Mode Support
3. Problem Selection Extensions
4. Multiple Learner Models

3.1 SCALE Dashboard Visualizations & Functionality.

We have added dashboard support for all of the SCALE APIs, providing visual representations of content structures, including problems and knowledge components (KCs), such that tagging and relationships can easily be entered or maintained. SCALE import/export capabilities have been expanded for support of all data elements and refactored to improve performance. Finally,

visualization features were extended to provide more options for viewing learning curves, such as by allowing any level of content groupings for all students or sets of students. This is helpful when performing testing with experimental and control groups to compare and contrast the learning outcomes for students.

3.2 Learner Mode Support.

In working with various learning platforms, we determined that edtech is often used differently by students, depending on which stage they are in in their learning, and how the technology is integrated with the teaching strategies in the classroom. We have therefore added the ability to support multiple learner modes, triggering SCALE to react in a tailored way to the student's current mode. Although the architecture is built to support any number of learner modes, we have currently identified and implemented the following three: Mastery Learning, Review, and Static Assessment. Mastery Learning mode is used to support the formative stage, so that the learning process and activities can be modified, adapting to the students' current state of knowledge, and thus increasing the efficacy of the learning system and promoting improved student attainment. Review mode is used to support students when they are reviewing previously learned concepts, such as during their preparation for a formal assessment. Review mode is still adaptive, in that the system adjusts what the student will see based on the current student mode. However, during review mode the student will have the opportunity to see every concept (skill or KC) at least once, essentially resetting the thresholds used by the system to determine what the student knows. Static Assessment mode can be used for a summative assessment of the students. Unlike Mastery Learning mode or Review mode, by default Static Assessment does not perform adaptive problem selection. We envision adding support for several additional learner modes as we complete additional research.

3.3 Problem Selection Extensions.

In order to provide the best learning experience for students, the problem selection methodology and algorithms have been enhanced and expanded. First, we have added the ability for the system to determine and record problem difficulty. While a theoretical view would manage this metric based on the number of KCs required to solve a problem, in practice, problems are not always tagged at the level of granularity or thoroughness required to capture difficulty level. So, SCALE now provides a means for problems to be categorized with a difficulty level to help align the problems, in order to deliver content to students at the correct level for each student. This determination is made based on analysis of the logged student data, using KC information and updated student models, and can also be informed by expert entry and tagging.

Second, we have also created a new ensemble selection method. Originally, SCALE used an optimized problem selection methodology, that integrated a moderate amount of random selection, in order to drive efficiency in learning. However, we have found that different students will thrive and excel in distinct ways, based on a variety of problem selection approaches. And so now, SCALE provides the framework for managing different problem selection methodologies using our patent-pending ensemble architecture. This involves supporting multiple learner personas associated with corresponding problem selection methods. While interacting with the learning platform, SCALE will align the problem selection methods based on the learning persona that is connected to each

student. This alignment is determined automatically and in real-time as the student uses the SCALE-enabled learning system. Over time, and potentially depending upon content groupings, student personas may change. In this way the problem selection method will adapt as necessary, to align students with the appropriate persona, and thus the associated problem selection method that is best for each student while learning the subject matter.

3.4 Multiple Learner Models.

Models are built using the data as described previously. SCALE supports managing multiple student models for the same platform and content such that SCALE will learn about the student with each student interaction. On a period or ad-hoc basis, log data is used to refine existing models, sometimes splitting existing models or creating new models. This could be automated to be done in a real-time manner, but to date, content developers and teachers seem to want students to experience the learning through controlled model releases.

4 CONCLUSIONS AND FUTURE WORK.

The process of taking cutting edge research out of the university labs and using it to create a new product has been illuminating in numerous ways. Regardless of how “finished” new technology seems in the experimental stage, there are always refinements that only the real-world testing can discover. We received invaluable feedback and insight from both developers - who implement SCALE to seamlessly incorporate it into their own products - and teachers, who will be using SCALE-enabled technology and dashboards in the classroom itself. On the developer side, we found ways to streamline and improve how SCALE works, what information it delivers, and how both students and problem sets can be grouped together in different ways for deeper analysis. For teachers, we discovered that multiple learner modes is very important, due to the fact that students and teachers use the learning system in different ways for different purposes. Sometimes the material is new and being learned for the first time; sometimes the student needs to review for a test; and sometimes the teacher needs an objective, fixed assessment, that won’t change and adapt. (Of course, it is entirely possible to have adaptive assessments, too, such as with standardized test like the SAT, GRE, and versions of the ACT. However, to create a robust adaptive assessment requires much more student data that is generally available for a particular set of content material. Further, this requirement puts constraints on how and how quickly the material can be updated. And so a static assessment may work better in many situations.)

Our ensemble method of problem selection is quite promising, and also grew out of discussions and feedback with both teachers and developers. For example, students who learn Algebra I in 7th grade may be fundamentally different than student who learn the same content in 10th grade. And thus, these two different groups of students may need different support structures, and our models may require different assumptions in order to optimize learning. This work is very promising but still preliminary. We are in the process of conducting further experiments and refinements on these methods for determining the right selection method for individual students, including retrospective studies using our Replay Tutor and already-collected data; and new studies in participating classrooms. The ultimate goal will be to properly refine the personas: how many are appropriate or necessary, and in what ways is learning improved through their use? Is it more effective, more

efficient, or both? And, importantly, to what degree are these personas generalizable, across content areas and across Learning Systems?

We will continue to address the needs of the market and will continue to conduct targeted pilot tests to assess results of new features and functionality. We are also addressing the need for adding activity selection, and not just problem selection, to find the next best activity to maximize learning. And finally, we have been working to integrate non-cognitive factors into the engine to provide additional data points of student affect and behaviors, to continue to contribute ways to keep students actively engaged and on improved trajectories of learning.

5 ACKNOWLEDGEMENTS

TutorGen gratefully acknowledges support of SCALE® from the National Science Foundation award numbers 1346448 and 1534780 and from the Commonwealth of Kentucky Cabinet for Economic Development, Kentucky Science and Engineering Foundation, and The Kentucky Science and Technology Corporation, award numbers KSTC-184-512-14-182 and KSTC-184-512-16-241.

REFERENCES

1. Barnes, T. & Stamper, J. (2009). Automatic hint generation for logic proof tutoring using historical data. In *Journal of Educational Technology & Society*, Special issue on Intelligent Tutoring Systems, Vol. 13, Iss. 1. 2010.
2. Koedinger, K., Stamper, J. (2010). A Data Driven Approach to the Discovery of Better Cognitive Models. In Baker, R.S.J.d., Merceron, A., Pavlik, P.I. Jr. (Eds.) *Proceedings of the 3rd International Conference on Educational Data Mining. (EDM 2010)*, pp. 325-326. Pittsburgh, PA.
3. Stamper, J., Koedinger, K.R. (2011) Human-machine Student Model Discovery and Improvement Using DataShop. In Kay, J., Bull, S. and Biswas, G. eds. *Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011)*
4. Blink, M.J., Stamper, J., and Carmichael, T. (2014) SCALE: Student Centered Adaptive Learning Engine. In S. Trausan-Matu et al. (Eds.) *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, pp. 654-655, 2014. Springer.
5. Clark, R. E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2007). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593). Mahwah, NJ: Lawrence Erlbaum Associates.
6. Lee, R. L. (2003). Cognitive task analysis: A meta-analysis of comparative studies. Unpublished doctoral dissertation, University of Southern California, Los Angeles, California.
7. Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck, & M. Wilson, (Eds.) *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.
8. Stamper, J., Koedinger, K., McLaughlin, E. (2013) A Comparison of Model Selection Metrics in DataShop. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*. Memphis, USA. Jul 6-9, 2013. pp. 284-287.
9. Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, T.-W. Chan (Eds.)

- Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 164-175. Berlin: Springer-Verlag.
10. Stamper, J., Barnes, T., and Croy, M. (2011) Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. In Kay, J., Bull, S. and Biswas, G. eds. Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011). pp. 345-352. Berlin Germany:Springer.
 11. Sutton, S. and Barto, A.(1998). Reinforcement Learning: An Introduction. Cambridge: MIT Press.
 12. Barnes, T., Stamper, J. (2008). Toward Automatic Hint Generation for Logic Proof Tutoring Using Historical Student Data. In Procs of the 9th International Conference on ITS, pp. 373-382. Berlin, Germany: Springer.
 13. Stamper, J., Barnes, T., Croy, M. (2011). Enhancing the Automatic Generation of Hints with Expert Seeding. In The International Journal of Artificial Intelligence in Education (IJAIED), Special Issue on the Best of ITS 2010.
 15. Stamper, J., Barnes, T., and Croy, M. (2010) Enhancing the Automatic Generation of Hints with Expert Seeding. In Alevan, V., Kay, J., and Mostow., J eds. Proceeding of the 10th International Conference on Intelligent Tutoring Systems (ITS2010). vol. II, pp. 31-40. Berlin, Germany: Springer Verlag.
 16. Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, c., Chen, L., Cosejo, D. I learn from you, you learn from me: How to make iList learn from students. In V. Dimitrova, R. Mizoguchi, B. Du Boulay and A. Graesser (Eds.), Proc. 14th Intl. Conf. on Artificial Intelligence in Ed, AIED 2009, pp. 186—195., Brighton, UK. IOS Press (2009)
 17. Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI). 19(3), pp. 243-266.
 18. Koedinger, K. R. & Alevan, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. Educational Psychology Review, 19 (3): 239-264.
 19. Croy, M., Barnes, T., and Stamper, J. (2008). Towards an Intelligent Tutoring System for Propositional Proof Construction. In A. Briggie, K. Waelbers, & E. Brey (Eds.) Current Issues in Computing and Philosophy, pp. 145-155. IOS Press: Amsterdam, Netherlands.
 20. Corbett, A.T. and Anderson, J.R.: Knowledge tracing: Modeling the acquisition of Procedural knowledge. User Modeling and User-Adapted Interaction, 4. (1995) 253-278.
 21. Rosen, Y., Rushkin, I., Ang, A., Fredericks, C., Tingley, D., Blink, M.J., Lopez, G. (2017) Adaptive Assessment Experiment in a HarvardX MOOC. Proceedings of the 10th International Conference on Educational Data Mining. (EDM), Wuhan, China, June 25-28, 2017. pp. 466-471.
 22. Rosen, Y., Rushkin, I., Ang, A., Fredericks, C., Tingley, D., Blink, M.J. (2017) Designing Adaptive Assessments in MOOCs. Proceedings of the 4th ACM Conference on Learning @ Scale (L@S 2017). (ACM, 2017), Cambridge, MA, pp. 233-236, ACM.

The ASSISTments TestBed: Opportunities and Challenges of Experimentation in Online Learning Platforms

Anthony F. Botelho¹, Adam C. Sales², Thanaporn Patikorn¹, Neil T. Heffernan¹

¹Worcester Polytechnic Institute, Worcester, MA

²University of Texas at Austin, Austin, TX

abotelho@wpi.edu; asales@utexas.edu; tpatikorn@wpi.edu; nth@wpi.edu

ABSTRACT: The ASSISTments TestBed is a platform for conducting small-scale, short term randomized trials within the ASSISTments online learning platform. Any education researcher may propose an experiment, which will be run at no cost. As a learning system, ASSISTments is positioned to augment teacher instruction and help students learn. As a shared scientific instrument, the system aims to facilitate the running of studies to learn what types of instructional strategies and content helps which students most and openly share such information and tools to benefit educational research. Through the exploration and analysis of 9 experiments run within ASSISTments, we describe how these tools are being combined with multiple methods to better identify what works for whom. Toward the goal of more precisely measuring treatment effects, this paper acts as an overview of some of the scientific and statistical opportunities that the TestBed system affords when compared to traditional randomized trials in education. We will argue that this framework represents a promising, if uncharted, avenue in the science of education, and merits the attention of both methodologists and substantive education researchers.

Keywords: randomized controlled trials, testbed, treatment effects, heterogeneous treatment effects

1 INTRODUCTION

The benefits and opportunities made possible through computer-based learning platforms such as ASSISTments extend beyond scientific discovery to include much more practical applications by providing the means to learn what content and instructional practices lead to better student learning. The running of randomized controlled trials has long been the quintessential method of determining the causality of an intervention, and is only augmented through such computer-based systems. The benefit of running RCTs within such systems is not limited to just the scale of the population of students that can be included in a conducted trial, although this too can provide sufficient statistical power beyond what traditional orchestrated studies commonly observe, but rather the benefit is truly in the breadth of data collected for each student, consistency of measures as recorded within the platform, and depth of historical data available within the system that can be leveraged to learn what works best for whom.

A focus on developing methods to more precisely estimate treatment effects is essential in identifying instruction that may be more effective for one group of students than another, and a significant amount of research has been devoted to discovering and developing interventions with heterogeneous effects. Other fields such as marketing and economics arguably have an even longer history of this research leading to methods aimed at measuring such effects (Wager & Athey, 2017). Paying attention to context can help identify the situations and for which subgroups a treatment may have an effect to incorporate more personalized interventions to help students.

Through a series of descriptive and empirical examples using 9 studies run within ASSISTments, the goal of this work is to highlight the importance of developing infrastructure to support the running of randomized controlled trials for the purpose of discovering which instructional practices work, and highlight several methods being applied to more precisely measure treatment effects toward the goal of identifying heterogeneous effects where they may exist.

2 THE ASSISTMENTS ECOSYSTEM

The use of computer-based learning platforms in real classroom settings offer the opportunity to not only test and learn what content and instructional practices benefit students, but also to complete the loop by then deploying successful interventions back to students. It is in this iterative feedback loop that these systems are, at least in theory, able to grow and eventually be able to adapt to meet the needs of students.

The primary goal of this paper is to describe the types of benefits a computer-based learning platform can offer in facilitating scientific discovery and turning research into practice, using a system called ASSISTments to exemplify these opportunities. ASSISTments is a free web-based learning platform made available through Worcester Polytechnic Institute. It is used by teachers and students across the United States for homework and classwork, and has been shown to nearly double student learning over the course of a school year as compared to traditional teaching methods (Roschelle et al., 2016).

The whole of ASSISTments extends beyond a computer-based learning system to form an ecosystem (Heffernan & Heffernan, 2014) of tools that are focused on providing immediate feedback to students in an effort to augment the teacher's ability to provide instruction in a more data-driven procedure. This teacher-focused approach allows teachers using the system to follow the same curricula as would otherwise be used, but, as students are working on the content within the system, immediate correctness feedback can be provided in addition to other forms of aid including hints and scaffolding where such content has been authored (this additional aid is also pertinent to the idea of conducting trials to learn what types of content benefits students most and will be addressed further in the next section). Even without additional student aid, however, just immediate correctness feedback can help a student understand where he/she needs additional instruction and, through reports provided to teachers through ASSISTments, the teacher can too understand where students need further support; instead of going over homework during class, the teacher can know what content was most troublesome for students beforehand and direct time, attention, and remedial instruction during class to address these areas.

2.1 The ASSISTments Testbed

Aside from these attributes that exemplify how a system such as ASSISTments can be used to run RCTs, it is important to further describe the ASSISTments Testbed as this tool extends these benefits to researchers external to the developers of the platform. The testbed defines a process and set of tools that allow researchers to propose, build, and run RCTs through ASSISTments, and also open supplies the researchers with the Assessment of Learning Infrastructure (ALI) tool (Ostrow et al., 2016) that provides a series of automated analyses and access to the anonymized data from the system associated with their study. The testbed therefore provides researchers with the tools

necessary for each aspect of the study design and deployment processes as well as aids in the analyses of such studies; the tool has facilitated over a dozen studies since its deployment resulting in several notable published studies (Fyfe, 2016; Koedinger & McLaughlin, 2016; McGuire et al., 2017).

The ASSISTments Testbed defines a set of 5 steps aimed to guide researchers who wish to propose a study from a research idea through to the publication phase of that study. In this way, its goal is to facilitate the running of randomized controlled trials and openly publishing upon the findings. The aim of the testbed is to make it easy for researchers, both those working with ASSISTments and others external to Worcester Polytechnic Institute from where the system is provided, to run numerous RCTs to test the effectiveness of different learning interventions with teachers and students using the software in real classroom settings. In addition, this further makes it easier to replicate studies on different populations and content within the system, as will be the basis of the example analyses described in the later sections of this paper.

The testbed and reporting infrastructure also acts as the facilitator of the 9 studies exemplified in this work to illustrate the benefits and opportunities made possible through computer-based systems. The next section describes these studies in larger detail.

3 VIDEO VS. TEXT FEEDBACK: A CASE STUDY WITHIN ASSISTMENTS

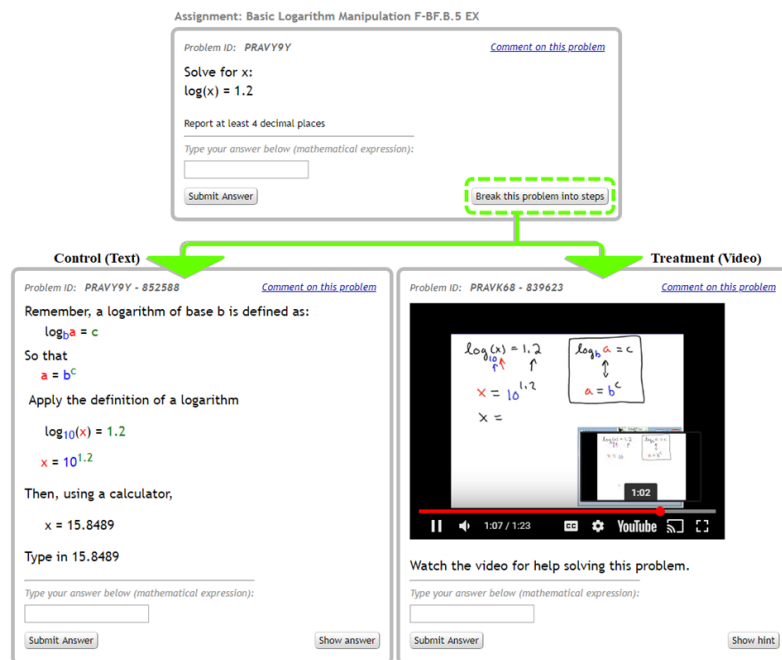


Figure 1: An example experimental design with ASSISTments comparing text-based feedback with video-based feedback when students request help.

To give a better idea of the process through which a study can be proposed, deployed, and analyzed through the testbed, we will describe the steps using an example intervention. Let's say that a researcher comes to the ASSISTments testbed and wants to run a study to test the effectiveness of video feedback for students as opposed to a text-based explanation given to students who need additional help to learn the material. In other words, the researcher wants to randomize what

happens when a student asks for help, giving either a text-based worked example to explain the correct procedure to solve a problem, or a video containing the same information delivered as a video in a more paced manner; this certainly seems like a reasonable comparison as both methods are commonly used in various systems to supplement teacher instruction. With this idea, the researcher proposes to run an RCT within ASSISTments and is given the choice to use the normal population of teachers and students who already use the system for homework and classwork daily, or the researcher can recruit his/her own set of teachers to run a more orchestrated study; for sake of example, we will consider that the researcher chooses to use the teachers and students who normally use ASSISTments. As such, the researcher creates an ASSISTments account and chooses the subject matter on which to run the experiment, and, again for example, let's say that the researcher chooses logarithms as this is a subject that may be difficult for some students and learning what types of aid helps students learn this topic could be meaningful and impactful.

The researcher then creates a problem set using the set of assignment-building tools within ASSISTments aligned to the experimental design; such tools allow the researcher to define, for example, "if-then-else" style and "randomly choose" style rules to define where in the problem set randomization occurs. For instance, a reasonable design may first include a question designed to check if students can see video (as some schools may block such content from sites such as YouTube), and only randomized students who have the ability to see video. After this "video-check" the researcher may define a "randomly choose" section that will randomly assign students to either a set of problems containing text feedback or another, almost identical set of problems containing video feedback; an example of such a condition is illustrated in Figure 1, where a student may be randomized to see either a text-based worked example or a video of the same content when requesting help from the system. Of course more complex designs could also include common design elements such as pretests and posttests, but this example will keep the design simple (and it also represents the general design of each of the studies that will be exemplified in the following section). A problem set created in this way performs student-level randomization, mitigating the need to block students by locale and other factors; although, a researcher may still be able to do so, albeit through a slightly more complex orchestrated design.

Once the problem set is created and approved by a team of researchers and content experts working with ASSISTments (to ensure that the content is not inherently harmful, broken, offensive, or otherwise in violation of IRB terms), the problem set can be deployed amongst the ASSISTments-certified content within the system. While teachers have the ability to create their own content with ASSISTments, many simply choose to use the existing content that has been implemented into the system. When a teacher assigns the particular research-created content, students are randomized and the data is recorded. After a predetermined amount of time, the study is retired and the researcher can begin the planned analyses.

As mentioned above, a tool, called the Assessment of Learning Infrastructure (ALI) aids researchers in the collection and initial analyses of data. Researchers request the data from their experiment by providing ALI with the problem set information and then receive an email containing some initial basic analyses and statistics (e.g. the number of students randomized to each condition as well as completion rates split by condition with a chi-squared test to identify if there is differential attrition between the two conditions). In addition to these descriptives, the researcher gains access to

anonymized datasets containing the student data at various granularities including problem-level, action-level, and also student-level covariates generated from data before random assignment to condition (i.e., the student's prior percent correct, prior completion, etc.). With this data, the researcher can perform the planned analyses and write the report on their results, citing the initial design document and ALI report to promote open data and science.

This In continuing our example of experimentation through the ASSISTments Testbed, we exemplify a set of nine studies run in ASSISTments comparing text-based and video-based feedback for students. Data from experiments run on the platform of ASSISTments have long been made open and available for researchers to analyze. In 2016, for example, a dataset of 22 such experiments run within the system were published (Selent, Patikorn, & Heffernan, 2016) and made open in the hopes that interesting analyses and methods could be applied to better estimate treatment effects and also to motivate other companies and institutes who run RCTs on their own respective platforms to similar see value and make such data open and available. The nine studies observed here are amongst the 22 and are particularly of interest as they apply the same comparison of video versus text feedback. In this way, they act as 9 replications of the same idea and can be used to exemplify some of the challenges and applicable methods available to address such challenges.

These studies were run in mastery-based assignments called "skill builders," where the system provides students with problems until they are able to demonstrate sufficient understanding of the material (e.g., a student must answer three consecutive problems correctly without the use of computer-provided aid), and each student must meet this threshold in order to complete the assignment. Students who are unable to learn the material by the tenth problem are asked to seek additional help, and the assignment is left incomplete (while there are various settings that allow teachers to control each threshold and how to address struggling students, the data used here aligned to the described defaults). We observe the effectiveness of the treatment with regard to the outcome measures of student completion as well as a measure called "inverse mastery speed," calculated as 1 divided by the number of problems needed to complete the skill builder assignment.

3.1 Methods to Reduce the Standard Errors of Effects

While ASSISTments and the accompanying ASSISTments Testbed provide infrastructure and tools to run experiments, these alone are not the entire solution to the problem of finding which interventions work for which groups of students. What are missing from these examples thus far are methods that can help to more precisely measure the effects of a particular treatment. Whenever calculating a treatment effect, the ability to accurately measure the impact that the treatment has on any particular outcome is dependent on the magnitude of the effect, but perhaps more importantly, the scale and variance of the population of students included in the study; the more students included in an experiment, the smaller the standard errors on that effect tend to be (i.e. larger samples tend to allow for more precise estimates of the effect). While this goal of reducing standard errors is applicable to any experiment, it becomes much more important to consider when exploring potential heterogeneous effects. If it is difficult to precisely measure a treatment effect across the entire population of students in a particular study, it is much more difficult to measure such effects when observing smaller sub-groups of students.

The next 3 sections therefore describe and compare two methods that are being applied with this specific goal of measuring treatment effects with greater precision. While the examples themselves will not explicitly explore the potential heterogeneity of the interventions, this paper presents some of the pilot work in this area.

3.1.1 *Regression to Mediocrity*

It is a well-documented issue that a crisis is currently affecting several scientific fields in that, for any number of reasons, experimentation across fields is failing to hold to replication (Ioannidis, 2005a, 2005b). If we wanted to know the true effect of video feedback as compared to text feedback on the outcome measure of completion, for example, due to random variation in content, population, measures, etc., we are likely to observe varying estimates with each replication. In some cases, a replicated effect may appear to have a statistically reliable positive result, while another may show the exact opposite, with many others may show no statistical reliability.

A range of statistics research has been devoted to this and similar problems (Rubin, 1981), but the concept for which we are focusing is that of “shrinkage” (James & Stein, 1961; Efron & Morris, 1973). Also referred to as regression toward mediocrity (or regression to the mean) (Galton, 1886; Stigler, 1990), the idea is that if we run multiple replications, sometimes our estimate will be too high and other times too low; as we run more replications, the average of our estimates will begin to regress toward the average true effect. Other work has been inspired by the same idea, attempting to use the consistency of data collected across experiments to increase power in estimating effects for individual experiments (Patikorn et al., 2017). Here, however, we describe a different approach called “partial pooling.” The idea of this method is, instead of analyzing each experiment individually and independently, we can pool together similar experiments that we think should have the same effect at once (e.g. replications of the same or similar treatment) in order to better estimate the effects for all pooled experiments. Partial pooling reduces the variance of the estimated effect size of each experiment by looking at the variances and the estimates effect sizes of other experiments, causing the new estimates to shrink toward the mean of the estimated effect distribution.

A drawback to this approach, however, is that it does bias the new estimates toward the overall mean; such is, after all, the purpose as the mean of effect estimates is believed to be a closer estimate to the true effect. Despite this, yet another method may be used to better estimate effects without such a bias. We describe this method in the next section.

3.1.2 *A Role for the Remnant: A Model-Based Approach*

The idea of applying partial pooling works well in the case of computer-based systems running experiments due to the consistency of measures collected across students (although the method itself is not inherently limited to cases where the measures are as consistent as used here), as the system records the same information for each student. However, this is also true for all students using the system, not just those who participate in an experiment. So what, then, can be learned from all the students who are not randomized to condition? In the case of ASSISTments, there are hundreds if not thousands of students using the system every day, and if we could utilize their data to better analyze experiments, the added power is likely to help reduce standard errors on the estimated treatment effect.

Previous work explored the use of this population of students external to the experiment, which has been referred to as the “remnant,” to more accurately estimate treatment effects (Sales, Hansen, & Rowan, 2018). The remnant essentially consists of all students who have ever used the system that were not a part of any of the current experiments under analysis; they may have been a part of previous experiments but, for instance in the case of our example, it includes a large sample of students disjoint from those who participated in any of the 9 example RCTs. But what, if anything, can be learned from this group? No randomization occurred for these students, and there is no guarantee that a condition in the experiment is “normal” behavior, meaning that the manner in which students interacted with the system during the experiment as compared to normal usage may be very different. What we do know, however, is that data pertaining to outcomes of interest (i.e. assignment completion, knowledge level and correctness, number of problems needed to complete mastery-based assignments) is available for the remnant as well as those in the experiment.

It is from this idea that a method called “remnant based residualization,” or REBAR (Sales, Hansen, & Rowan, 2018; Sales, Botelho, Patikorn, & Heffernan, 2018), was developed. The process is rather intuitive. First, we can build a model using the remnant to predict an outcome measure of interest. In our example case, we use the remnant to train a model to predict whether a student will complete an arbitrary next assignment. Second, the trained model is applied to predict the outcome measure for those in the experiment. Third, the estimates of the model (our prediction of whether each student will complete the experimental assignment), are subtracted from the actual outcome; this step is essentially removing variance from the outcome measure of interest that can be explained away by the model trained on the remnant. From this point, the last step is to simply analyze the experiment using any desired method using the residual in place of the actual outcome. As the model is trained on a population completely external to the participants in the experiment, the estimates are unbiased. For this reason, the estimates themselves do not even need to be accurate; a bad model should be just as bad for everyone (on average). However, the better the model is at predicting the outcome, the more variance that can be accounted for within the experiment leading to more accurate treatment effect estimates.

3.1.3 *Why Not Both?*

As mentioned in the previous section, the last step of the REBAR method uses the residual to run any set of desired analyses. For this reason, the REBAR process and the described partial pooling method are disjoint approaches and therefore could be combined to even further reduce standard errors of the estimated treatment effects. In this way, we can take advantage of both the scale and breadth of data made available through the use of the remnant, while also taking advantage of the consistency of measures across the experiments.

We use the model estimates from the REBAR method for both outcomes measures of completion and inverse mastery speed as described in the previous section. The estimates are subtracted from the observed outcomes following the REBAR methodology, and then the resulting residual is used in the Bayesian partial pooling approach. The combination of these two approaches results in the reduction of standard errors across all example studies. As shown in Figure 2, the combination of methods reduces the standard errors of all experiments when compared to the traditional method and is superior or at least comparably similar to either method alone.

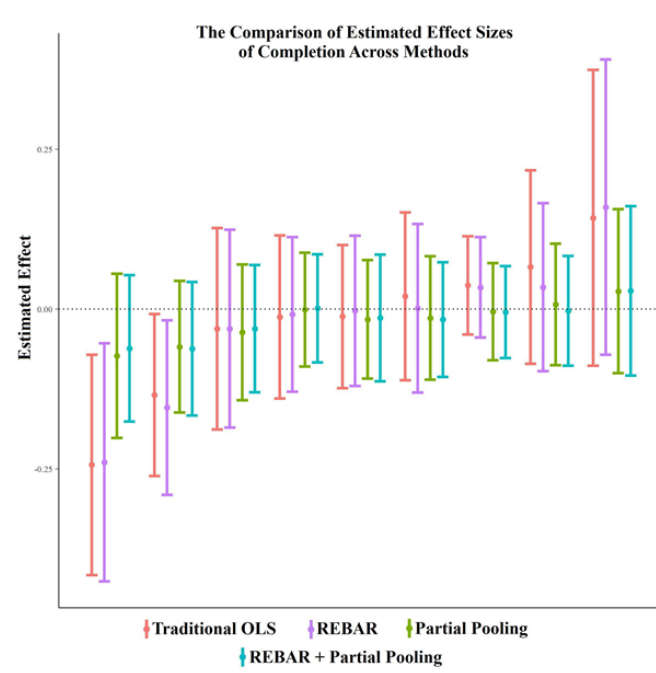


Figure 2: The estimated treatment effects of student completion for each of the 9 experiments across all methods.

In consideration of the second outcome measure of inverse mastery speed, the combination of methods again leads to considerable reductions of standard errors beyond that of the traditional method in all experiments, as seen in Figure 3. Similarly to that of Figure 2, the combined method performs better or comparably similar to either other method alone.

It can be seen in both analyses, however, that the combined method does not lead to the smallest standard errors in every case. It is important to explore and understand, as is the goal of ongoing and future work, when each method is likely to lead to improvements in precision. Regardless, these methods show promise in their ability to aid in the analysis of experiments and discovery of potential heterogeneity in the measured effects. It is also the case that the methods helped to remove some of the variation of the 9 replicated studies, where the combined method no longer results in statistically reliable effects in any of the experiments; it is important to emphasize, however, that this is largely influenced by the partial pooling methods bias toward the mean effect measured across all experiments and that these particular experiments were chosen for these analyses in-part for exemplary purposes. Ongoing and future work is further exploring the application of these methods at larger scale across multiple experiments running through the ASSISTments Testbed.

4 CONCLUSIONS

The issues and challenges faced by the field as it moves toward new experimental environments and, through these, new data environments, are by no means novel, but rather tools such as computer-based platforms are merely allowing us as researchers to finally address these problems in more practical ways. It has always been a challenge to design replicable RCTs to test ideas; this is a challenge for replicability of results (i.e. the same or similar findings and conclusions are reached

after additional trials), but also in a much more direct interpretation of replicability, where a design can be replicable. Computer-based systems offer new ways to allow for clear replication, using the same design for new populations or contexts, using the same measures calculated in the same ways across all experiments. This consistency alone offers new opportunities to more accurately evaluate instructional strategies and the like.

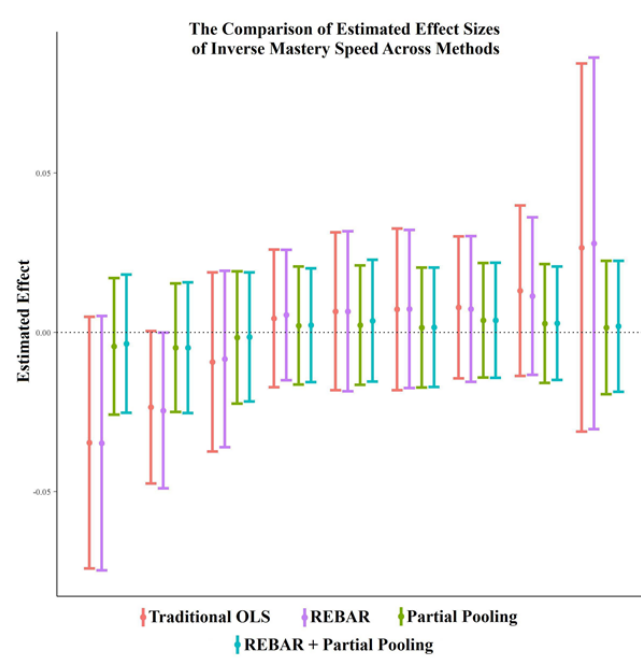


Figure 3: The estimated treatment effects of student inverse mastery speed for each of the 9 experiments across all methods.

Just as replication has been a challenge, the ability to accurately estimate treatment effects is another long-standing issue. It is important to consider how existing methods can best be combined with the opportunities that computer-based systems offer. Where in the past collecting data from several dozen students served as a challenge to any researcher intending to run a randomized controlled trial, it is now more trivial to collect data from several hundred students, if not more, through such systems allowing us to direct more focus to the other prevalent challenges. Issues such as testing ideas in new contexts or identifying heterogeneity become much more feasible as the scale and replicability of studies becomes easier.

We refer to and describe a number of studies and research in this article that have been facilitated by ASSISTments and the ASSISTments Testbed, but these are small examples compared to what is currently possible with these and similar tools. These tools in combination with the development and application of methods to more precisely measure treatment effects holds great promise in regard to the goal of discovering what works (and what does not work) for particular groups of students.

REFERENCES

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341), 117-130.

- Fyfe, E. R. (2016). Providing feedback on computer-based algebra homework in middle-school classrooms. *Computers in Human Behavior*, 63, 568-574.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246-263.
- Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4), 470-497.
- Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2), 218-228.
- Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLOS Medicine*, 2(8), e124.
- James, W., & Stein, C. (1961, June). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1961, pp. 361-379).
- Koedinger, K. R., & McLaughlin, E. A. (2016). Closing the Loop with Quantitative Cognitive Task Analysis. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 412-417).
- McGuire, P., Tu, S., Logue, M. E., Mason, C. A., & Ostrow, K. (2017). Counterintuitive effects of online feedback in middle school math: results from a randomized controlled trial in ASSISTments. *Educational Media International*, 54(3), 231-244.
- Ostrow, K. S., Selent, D., Wang, Y., Van Inwegen, E. G., Heffernan, N. T., & Williams, J. J. (2016, April). The assessment of learning infrastructure (ALI): the theory, practice, and scalability of automated assessment. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 279-288). ACM.
- Patikorn, T., Selent, D., Beck, J., Heffernan, N., & Zhou, J. (2017, January). Using a Single Model Trained across Multiple Experiments to Improve the Detection of Treatment Effects. In *10th International Conference on Educational Data Mining*.
- Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, 2(4), 2332858416673968.
- Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4), 377-401.
- Sales, A. C., Hansen, B. B., & Rowan, B. (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, 43(1), 3-31.
- Sales, A., Botelho, A. F., Patikorn, T., & Heffernan, N. T. (2018, July). Using Big Data to Sharpen Design-Based Inference in A/B Tests. In *Proceedings of the Eleventh International Conference on Educational Data Mining*, 479-485.
- Selent, D., Patikorn, T., & Heffernan, N. (2016, April). Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 181-184). ACM.
- Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5(1), 147-155.
- Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, Doi: 10.1080/01621459.2017.1319839

Instrumenting Courseware and Leveraging Data with the Open Learning Initiative (OLI)

Norman Bier, Stephen Moore, Martin Van Velsen

Carnegie Mellon University

nbier@cmu.edu, {stevenmo, martin.v}@andrew.cmu.edu

ABSTRACT: Founded in 2002 as part of the Hewlett Foundation's inaugural open education grants, the Open Learning Initiative (OLI) is a recognized leader in adaptive courseware and learning engineering, combining leading research in cognitive and learning science with state-of-the-art technology to create adaptive, open courseware that enacts instruction. By rigorously capturing and evaluating learner data, OLI drives powerful feedback loops that assist learners, support educators, improve courses, and drive learning science research. This workshop will provide an overview of creating instrumented courseware with OLI's tools, aligning measurable, student-centered learning outcomes with active learning activities and assessments. We will provide examples of the data generated by OLI learner interactions and show how this data is used to provide feedback to learners and drive analytics for both instruction and course improvement. Finally, we will show how OLI data is made available for research, teaching participants how to access this information and providing examples of how this data has been used to support primary research, secondary analysis, and ongoing analytics work. Participants will leave with the ability to build their own OLI courses, the ability to access OLI data for their own work, and contacts for ongoing engagement with the OLI team.

Keywords: OER, Instrumented Courseware, Iterative Improvement, Learning Engineering, Learning Analytics

1 INTRODUCTION – ENGINEERING LEARNING

The Open Learning Initiative (OLI) serves as a combination research and development project at Carnegie Mellon University (CMU), integrating with the larger work of the university's Simon Initiative. OLI focuses on developing, using, improving, and researching science-informed, open courseware as a key element of a community-based research activity focused on understanding and improving human learning.

Central to the Initiative is an approach, born at CMU, that Nobel Laureate Herbert Simon dubbed *Learning Engineering*: the use of learning research and the affordances of technology to design and deliver innovative, instrumented educational practices with demonstrated and measurable outcomes. This close integration of research, data, and instructional practice contrasts with the approaches of many other institutions, where instructional design is frequently based on intuition rather than research, and where technology is often implemented for its own sake rather than as a reasoned, supportive part of a larger instructional research agenda. From its home in the Simon Initiative, OLI offers an exemplar of the success of the learning engineering approach.

2 THE OPEN LEARNING INITIATIVE

Founded in 2002 as part of the Hewlett Foundation's inaugural, pioneering open education grants (Kernohan, & Thomas, 2018), OLI is a recognized leader in adaptive courseware, learning engineering, and open education, combining leading research in cognitive and learning science with state-of-the-

art technology to create adaptive, open courseware that enacts instruction. By rigorously capturing and evaluating learner data, OLI drives powerful feedback loops that assist learners and educators, improve courses, and contribute to our larger understanding of how humans learn. Developed by multi-disciplinary teams, OLI courses can be used to support independent learners, but are primarily designed to support a hybrid instructional model and toolset that maximizes faculty time and expertise. This approach makes OLI unique in the open educational resources (OER) space; while many open projects focus on a loose collection of openly licensed assets, or on developing static OER textbooks, OLI's courseware offers a fully designed learning experience. This experience combines expository content, dynamic activities, and specialized technologies (including labs, simulations, tutors, and other domain-specific learn-by-doing activities). While the expository materials can be downloaded from OLI to create a traditional OER textbook, the complete courseware offers a much greater set of benefits. Data from learners' interactions with these activities, in conjunction with the model of expertise developed as part of the course's design, supports a wide variety of opportunities to adapt to learners' needs. These can include targeted feedback and hints that address demonstrated learner misconceptions, as well as sequencing of problems and activities based on learner achievement, all presented within the context of developing better metacognitive skills and awareness on the part of the student. This same information supports faculty as they design their classroom instruction, with an advanced analytics dashboard that provides detailed learning estimates in relation to the skills and learning objectives specified in the cognitive model. Many analytic systems focus only on engagement or performance metrics; the OLI dashboard estimates learning based upon all aspects of a learner's interactions with assessments (Lovett, 2012). In addition to the benefits for learners and classroom educators, these data also offer benefits in the aggregate, providing insights on course performance that can support faculty in empirically improving the design of a course over time.

2.1 OLI Results

Extensive research has demonstrated the success of the OLI approach in postsecondary education. Studies show dramatically improved outcomes, savings in cost and time, and improved learning productivity over time. Perhaps the best known of this work has focused on the use of the OLI Statistics course; this accelerated learning study demonstrated improved outcomes for CMU learners spending less than half the time of their traditional peers (Lovett et al., 2008). Studies of OLI in collaboration with larger public universities have also demonstrated the scale interventions to large numbers of learners, improving outcomes while lowering costs (cf., Bowen et al., 2014; Griffiths et al., 2014). Recent studies have found that the impact of OLI's learn-by-doing activities can be six times that of other instructional approaches (Koedinger, Kim, Jia, McLaughlin, & Bier, 2015), and follow-on studies have indicated that this doer-effect is both causal and is observable in a multiple number of domains and learner contexts (Koedinger, Jia, McLaughlin, & Bier, 2016).

Ongoing studies continue to investigate the role of OLI with different learner populations, and results suggest that the use of OLI activities can help to smooth out expected negative outcomes often associated with vulnerable and under-prepared learner populations (Evans, Leinhardt, & Yaron, 2008; Kaufman, Ryan, Thille, & Bier, 2013; Ryan, Kaufman, Greenhouse, She, & Shi, 2016). Over the past decade, 40 OLI courses have seen enrollments from over four million independent learners. These same courses have been used to support academic classes in hundreds of institutions of higher

education and high schools, with more than 500,000 enrollments in these types of credit-bearing contexts. This effort has also contributed to extensive research in understanding how human beings learn, including the generation of hundreds of learner interaction datasets that have been used for primary and secondary analysis. This represents an exceptional community of educators, learners, and researchers with whom workshop participants can engage (and who will benefit from the project).

2.2 Scaling OLI Course Development

As the OLI project has grown, it has become increasingly clear that the need for large teams, extended timelines, and deep technical expertise has been a barrier for scaling the community involved in OLI course development (Herckis & Smith, 2018). Similarly, though learning engineering tools and approaches to leveraging data for iterative course improvement are remarkably sophisticated, these tools have often required more time and expertise to implement than is reasonable for most faculty. This challenge has been compounded by the multiple systems and interfaces required to leverage these improvement tools (Bier & Jerome, 2012). To address this need, OLI has made significant investments in developing an integrated authoring suite to support a broader community in the development, improvement, and refinement of open courseware for the OLI system. Preliminary development efforts focused on easy, WYSIWYG authoring capabilities that allow any faculty member to easily develop OLI course materials, not merely as a set of content, but as an integrated learning experience that provides appropriate semantic context to the materials and supports the easy tagging of skills and learning objectives to all learning activities, providing a foundational cognitive model of expertise for the course.

Subsequent development has focused on upgrading this suite into a more thorough workbench for supporting all faculty in learning engineering. This has focused on two major components: 1) embedding into the system elements of instructional design intelligence and learning engineering support that have traditionally been provided via human consultation, thereby scaffolding the authoring process to encourage best practices for learning from the beginning; and 2) embedding analytics for course improvement directly into the authoring view, making learning data actionable for faculty and lowering barriers for continuous, iterative improvement.

2.3 OLI Course Improvement Analytics

These **course improvement analytics** build on the successful prototypes developed under NSF Grant 1418244 (Data-Driven Methods to Improve Student Learning from Online Courses) and provide a range of insights into the underlying design and effectiveness of the course. These improvement analytics include three core elements:

- **Course Design Analytics**, showing the breadth of learning activities and assessment opportunities in relation to the skills and learning objectives that constitute the cognitive

model of the course. This view supports improvements in the robustness of the course's design and can be used even before student learning data is available. Such use supports more effective preliminary design and ensures that the data gathered from student use will offer a fuller set of actionable improvement opportunities.

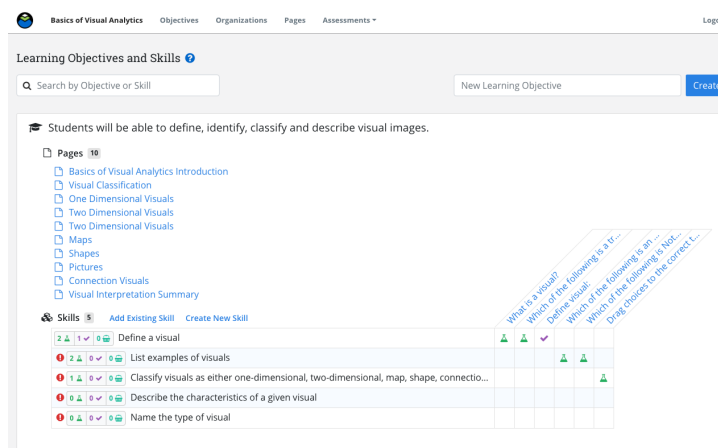


Figure 1: Analytics for Course Design

- **Effectiveness Analytics**, offering insights into both larger learning activities and individual questions. These analytics include traditional item-response theory (IRT) models, difficulty analysis, and views of student use and engagement patterns. These views can also provide insight into the role of individual activities in relation to the larger learning objectives and skills with which they are associated. The interface supports improvement and investigation at a variety of scales and levels of detail, from course-level (“Show me the units with the largest disconnect between practice and exam success”) to objectives (“Which objectives are students not succeeding in”) to individual questions (“This question is one that most students are not getting correct, even after multiple attempts”), and includes summary-level dashboards and analytics embedded directly in the authoring interface.
- **Cognitive Model Analytics**, building upon ongoing work at the Simon LearnLab (Pittsburgh Science of Learning Center) to understand learning model behavior and identify mismatches between expected and actual learner behavior, offering opportunities to improve the underlying model of expertise that is represented by the course. These elements build on decades of tools and methods for optimizing learning through cognitive model discovery and refinement (Stamper & Koedinger, 2011), particularly Learning Curve analysis, a method to identify latent variables in a logistic regression model called the Additive Factors Model (AFM), which is a generalization of IRT (e.g., Wilson & de Boeck, 2004)

By embedding these analytics tools directly into the authoring interface, we make it more likely that instructors will use them. By carefully leveraging design and user experience expertise from the Simon Initiative community, we build these tools so that any faculty member can interpret and act upon the insights they provide. And by engaging with a larger community, these tools are successfully tested, improved, and used by faculty at a diverse array of institutions (Shestak, 2017; Richie, 2018). Together, this suite enacts core elements of instructional design, guiding and scaffolding authors in the development of instructional materials that are as robust as possible, which will provide sufficient data to engage in an iterative improvement process which leverages that data for an empirical approach to course improvement.

3 OLI LEARNING ACTIVITIES AND DATA

The OLI platform is a collection of tools for creating and delivering online instruction that embeds core learning science principles in the system’s design, capabilities, and navigation. Content in the system combines *structure*, *learning objectives*, and traditional *expository materials* (text, examples, images, videos, etc.) with *native activities*—learn-by-doing interactions which offer practice, targeted feedback, and robust hints. Together, these components provide a structured, complete, and supported learning experience. As part of the course development process, the *semantic context* for each of these elements is also captured; OLI defines a learning taxonomy using a series of DTDs¹ that provide additional structure to the learning environment and capture the pedagogical intent of specific components. For example, exposition is captured not merely as a series of textual elements but rather is specified as worked-examples, theorems, learn-by-doing opportunities, self-assessments, and many other semantic elements. This semantic context informs the data that is collected from learners’ use of the course, allowing for more meaningful research and analysis than that offered by more free-form design and click-stream collection approaches. The design of the system has been further enhanced by UDL principles to increase flexibility, address learner variability, and allow learners multiple ways to recognize, act on, and engage with knowledge. These pre-defined capabilities may not always provide the full capabilities necessary for new approaches, domain-specific activities, or experiments. Therefore, the system also provides mechanisms for incorporating other non-core technologies, via APIs. Such non-core technologies include standard elements that are used frequently in courses, including certain types of labs, simulations, and cognitive tutors. These technologies can also include less standard, more experimental elements; as technologies and their associated pedagogical approaches become less experimental and better tested, their use becomes more standardized, eventually moving towards integration with the core system.

3.1 Native OLI Activities

Expository content forms an important part of OLI learning environments, but more important are OLI’s native activities. These active learning activities provide students with opportunities to answer questions and solve problems, with targeted feedback and help. By aligning these activities with the course’s student-centered, measurable learning outcomes, the OLI system is able to continuously assess student learning. These activities support a range of machine-evaluated question types (multiple choice, fill-in-the-blank, short answer, multi-select, ordering, hot-spot, and others), along with feedback and hints. Formative assessment within the OLI system is provided in-line, with these activities presented within the flow of other expository elements. Such activities are considered “low-stakes” within the OLI context—learners do not receive a score for these activities, activities support an infinite number of attempts, and instructors are unable to see the specific results of an individual student’s success or failure for a given activity. (Instructors are able to see that a student has completed an individual activity, and they are presented with an aggregate view of their class’s success; in this way, low-stakes activities present students with a “safe space” to practice, without being penalized for mistakes.) Low-stakes activities have two different semantic contexts, called purpose types, based on the pedagogical intent of the activity: **Learn By Doing (LBD)** activities are

¹ <http://oli.cmu.edu/dtd/>

inserted to provide students with opportunities to practice or master new knowledge and skills, and assume that the learner will make mistakes along the way. **Did I Get This (DIGT)** activities are presented as self-assessment opportunities, provided at points where it is anticipated that the learner should have mastered a specific skill (and offering additional context to support the learner in metacognitive skills development and guidance for self-remediation, if needed). By design, the student is assumed to not yet have mastery, so the exercises should be tailored to include some instruction and reinforcement through the question content itself and from the immediate feedback offered for correct and incorrect answers. Questions can include hints to provide support to students as they learn. OLI adds one to three hints where appropriate, following this pedagogical rule-of-thumb: (1) Restatement: What is the question asking? (2) Cognitive hint: Here are steps you should take. (3) Bottom-out hint: Used in numeric or text-input only, where the student may not be able to get to the correct response or feedback on their own. This same approach also supports summative assessment—high-stakes **Quizzes** and **Checkpoints**—which provide more detailed scores and information for instructors and can be used to calculate grades in formal learning environments.

Compared to many online systems, such as learning management systems which focus on collecting navigation and clickstream data, OLI's native activities form the heart of a richer dataset. Each activity is broken down into one or more problem steps (Antonenko et al., 2012; Psaromiligkos et al., 2011). For instance, if a question asks a student to set the value of three dropdown boxes, then that question has three steps. In addition to the traditional timestamps and UI elements the student interacts with, each step is assigned a set of one or more hypothesized competencies or knowledge components (KCs) required by the student to answer the question (Stamper & Koedinger, 2011; Koedinger, Corbett, & Perfetti, 2012). This KC tagging of the questions, in conjunction with their accuracy, time on task, and number of attempts, provides detailed insights into which concepts students struggle with the most. In particular, the KC mapping provides a comprehensive modeling of the student learning process and enables both students and instructors to better assess their learning. Moreover, when used in conjunction with the additional semantic context provided by the OLI course structure, this data can be used to more meaningfully understand demonstrated learner misconceptions, evaluate course design elements, and provide information for primary and secondary learning science research and analysis.

3.2 Integrating Custom and Third-Party Activities

As a result of the increasing specialization of learning technologies, most current learning platforms depend on external learning tools, the consequence of individual companies and organizations tackling a unique type of student interaction or learning domain and its resulting technology. Additionally, the vast diversity of available interactive learning content makes it impossible for any single platform to support it all natively. The OLI platform is designed to build weak links and strong bonds to externally provided learning tools. The platform does not place many software constraints on the technologies it integrates with, but through usage of APIs, it creates a strong bond to its student performance analysis system (Dashboard, Logging, DataShop, etc.).

One benefit of the OLI platform's integration mechanism is that it easily allows the inclusion of research prototypes. The platform currently supports approximately 40 custom integrations, the majority of which are research-oriented. Some of these research projects are small add-ons which, for example, log specific user interactions (such as page interaction behaviors), but many of the

integrations are full-scale research platforms in and of themselves (e.g. VLab; Aleven et al, 2016; Blink et al, 2014).

4 OLI DATA

Broadly, OLI data is classified into three categories of learning interaction analysis, each with its own client, log service, and processing components.

- **Student page interaction data** is captured as a log stream that records students' basic interactions with learning content. Questions such as, "How are students navigating the course materials?" and, "How much time is spent on learning activities?" can be answered from this data.
- **Student learning in activities** is captured such that feedback can be provided, student responses can be graded, and skill data can be updated. Within OLI, log data can capture additional learning behaviors, and can capture arbitrary additional data elements (specific to individual activity types); this approach provides rich source of information from which many different views of a learner's performance can be extracted.
- **Student problem navigation data** is produced by learning activities that require students to engage in more expanded and involved interactions with problem materials. For example, a math problem might require multiple simplification steps. For such multi-step problems, OLI logs data in DataShop/Tutor messaging format, which is a transaction-based format that can capture the precise way in which a student arrives at an answer.

4.1 Emerging Data Trends

Learning analytics and algorithms continue to provide a deeper view into student learning. In order to support new tools and techniques, OLI provides an extendable and interoperable method of logging data. See the figure 2, below, which outlines how the OLI architecture supports semantic data analysis.

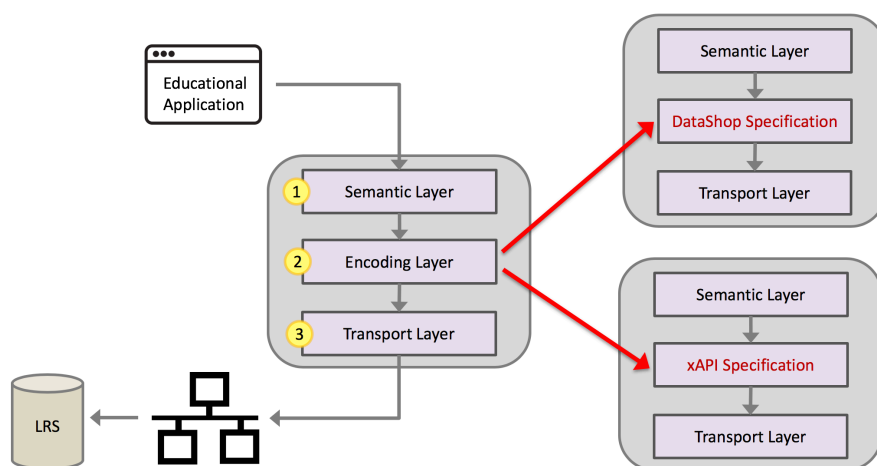


Figure 2 Data Logging Architecture

Learning content software developers have access to three layers of data encoding (note that in Figure 2, the term *Educational Application* denotes integrated learning activities as well as content pages and native OLI activities).

- 1) At the highest level, developers are concerned with ensuring that the meanings of student interactions are preserved. For example, it is important that analysis can discern what the state of a learning activity is when a student asks for a hint. Knowing the state defines what

feedback is given, and paired with student request and tutor response, allows analytics to understand where and why a student is struggling.

- 2) Each message needs to be encoded such that the receiving end can determine the intent of both the student and responses by an automated feedback system. A number of different formats are available, each with slightly different goals and specifications. Since the OLI platform possesses no *a priori* knowledge of which analysis system will be used and which data format it uses, it ensures that an abstraction within its logging code can switch the data encoder. It is our intention to make these tools public and accessible to the greater technology-enhanced learning community, and we are therefore in the process of making these tools open source².
- 3) The message delivery from web browser to log service needs to be robust and efficient. This layer of the architecture supports message bundling to ensure that the browser has as few connections to the log service as possible, and the transport layer also supports retries and a local queue in the event of a log service becoming unresponsive due to networking issues.

4.2 Data Formats

A core aspect of the OLI approach to learning is our data-driven student model; the OLI platform captures exhaustive, real-time data on student interactions with learning materials and instructor interactions, and with learning materials and analytics tools. This data is used to drive feedback loops for learners and instructors (often in real time), as well as for Learning Engineers (for iterative course improvement) and Learning Scientists (for ongoing research and evaluation).

This exhaustive approach to data capture means that, in theory, any researcher, designer, or engineer can assemble the necessary data components for their current tasks or inquiries; in practice, however, the components are captured at a grain size fine enough that significant amounts of aggregation and pre-analysis are necessary to provide information in a useful form. To that end, OLI has a number of standard reports that capture the most frequently used approaches to our data.

4.2.1 Course Design and Improvement

OLI offers a number of reports that provide insight into the performance of learning materials and highlight potential areas for improvement. These reports range from raw numeric data to more carefully processed and designed spreadsheets which have been refined over multiple iterations. These more heavily processed reports exist to make the data visualization easier and more accessible to course authors and learning engineers, with color coding used to provide a first pass at interpretation and identification of potential problem spots in the course. Data for design and improvement includes:

- Number of students
- Average number of attempts
- Average help need

² <https://github.com/Simon-Initiative/DataShopLogger>

- Eventually correct
- First try correct
- Utilization, completion, and accuracy rates
- Chart of low- and high-stakes performance per skill
- Chart of low- and high-stakes performance per learning objective
- Aggregate skill view showing potentially problematic skills where:
 - Assessments are missing
 - Practice is inadequate
 - Assessment and practice may be misaligned
- Aggregate learning objective view showing potentially problematic objectives where:
 - Assessments are missing
 - Practice is inadequate
 - Assessment and practice may be misaligned

Current plans for the improvement of course design and improvement analytics include embedding information from these reports directly into the course authoring platform.

4.2.2 Research and Evaluation

DataShop: OLI course data can be loaded into LearnLab's DataShop³, providing an extensive range of analytic and reporting tools. DataShop spans the gap between research and improvement, with capabilities and methods that can be used for research and evidence-based course improvement. Tools include:

- Knowledge Component Modeling
- Learning Curve Analysis
- Problem Breakdown
- Performance Profiler
- Error Report

See: <https://pslcdatashop.web.cmu.edu/Project?id=122>

Evaluation Dataset: This data set is used when conducting more formal evaluation studies; it's an exceptionally large set normally accessed via a database, though export to CSV is possible. It contains aggregated information that can be used to analyze and answer questions including:

- To what extent did students access OLI content?
- To what extent did students complete the high-stakes assessments?
- To what extent did students' use of the course go beyond simply accessing/completing activities and assessments in a way that could have led to gains in their learning?
- What were student success, assistance, and help-seeking behaviors for low-stakes activities?
- How well did students perform on their initial attempt at any high-stakes assessment?
- How well did students perform on their last attempt at any high-stakes assessment?
- What were faculty access and use patterns for tools, analytics, and content?

³ <http://pslcdatashop.web.cmu.edu>

5 BUILDING COURSES WITH OLI

While the traditional OLI design process has proven successful in creating online learning experiences that demonstrably enact learning and support instruction (Thille & Smith, 2011), the process for developing OLI courses has continued to be time- and resource-intensive. Traditionally, this course design and implementation process hinged upon the role of the OLI Learning Engineer, whose task was to work closely with faculty, domain experts, and a larger course development team (potentially including learning scientists, instructional designers, assessment specialists, technologists, and other experts, as appropriate) to collaboratively design the learning experience, and then implement that design in OLI's XML structure. In addition to the XML authoring requirements, the system's build, deploy, and publishing process required additional expertise with subversion control systems and Linux-based command line tools. Furthermore, the deployment process itself often created extended intervals between design and implementation and finally publishing the completed, rendered courseware. Beyond the challenge of finding sufficient numbers of learning engineers possessing the requisite talents in learning design, project management, and technology, the process also made ongoing editing and revision challenging, and created a barrier for many educators who were interested in participating more directly in the authoring and improvement process. These hurdles have limited OLI in its ability to fully engage in the reuse/revision/remix approaches that are such an essential part of Open Education; developing and expanding the number of participants who use OLI as a community-based research activity is a core part of the Initiative's mission (Thille, 2012), and these barriers to authoring courses have slowed participation in the project by potential authors, hindering this part of the mission.

To address these challenges, OLI has invested heavily in developing an accessible, WYSIWYG authoring platform. This set of tools provides a better architecture to scaffold the design and development process, walking course developers and faculty through the process of articulating student-centered, measurable learning objectives and sub-skills; developing aligned practice and assessment opportunities with targeted hints and feedback; authoring expository learning elements; tagging course elements with the knowledge components represented by learning objectives and skills; and organizing these elements into a structured, coherent learning experience. The authoring tool is publicly available at <http://echo.oli.cmu.edu>. Beyond engaging with a larger community of authors and educators, the tool should also serve the learning analytics community by streamlining the process of developing well-instrumented learning experiences. Current development efforts are focused on expanding the design and improvement analytics that are embedded in the authoring tool, and on developing more thoughtful scaffolding for the authoring and developing process.

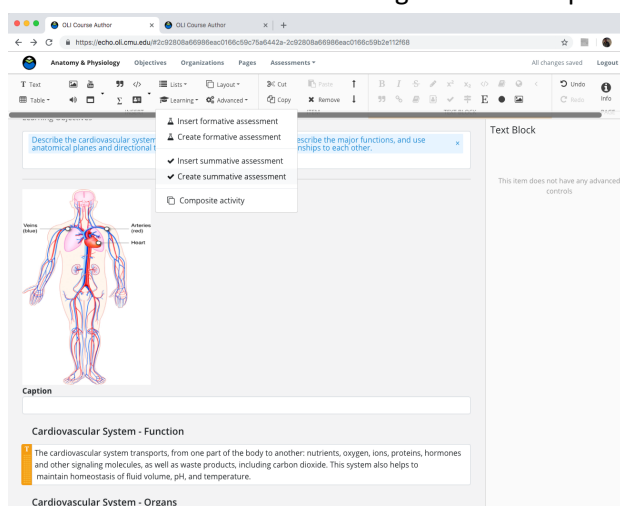


Figure 3: OLI Web-Based Course Authoring Platform

REFERENCES

- Aleven, V., Sewall, J., Popescu, O., Ringenberg, M., van Velsen, M. and Demi, S., Embedding Intelligent Tutoring Systems in MOOCs and e-Learning Platforms, In Proceedings of the Thirteenth International Conference on Intelligent Tutoring Systems (ITS) (2016)
- Bier, N. and Jerome, W. (2012). Learning Data Visualization. Open Education Annual Conference. Vancouver, Canada, October 17, 2012.
- Bowen, W. G., Chingos, M. M., Lack, K. A. and Nygren, T. I. (2014), Interactive Learning Online at Public Universities: Evidence from a Six-Campus Randomized Trial. *J. Pol. Anal. Manage.*, 33: 94111. doi: 10.1002/pam.21728
- Blink M.J., Stamper J.C., Carmichael T. (2014) SCALE: Student Centered Adaptive Learning Engine. In: Trausan-Matu S., Boyer K.E., Crosby M., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science, vol 8474. Springer, Cham
- Griffiths, R., Chingos, M., Mulhern, C., & Spies, R.. (2014). Interactive Online Learning on Campus: Testing MOOCs and Other Platforms in Hybrid Formats in the University System of Maryland. New York: ITHAKA S+R. Available at: <http://sr.ithaka.org/research-publications/Interactive-Online-Learning-on-Campus>
- Herckis, Lauren and Smith, Joel. (2018). Understanding and Overcoming Institutional Roadblocks to the Adoption and Use of Technology-Enhanced Learning Resources in Higher Education. Report to the Carnegie Corporation of New York. <https://www.cmu.edu/simon/news/docs/ccny-report.pdf> (Accessed 11 December, 2018)
- Kaufman, J., Ryan, R., Thille, C., & Bier, N. (2013). Open Learning Initiative Courses in Community Colleges: Evidence on Use and Effectiveness. Carnegie Mellon University, Pittsburgh, PA. Available online: http://www.hewlett.org/sites/default/files/CCOLI_Report_Final_1.pdf
- Kernohan, David and Thomas, Amber. (2018). *OER: A Historical Perspective* (Jisc, 2012): <http://repository.jisc.ac.uk/4915/> (accessed 27 August, 2018)
- Koedinger, K. R., Corbett, A. T., & Perfetti, C.(2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36, 757–798.
- Koedinger, K., Kim, J., Zhuxin, J., McLaughlin, E., and Bier, N. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC, *Proceedings of the ACM Conference on Learning @ Scale*.
- Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016). Is the Doer Effect a Causal Relationship? How Can We Tell and Why It's Important. In *Proceedings of the Sixth International Conference on Learning, Analytics and Knowledge*, pp.388-397. Edinburgh, UK.
- Leinhardt, G., Cuadros, J., & Yaron, D. (2007). "One Firm Spot": The Role of Homework as Lever in Acquiring Conceptual and Performance Competence in College Chemistry. *Journal of Chemical Education*, 84(6), 1047.
- Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*.
- Lovett, M. (2012, July). Cognitively informed analytics to improve teaching and learning. Presentation at EDUCAUSE Sprint, July 25, 2012: <http://www.educause.edu/ir/library/powerpoint/ESPNT122Lovett.ppt>

- Ritchie, Ann Lyon. (2018). Educators Return to LearnLab Summer School. <https://www.cmu.edu/simon/news/stories/2018-learnlab.html> (accessed 27 August, 2018)
- Ryan, S., Kaufman, J.H., Greenhouse, J.B., She, R., & Shi, J. (2016). The effectiveness of blended online learning courses at the community college level. *Community College Journal of Research and Practice*, 40 (1) pp. 285-289
- Shestak, Elizabeth. (2017). Simon Summer Institute Helps Pittsburgh-area Professors Build Ed Tech Tools for Their Classrooms. <https://www.cmu.edu/simon/news/stories/2017-pche-summer-institute.html> (accessed 27 August, 2018)
- Stamper, J., Koedinger, K.R. (2011). Human-machine Student Model Discovery and Improvement Using DataShop. In Kay, J., Bull, S. and Biswas, G. eds. *Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011)*. pp. 353-360. Berlin Germany:Springer.
- Thille, C. & Smith, J. (2011) Cold Rolled Steel and Knowledge: What Can Higher Education Learn About Productivity?, *Change: The Magazine of Higher Learning*, 43:2, 21-27, DOI: [10.1080/00091383.2011.556988](https://doi.org/10.1080/00091383.2011.556988)
- Thille, C. (2012). *Changing the Production Function of Higher Education*. Part of the Series: Making Productivity Real: Essential Readings for Campus Leaders. Washington, DC: American Council on Education.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models* (pp. 43-74). Springer New York.

LearnSphere: Learning Analytics Development and Sharing Made Simple

Kenneth R. Koedinger, John Stamper, Paulo Carvalho

Carnegie Mellon University

koedinger@cmu.edu, jstamper@cs.cmu.edu, pcarvalh@andrew.cmu.edu

ABSTRACT: LearnSphere.org provides a web-based learning analytics authoring environment where non-programmers can build, share, and modify novel combinations of a rich and growing set of methods. Methods for data import, transformation, statistical analysis, machine learning inference, and visualization and reporting can be combined in novel workflows. These workflows are linked to data and both workflow analytics and data can be shared and modified. A wide variety of workflows exist corresponding with techniques used a wide variety of published analytics.

Keywords: Learning Analytics, Analytical Workflows, Learning Data Sharing.

REFERENCES

- Koedinger, K. R., Stamper, J. C., Leber, B., & Skogsholm, A. (2013). LearnLab's DataShop: A data repository and analytics tool set for cognitive science. *Topics in cognitive science*, 5(3), 668-669. <https://doi.org/10.1111/tops.12035>

AskOski: Using University Enrollment Data to Surface Novel Semantic Structure and Personalized Guidance

Zachary Pardos

University of California Berkeley
pardos@berkeley.edu

ABSTRACT: This paper presents ways in which the synthesis of data from higher-ed can illuminate the terrain of the university and support students in their decision making and wayfinding. A novel application of recurrent neural networks and skip-grams, techniques popularized by their application to modeling language, are brought to bear on millions of historic student course enrollments to create vector representations of these objects. Analysis of the produced vector space reveals predictive information about students' on-time graduation and a high degree of emergent semantic relational information about courses which can be visualized, reasoned about, and surfaced to students. A course information platform, adopted by the UC Berkeley Office of the Registrar, uses this automatically inferred semantic information to help students navigate the university's offerings and provides personalized course suggestions based on topic preference, course history, and program requirements. Considerations for scaling such a system across the system will be discussed, as well as its place in the multi-stakeholder environment of the university.

Keywords: Recommender systems, Distributed representation, Recurrent neural networks, Skip-gram Scrutability, Usability study, Higher education.

REFERENCES

Pardos, Z.A., Fan, Z., Jiang, W. (2019) Connectionist Recommendation in the Wild: On the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*. <https://doi.org/10.1007/s11257-019-09218-7>

Enhancing Test Preparation via Continuous Tracking of Practice Assessment Analytics & Personalized Resource Recommendations

Stephen Polyak, Michael Yudelson, Kurt Peterschmidt,
Benjamin Deonovic, Alina von Davier

ACTNext by ACT, Inc.

{steve.polyak, michael.yudelson, kurt.peterschmidt,
benjamin.deonovic, alina.vondavier}@act.org

ABSTRACT: Learners preparing to take summative, high-stakes assessments such as The ACT College Readiness Assessment will typically use resources to review the knowledge and skills that are associated with the requisite academic subjects. Given the broad scope of these subject domains, learners would benefit by receiving targeted, personalized lists of recommended resources that align with their individually diagnosed area needs. In our work, we have created a Recommendations and Diagnostics (RAD) API that can be plugged into a learning and assessment system to continuously track a learner's practice assessment analytics and translate that into predictions of skill mastery. Using these predictions, we drive a recommendation engine that prioritizes areas of need based on ACT's Holistic Framework and delivers sets of tagged open educational resources for learners to review. We discuss our hierarchical model that is based on LLTM and uses Elo ratings. Also, we discuss the role of industry standards such as IMS Global Caliper and the Competency & Academics Standards Exchange (CASE) as part of our initial integration into ACT's free test preparation solution.

Keywords: Assessment Analytics, Resource Recommendation, Diagnostics.

1 INTRODUCTION

Our research involves the use of advanced psychometrics, machine learning techniques and algorithmic development based on the application of artificial intelligence in the education/learning space. Our development team is focused on well-defined, robust, modular, mobile and web-based solutions that can bring our research to products and services at scale. In this article, we briefly overview an initial solution aimed at bridging the link between measurement and learning that can recommend a set of open educational resources (OERs) for a specified learner based on a continuously updated diagnostic model that uses quiz/test item responses to measure mastery of learning objectives. This solution is packaged as an application programming interface (API) known as the Recommendations and Diagnostic (RAD) API. We then present our hierarchical skills-based variation on the use of Elo ratings. We discuss several ideas we are evaluating to track metrics relating to the use of RAD in a free test preparation solution. We also detail the role industry standards such as IMS Global Caliper and the Competency Academics Standards Exchange (CASE) play in our solution. We also present our initial plans for establishing metrics to evaluate RAD in the context of its initial integration for test preparation and skills practice/review.

2 RAD IN TEST PREPARATION

ACT offers a free learning and assessment platform, ACT Academy, that helps learners review the skills that are assessed on the ACT college readiness assessment. It consists of short (5-10 item) practice quizzes as well as full length practice tests that users can select in a self-directed manner. Our RAD API can enhance systems like ACT's test preparation by processing the assessment response data, maintaining a diagnostic perspective on the learner's evidence and using that data to then generate personalized lists of instructional content. The RAD life cycle is illustrated in Figure 1.

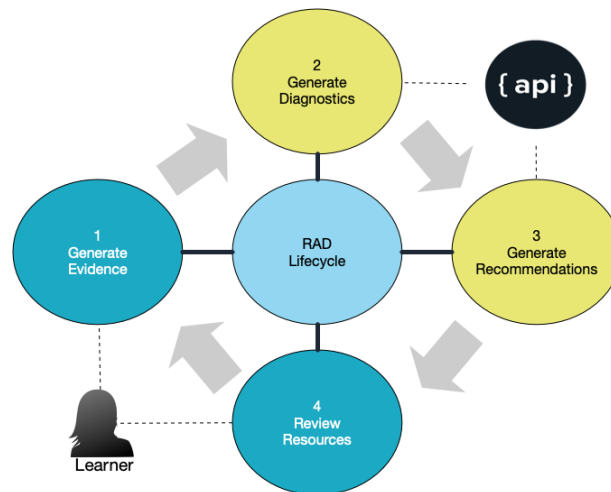


Figure 1: RAD Learner/API Lifecycle Steps

As users present evidence of learning in a learning and assessment system (1) (e.g. scored item responses), RAD inspects the item metadata and assesses which skills were required to correctly answer the item. RAD retrieves information on how difficult the skills for that item are based on population sampling over the items/skills. Using this knowledge and prior estimates of the learner's ability on those skills, RAD API combines this data using an Elo-based algorithm to update diagnostic records (2) that can predict learner mastery of skills in a continuous, real-time fashion. These diagnostic records are then used when learners request instructional resources in some area of English, math, reading or science (e.g. Geometry). RAD uses its hierarchical knowledge of the subject domain to inspect the category of knowledge and evaluates which skills/ skill areas would be the most helpful for the learner to review. These recommendations draw on the OpenEd catalog of instructional content to deliver personalized recommendations (3). After learners interact with the learning resources (4) they continue the lifecycle by continuing their progress with more test preparation and practice with ACT Academy quiz/test items.

RAD's current Elo-based algorithm was evaluated against a range of alternatives preceding the lead up to its initial operational use. In September 2018, the RAD API was integrated into ACT Academy to power its recommendations and diagnostics. This integration is illustrated in Figure 2. The figure shows that ACT Academy utilizes a custom Open Source TAO test delivery engine that has been enhanced to produce IMS Global Caliper event data (AssessmentEvent, AssessmentItemEvent).

This data is sent from Academy's TAO platform to the ACT Learning Analytics (LEAP) platform. LEAP is then responsible for forwarding the data on to RAD for diagnostic processing in real-time. As

mentioned earlier, RAD integrates with the OpenEd catalog to locate resources for recommendations. This is done in a learning object repository (LOR) independent manner by using the IMS Global LTI Resource Search API.

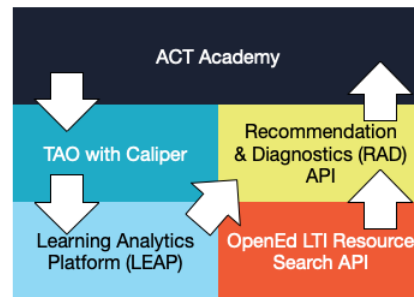


Figure 2: RAD Learner/API Lifecycle Steps

Since the initial integration of RAD API in September 2018 and as of December 4, 2018, the RAD lifecycle has processed: 309,216 assessments consumed by 102,394 learners and made 363,657 recommendations.

3 ELO RATING SCHEMA

Elo, named after its inventor Arpad Elo, is, technically, not a model, but a rating system that tracks rating values of two classes of variables for the modeled events (Elo, 1978). In chess, where Elo found first use, the events are chess matches and the variables are opponent 1 ability and opponent 2 ability. After each match, the ratings of opponent abilities are updated based on the outcome (a win of either opponent or a draw).

Elo has been used in educational domain too. Here, an event is student's opportunity to answer a question item or a problem step correctly. Student is opponent 1, and item is opponent 2. Sometimes, a set of skills relevant to the question item are used to represent opponent 2. Also, student abilities could be represented hierarchically through a set of student-skill abilities together with an overall ability. For an extended discussion see an overview paper by (Gřihák et al., 2015).

Although not a model per se, Elo has a few desired properties that are relevant to us. First, Elo predominantly uses local updates of the tracked values. Second, it requires minimal fitting or tuning. And third, student success or failure always result in a respective increment or decrement of their tracked ratings. We tried several versions of Elo rating schemata that build upon LLTM model. See, for example (von Davier et al., 2019) where such model is discussed. Below, we describe several variants of Elo, including the hierarchical based on LLTM.

3.1 Simple Student-Item Elo

The simplest case of an Elo model is akin to the 1PL IRT (Rasch) model (Rasch, 1960) that is frequently used in psychometrics. In both cases, there is a notion of student's uni-dimensional ability θ_i and difficulty of a question item β_j . A probability of student i answering item j correctly is computed as shown in Equations 1a and 1b.

$$p_{ij} = \sigma(m_{ij}) \quad (1a)$$

$$m_{ij} = \theta_i - \beta_j \quad (1b)$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (1c)$$

The difference between Rasch model and simple student-item Elo is in how values of interest (student ability and item difficulty) are computed. Rasch model is fit as a mixed-effect logistic regression, while Elo updates values as student/item data is exposed to it.

3.2 Student-Skill Elo

Instead of tracking item difficulties in the Simple Student-Item Elo, one might want to replace them with skill difficulties if skill labels are available for all question items. This could be done for several reasons. One, if the data is coming from a system that has longer student exposure to skills, skills could be used as

units of transfer to better track learning rather than using items that students are often interact with once or twice. Two, if the item pool is heterogeneous, less reliable, or extremely large and it is less efficient to track item properties. Equation 2 shows how to compute probability of student i answering item j correctly, when skill difficulties are used. Variable q_{ik} is an element of a Q-matrix -- a matrix of 1's and 0's, where a value of 1 means that a skill is relevant to a question item.

$$m_{ij} = \theta_i - \sum_k q_{ik} \cdot \beta_k \quad (2)$$

3.3 Hierarchical Elo

The-so-called hierarchical Elo tracks student abilities at two levels -- overall, and per-skill. Namely, there is a global student θ_i , as well as θ_{ik} values one per each student-skill tuple. Skill difficulties are also retained in this model. The form of the probability of correctness for the hierarchical Elo is given in Equation 3.

In general, a wide range of factors could be included into an Elo model. A rule of thumb is that student-level factors are included with a positive sign, and the factors of the environment that the student overcomes or competes against are included with a negative sign.

$$m_{ij} = \theta_i + \sum_k q_{ik} \cdot \theta_{ik} - \sum_k q_{ik} \cdot \beta_k \quad (3)$$

3.4 Updates to Elo-tracked Values

All of the values tracked by Elo are maintained on the log-odds scale and require a sigmoid transformation (rf. Equation 1c) to be converted to the probability scale. Initial values of all parameters are customary to be set to 0, before Elo has *seen* any data pertaining to those parameters. When new data on results of students answering items arrives, special rules are used to update tracked values. Equations 4a and 4b show an example of an update rules for student ability and item difficulty. Here, K is a sensitivity parameter which is, in this case, constant. Also, c_{ij} denotes actual

correctness of student's response (a value of 0 or 1), and p_{ij} is the prior estimate of the probability of correctness as it was defined in Equation 1a.

$$\theta_i = \theta_i + K \cdot (c_{ij} - p_{ij}) \quad (4a)$$

$$\beta_j = \beta_j + K \cdot (c_{ij} - p_{ij}) \quad (4b)$$

One could see that the difference between updating student and item parameters is the sign in front of the actual/expected value difference. When more student-level and environment-level parameters are used, for example student-skill ability Elo and skill difficulty respectively, the sign is set in a similar manner. In Equations 4a and 4b a single sensitivity K was used. One could use separate sensitivities for updating tracked parameters for students and items. There are also other ways to define sensitivity. An example of an alternative definition we used in our work is given in Equation 5. Here, K is redefined as a ratio, where the denominator – n_i – is a number of prior data points used to re-estimate student ability θ_i and a and b are parameters. Just like in the case with a single sensitivity K , parameters a and b could be universal for all classes of values tracked by Elo or be specific for student, item, or skill parameter values. There are other approaches to defining sensitivity K that could be found, for example, in (Gřihák et al., 2015).

$$\theta_i = \theta_i + \frac{a}{1+b \cdot n_i} \cdot (c_{ij} - p_{ij}) \quad (5)$$

4 METRICS

Our approach to selecting and evaluating RAD metrics is being applied to three main perspectives: predictive accuracy, platform usage, and student-skill rating distributions.

4.1 Prediction Performance

The current RAD API algorithm prediction is a straight-forward machine learning performance metrics question: we are interested in how well does the RAD API predict the correctness of learners' answers to question items.

4.1.1 RAD Classification Accuracy

One metric is classification accuracy for the total set or some subset of the RAD data. Classification accuracy is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ Made}$$

For this, we will need to select a mastery level threshold probability value.

4.1.2 RAD Confusion Matrix

A RAD Confusion Matrix will describe the complete performance of the algorithm using the following terms:

- True Positives : The cases in which RAD predicted the learner will get an item correct and the actual output was also correct.

- True Negatives : The cases in which RAD predicted the learner may miss and the actual output was incorrect.
- False Positives : The cases in which RAD predicted the learner will get the item correct and the actual output was incorrect.
- False Negatives : The cases in which RAD predicted incorrect and the actual output was correct.

		Prediction outcome		total
		p	n	
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

Confusion matrix accuracy can be calculated by taking the average of the values lying across the *main diagonal*, ie.

$$CM - Accuracy = \frac{True\ Positives + False\ Negatives}{Total\ Number\ of\ Samples}$$

4.1.3 RAD ROC

A receiver operating characteristic curve (ROC) for RAD is a plot that shows the diagnostic ability of RAD's current algorithm as its discrimination threshold is changed.

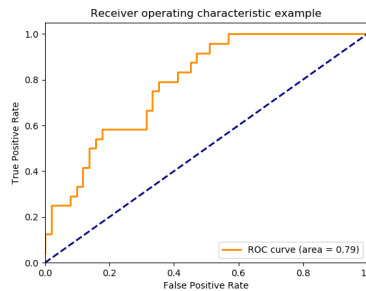


Figure 2: Sample Receiver Operating Characteristic curve

This involves:

- True Positive Rate (Sensitivity) : True Positive Rate is defined as TP/ (FN+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- False Positive Rate (Specificity) : False Positive Rate is defined as $FP / (FP+TN)$. False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

4.1.4 Table: Diagnostic Prediction Outcomes

In order to facilitate easier calculation of prediction performance metrics, we are proposing to add a new table that will be updated as RAD is continuously processing learner evidence. This is similar to Carnegie Learning's Cognitive Tutor *tutor event* data source.

User Id	Item Id	Timestamp	Prediction	Outcome
user1	item1	ts1	.72	1
user2	item2	ts2	.43	0

4.2 Platform Usage

We have identified several metrics that we are looking to gather based on ACT Academy data. This can help us ask/answer platform usage questions like:

What impact, if any, has RAD had on platform usage patterns in the host ACT Academy platform?

Data to support this can include:

- number of media resources consumed (per user, overall?)
- total amount of time spent on resources
- total amount of time spent on quizzes/tests
- number of unique items/user
- percent correct items answered
- number of days active
- diversity of resources shown (per user, per subject, etc.)

The last one in this list may in fact be one of the most important ones. Prior to RAD, Academy only had a tiny, fixed set of resources and only varied the number of those resources shown to a user based on a ratio correct diagnostic. RAD now draws on the much larger catalog of OpenEd resources that should be generating more diverse resource recommendations.

4.3 Ratings Distributions

RAD delivers diagnostic mastery estimates to ACT Academy as hierarchical probability values. Ultimately ACT Academy translates that into subject/category star ratings using cut points for stars. Currently this is defined as:

- One star = .1 to .59
- Two stars = .60 to .79
- Three stars \geq .80

Given this we should be able to answer questions such as:

What is the distribution of stars (per subject/area) for Academy learners given RAD's current collection of diagnostic records?

5 IMS STANDARDS

IMS Global is a non-profit entity that produces a set of standards to help advance interoperability within the educational focused software community. When fully embraced, it facilitates plug-and-play architectures and provides a foundation on which disparate systems can work together seamlessly. ACT is an active participant within this community, both contributing financially and with member's time on various committees.

5.1 CASE

The Competencies and Academic Standards Exchange (CASE) specification defines how systems exchange and manage information about learning standards and/or competencies in a consistent and machine-readable format. OpenSALT is a platform that supports this standard and houses the Holistic Framework, providing it to align with diagnostics in the RAD API, items in Academy, and resources in OpenEd.

5.2 Caliper Analytics

Caliper Analytics defines a standard to collect learning data from digital platforms. This is used to instrument, act upon, or visualize learning activity. It is this format that student response data from items presented in Academy is delivered to the RAD API. The sensor event type falls under the Assessment Profile defined within the Caliper standard. Other profiles that the standard provides are:

- Annotation Profile
- Assessment Profile
- Assignable Profile
- Forum Profile
- Grading Profile
- Media Profile
- Reading Profile
- Session Profile
- Tool Use Profile
- Basic Profile

All of these profiles are intended to cover the rich click stream data exhaust out of learning and assessment systems.

5.3 LTI Resource Search

Learning Tools Interoperability (LTI) Resource Search defines a standard by which digital repositories can be searched for a set of resources. It addresses searching Learning Object Repositories (LORs) and other catalogs of learning resources. Rich metadata about the resources can be obtained as well as URLs or LTI links that can be used to launch within another platform. RAD API uses this mechanism to find personalized, aligned content based on a particular diagnostic record.

6 SUMMARY

In this paper we outlined our initial work aimed at enhancing test preparation activity via continuous tracking of practice assessment analytics and providing personalized resource recommendations. This work has produced a reusable API that can support a range of learning and assessment systems. We presented our rating system that was selected from a variety of alternatives and configurations. We detailed the way that the system integrates with a hierarchical set of standards. An approach towards analyzing this work was described by means of various categories of metrics and we anticipate moving forward with this work. Our work is also supported by several industry standards which we described that will help provide portability of the solution beyond its initial implementation and scope. Looking ahead, we anticipate that we will be able to present more on the evolution of RAD both within its current use in test preparation and to additional learning scenarios. We are planning to enhance the RAD API with features that will provide an administrative dashboard and visualizations of the diagnostic record data over time.

Despite our initial focus on supporting test preparation, RAD API is capable of supporting a full range of diagnostic and recommendation needs of Learning and Assessment Systems (LAS) (Arieli-Attali et al., 2019). These systems bridge the gap between computerized testing and learning fields that, until very recently, developed in relative isolation. Our goal is to develop RAD API into a solution that helps adaptive educational systems become ubiquitous student support tools from K to Career and from summative and formative evaluation to learning.

REFERENCES

- A. E. Elo. The rating of chessplayers, past and present. Arco Pub, 1978
- Balakrishnan, R. (2006, March). *Why aren't we using 3D user interfaces, and will we ever?* Paper presented at the IEEE Symposium on 3D User Interfaces. <http://dx.doi.org/10.1109/vr.2006.148>
- Gřihák, J, Pelánek, R., Niznan, J,. Student models for prior knowledge estimation. In International Conference on Educational Data Mining (EDM 2015), pages 109–116, 2015
- Von Davier, A., Deonovic, B., Polyak, S.T., Woo, A. (2019). Computational Psychometrics Approach to Holistic Learning and Assessment Systems. *Frontiers in Psychology* (in press).
- G. Rasch. Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks Paedagogiske Institut, 1960.
- Arieli-Attali, M., Ward, A., Thomas, J., Deonovic, B., Von Davier, A. (2019). The Expanded Evidence-Centered-Design (e-ECD) for Learning and Assessment Systems: A Framework to Incorporating Learning Goals and Processes within Assessment Design. *Frontiers in Psychology* (in press).

Conceptual Change as Evidence of Learning

Steven Ritter, Stephen E. Fancsali, Michael Sandbothe, Robert G.M. Hausmann

Carnegie Learning, Inc.

{sritter, sfancsali, msandbothe, bhausmann}@carnegielearning.com

ABSTRACT: Extending prior work on knowledge component modeling via segmented learning curves, we consider properties of such learning curves that seem to indicate how heterogeneous populations of students learn over time. Pointing to the possibility that different cognitive models may be appropriate for different student populations, we use a concrete example and evidence from an initial pilot study to illustrate how conceptual change may provide evidence for learning.

Keywords: cognitive modeling, learning curves, mathematics education, knowledge components, skill modeling

1 INTRODUCTION

Knowledge component (KC) modeling has a long history within adaptive learning systems (e.g., Anderson, Conrad & Corbett, 1989). KCs, in such systems, are the basic elements of learning. As students work through (often) complex, multi-step problems, their actions (correct, incorrect, or requests for assistance) provide evidence of the student's knowledge of each KC. The set of KCs for a domain constitutes the definition of that domain's learning objectives, and these systems determine mastery of the domain based on the student's ability to master each of the components in that domain. Much recent work has focused on deriving the learning objectives for a domain, as well as the refinement of existing KC models, based on data (Cen, Koedinger and Junker, 2006; Koedinger, McLaughlin and Stamper, 2012).

These analyses treat KCs as parameters in a model of student learning and test to see whether learning is better modeled when components are split or merged. For example, Long, Holstein, and Aleven (2018) found better fits to data when they split the KC involved in removing a constant when solving an equation. Instead of a single skill that covers isolating the variable in equations like $x + 5 = 10$ and $x - 5 = 10$, their model includes two skills, one for subtracting when the constant is positive and one for adding when the constant is negative. The best-fitting model is often surprising because it violates expert intuitions about how students are likely to think about problem solving. Experts (or even those with a bit of algebra training), might focus on isolating the variable and think of adding or subtracting to do so as a relatively trivial distinction, but the data show that such distinctions may not be at all trivial to the student.

The theoretical basis for KC analysis and modeling (e.g., Anderson, 1993; Anderson, 2002) acknowledges that the KC model is always relative to the student's level of understanding. From this perspective, the reason that the KC model accurately explains student learning is that the KCs represent the actual cognitive steps that people use to solve the problem. But, as students learn, the steps they take can (and should) change. Although students learning equation solving may

distinguish between positive and negative constants, we hope that they will progress to the point where they see the operation of removing the constant as a single cognitive step, regardless of the sign of that constant. In other words, we expect a model of beginners to include two KCs, but a model of experts should include only one. In fact, an assessment of whether the student has merged these skills may act as an assessment of that student's conceptual knowledge. As Carnegie Learning has applied KC modeling to larger and more heterogeneous groups of students than have been used in prior work, we believe that the issue of detecting conceptual change of this kind has become more essential. In what follows, we present data from large and heterogeneous groups of students demonstrating these phenomena, explore the possibility of reliably detecting that students are at different conceptual levels of understanding, and discuss whether these levels may be reliable markers of student ability.

2 PRELIMINARIES

2.1 MATHia

MATHia is an intelligent tutoring system for mathematics, based on Carnegie Learning's Cognitive Tutor (Ritter, Anderson, Koedinger & Corbett, 2007). In MATHia, students work through curricula comprised of workspaces, each of which provides practice on fine-grained KCs via complex, multi-step problems (Figure 1). MATHia implements mastery learning (Bloom, 1968); in each workspace, students work through a series of problems selected so as to emphasize KCs that a student has yet to master. MATHia uses the probabilistic Bayesian Knowledge Tracing (BKT) framework (Corbett & Anderson, 1995) to estimate student knowledge of KCs based on their performance on individual problem-solving steps mapped to KCs. Once MATHia's probability estimate of a student's knowledge of a KC reaches 95%, mastery of that KC is achieved. When students are estimated to have achieved mastery of all KCs associated with a workspace, they "graduate" to the next topic. If students reach a set number of maximum problems within a workspace without having achieved mastery of all associated KCs, the software allows the student to advance to the next topic without mastery. In such cases, the teacher is notified that the student will require additional instruction in the unmastered topic.

2.2 Segmented Learning Curves

Extensive educational data science literature is devoted to the analysis and improvement of KC models based on empirical data (e.g., Junker, Koedinger & Trottini, 2000; Cen, Koedinger & Junker, 2005; Koedinger, McLaughlin & Stamper, 2012; Goldin, Pavlik and Ritter, 2016). Much of this literature bases improvements to KC models on the analysis of learning curves (e.g., Anderson, Conrad, & Corbett, 1989; Koedinger et al., 2011; Martin et al., 2011; Murray et al., 2013). These analyses are predicated on the idea that if KC modeling (the underlying basis for the approach of MATHia and Cognitive Tutor) correctly describes performance of complex tasks, then performance data on these KCs ought to follow a learning curve (Anderson & Lebiere, 1998; Anderson, 2002; Ritter, Anderson, Koedinger & Corbett, 2007). That is, with increased opportunities to practice a KC, student performance increases, and correctness (or error rates) plotted against opportunities to practice a particular KC ought to monotonically increase (or decrease, in the case of error rates) over increased opportunities. Deviations from "smooth," monotonically increasing learning curve indicate that the model mapping problem-solving steps to KCs may require improvement. Learning curves

are typically thought to follow a power law (Newell & Rosenbloom, 1981; Anderson, 2001), though this point is not without controversy (Heathcote & Brown, 2000).

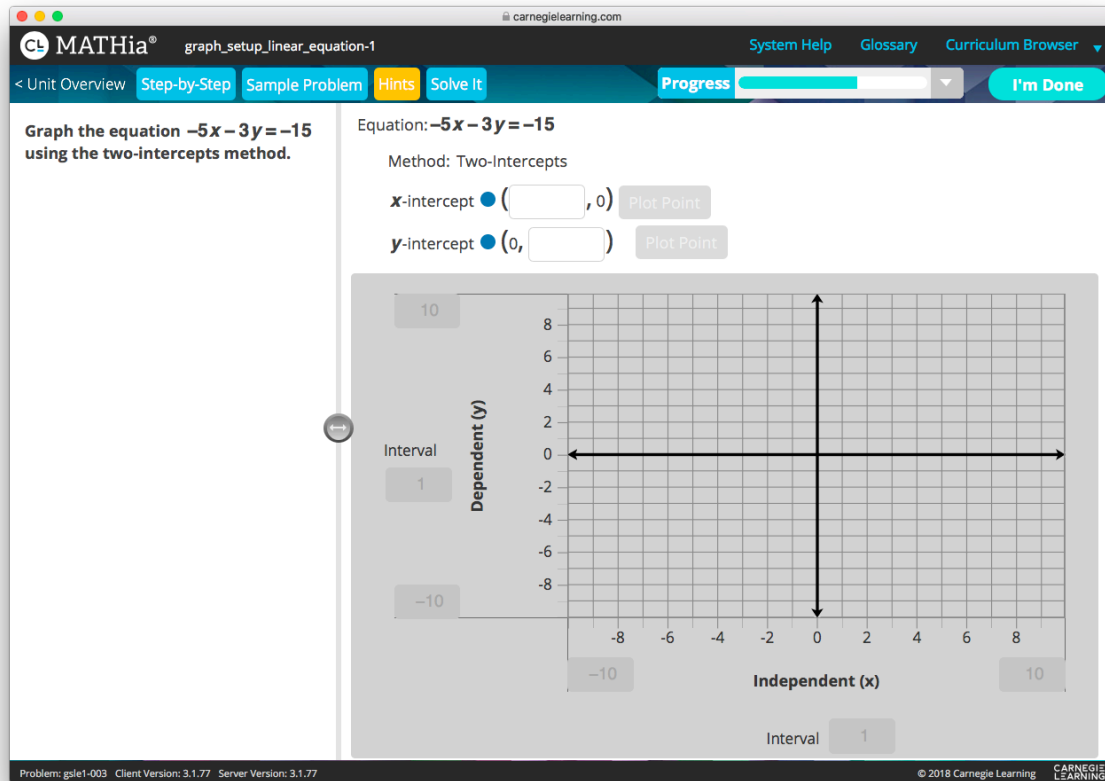


Figure 1: Screenshot of problem solving in MATHia from the workspace “Graphing Linear Equations Using a Given Method”

The present approach to modeling conceptual change builds on work due to Murray et al. (2013), wherein they presented learning curves “segmented” by student mastery. Rather than plot a single curve for a KC over an entire dataset, this approach calls for plotting multiple curves based on grouping students according to the problem-solving opportunity at which a student reached mastery, as judged by the learning system (e.g., one curve for all students who mastered on opportunities 3, another for all students who mastered on opportunity 4, etc.). Especially with increasingly large datasets from heterogeneous student populations, aggregate learning curves (i.e., those that plot all students in a single curve) may take on particular shapes that may indicate that a KC is not well-modeled (i.e., a “flat” learning curve that may indicate that students are not learning or mastering a KC over time) due to various effects unrelated to student learning, but rather due to student attrition in a system based on mastery learning (Murray et al., 2013; Nixon, Fancsali & Ritter, 2013).¹ We adopt a similar approach but “bin” or categorize students over ranges of opportunities at

¹ Notably, Murray et al. (2013) are not the first authors to point out that learning curves plotted over all students need not take on the same form as learning curves plotted for subsets of learners (Newell & Rosenbloom, 1981; Heathcote & Brown,

which mastery is achieved in our example (i.e., one curve for all students who mastered a KC at opportunities 1-5, another for those that mastered on opportunities 5-10, etc.).

Murray et al. (2013) focused on demonstrating that aggregate learning curves can be misleading with respect to whether or not learning had occurred within intelligent tutoring systems that implement mastery learning like MATHia; that is, the shape of an aggregate learning curve need not provide sufficient evidence that a KC is a target for cognitive/skill model refinement. In what follows, we consider properties of segmented learning curves that seem to indicate how heterogeneous populations of students learn over time, pointing to the possibility that different cognitive models may be appropriate for different student populations and illustrating how conceptual change, illustrated via segmented learning curves, may provide evidence for learning. In addition, the segmented learning curve may point to possibilities for improvements to cognitive models targeted to a particular sub-population of students.

3 AN EXAMPLE: CALCULATING INTERCEPTS USING GENERAL LINEAR FORM

3.1 The Problem Solving Context

Consider again Figure 1, which provides a problem-solving context instance within which we see the KC that is the target of the present work. In this workspace, students plot a linear function according to one of three methods given to the student: slope-intercept, two points, and two intercepts. These plotting methods differ between problems, and problem types are interleaved. Figure 1 shows a “two intercepts” problem in which students are asked to plot the linear equation $-5x - 3y = -15$ using the “two-intercepts” method. To carry out this method, the student provides the x-intercept and y-intercept for the given linear equation and then plots both points on a graph as a means by which to arrive at the line represented by that equation. The skill “Calculate intercept using general linear form,” as coded in MATHia’s cognitive model, applies to entering both the x-intercept and y-intercept; that is, students have two opportunities to demonstrate their knowledge or mastery of this KC within problems like this within this workspace.

Problems asking students to plot using two points ask the student to enter both x and y values for each point’s coordinates. Notably, students might, in fact, transform a “two points” problem into a “two intercepts” problem by choosing their two points as the intercepts. In fact, this strategy is recommended, since plugging in a zero for one of the variables usually leads to simple calculations. The KC on which we focus, “Calculate intercept using general linear form” only applies to two-intercepts problems. When students enter coordinates for two-points problems, their knowledge is tracked by a different KC, regardless of whether they choose to enter the intercepts as their two points.

Figure 2 provides the aggregate learning curve for “Calculate intercept using general linear form” over the population of 14,646 students who encountered this KC during the 2017-18 academic year

2000; Anderson, 2001). Zerr et al. (2018) also segment learning curves by performance to look for individual differences in heterogeneous populations.

while using Carnegie Learning’s MATHia, across a range of grade-levels and math courses in schools throughout the United States. On visual inspection, correctness rate for students over 26 opportunities to practice this skill never reaches 50% and is relatively “flat” with a modest “saw tooth” pattern of increases and decreases in the correctness rate. LearnSphere’s DataShop analytic toolkit (Koedinger et al., 2011) provides a learning curve categorization tool that, when applied to this learning curve, would categorize it as both “still high” (i.e., students’ last opportunities still exhibited an error rate about 40%) and as demonstrating “no learning” due to the lack of increase in the overall learning curve.² We now consider segmenting this learning curve³ by student mastery and consider the implications of this more nuanced visualization.

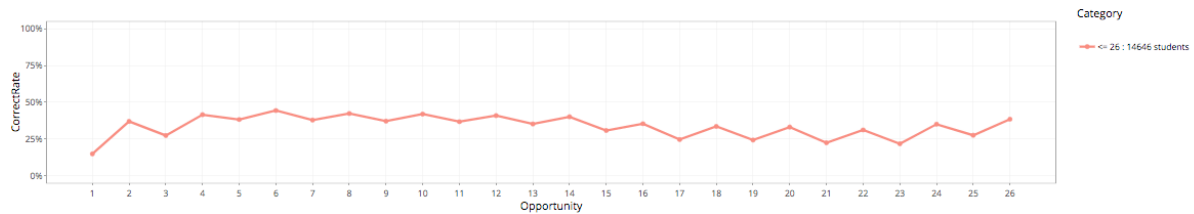


Figure 2: Aggregate learning curve for the KC “Calculate intercept using general linear form.”

3.2 Segmented Learning Curve

Figure 3 provides a segmented learning curve visualization for “Calculate intercept using general linear form” in this MATHia workspace. For ease of interpretation, we categorize students into groups according to the point at which mastery is achieved or the opportunity at which students reached the set maximum number of problems for this workspace (i.e., 26 problems) and were promoted. In addition to a relatively high correctness rate on the first opportunity (compared to other categories), we see a smooth, monotonically increasing learning curve for 4,364 students who achieved mastery within the first five opportunities to practice this skill (≤ 05 in the legend). Some other categories of students (i.e., those that master or got promoted⁴ in 6-10 opportunities and those that do so in 11-15 opportunities) had low correctness rates on the first opportunity but improved relatively steadily thereafter. In sum, upwards⁵ of 12,107 students (or approximately 83% of students) master the KC within 15 opportunities, but this fact does not seem obviously apparent from the aggregate learning curve.

² See “Learning curve categorization” at <https://pslcdatashop.web.cmu.edu/help?page=learningCurve>

³ For practical implementation of segmented learning curves within DataShop, see Fancsali, Sandbothe & Ritter (2019).

⁴ Due to nuances of problem selection within this MATHia workspace and the fact that not all KCs are practiced by every problem within a workspace, it is possible for students to be promoted from a workspace despite only seeing a relatively small number of opportunities for some KCs. Balancing the tradeoff of mixed practice of KCs with reasonable mastery learning criteria remains an on-going area of research and development. We are also exploring binning students within segmented learning curves by mastery vs. non-mastery status for a KC.

⁵ Approximately 1,400 students reached the maximum number of problems for the workspace within the first 15 opportunities at this particular KC. See footnote 4.

Of particular interest to us is the sawtooth pattern in these learning curves. The pattern is apparent but weak in the aggregate learning curve. When we look at the segmented curves, we can see that the sawtooth pattern is not at all evident in the learning curve for the 4,364 students in the highest learning curve. There is a slight sawtooth in the second highest group, representing a dip for the third opportunity and more extended sawtooth for students in the third-highest group. The three lowest groups, in contrast, show a sawtooth pattern that extends over nearly all opportunities to enter the intercept in two-intercepts problems.

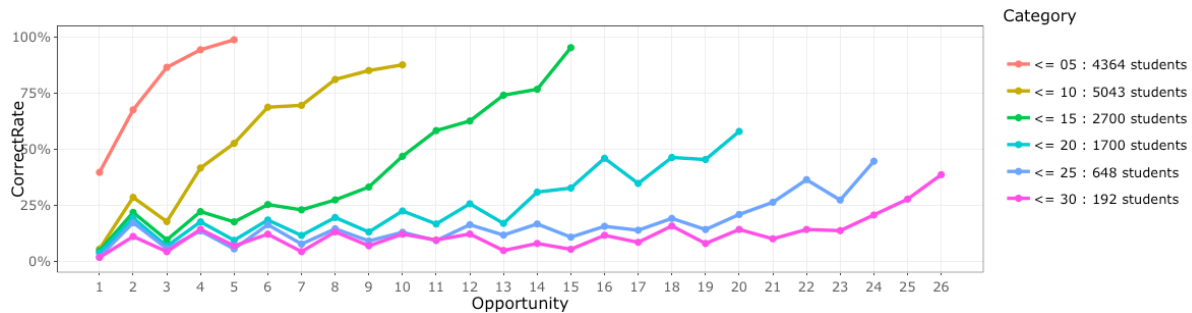


Figure 3: Segmented learning curve for the KC “Calculate intercept using general linear form.”

The sawtooth reflects the fact that the MATHia’s cognitive model considers calculating the x and y intercept to be two applications of the same KC. The first is for the x intercept, and the second is for the y intercept. All two-intercept problems have the layout shown in Figure 1, with the x -intercept above the y -intercept. Although students could compete the y -intercept first, the vast majority of students work top-to-bottom and complete the x -intercept before the y -intercept. The sawtooth pattern reflects the fact that students are somewhat more successful completing the y -intercept than they are in completing the x -intercept.

The sawtooth pattern is typically an indication that the cognitive model is not aligned with the actual KCs students are using to complete the task. In this case, the pattern indicates that calculating the x -intercept is, in some way, harder than calculating the y -intercept. If the cognitive steps required to calculate the x -intercept exactly coincided with the steps required to calculate the y -intercept, then we’d expect practice with calculating the x -intercept to transfer to calculating the y -intercept and thus see a monotonic increase in performance with experience (as, in fact, we see for the highest-performing segment). If, instead, calculating the y -intercept involves a completely different thought process than calculating the x -intercept, then we would see no transfer from one to the other. The extent of transfer from one step to the next depends on the extent to which they share KCs (Singley and Anderson, 1989; Koedinger, Yudelson and Pavlik, 2016).

Now consider a simple model of how students might find the x -intercept in the problem shown in Figure 1. They see the equation $-5x - 3y = 15$ and the term “ x -intercept.” They understand that x -intercept represents the point at which the line intersects the x -axis and recognize that y must be zero at that point. They can then substitute 0 for y (or a more expert student might directly understand that substituting 0 for y eliminates the y term) and recognize that the resulting equation is $-5x = 15$, which they can solve to find that $x = -3$. In a fully articulated cognitive model, each of these steps would correspond to one or more KCs. For practical reasons, cognitive tutors only track and represent a small subset of these steps. A student following this model understands what

“intercept” means and that the implication of a function intersecting an axis is that one of the coordinates is zero. For such a student, the process for calculating the y -intercept is much the same as that for calculating the x -intercept.

Suppose, instead, that the student only had shallow knowledge of the meaning of y -intercept. Such a student might only know the y -intercept as b in an equation of the form $y = mx + b$. Such a student could calculate the y -intercept by transforming the equation to slope-intercept form without knowing how to calculate the x -intercept. Alternatively, a student might recognize that the y -intercept is where $x = 0$ but not realize that x -intercept, a less-commonly encountered concept, is analogous. This kind of explanation is the most common for a sawtooth learning curve: two steps that are presumed to involve the same KCs turn out to crucially differ in one or more KCs.

A different kind of explanation for the sawtooth is that it reflects a repeated KC opportunity within a single problem, which typically leads to higher performance on the second opportunity (Martin et al., 2011). This explanation essentially says that, although the general KCs used to calculate the x and y -intercepts largely overlap, the second opportunity to apply those KCs within the problem is easier, because some aspects of the problem context are shared between the two applications of the KC. In this case, for example, the fact that the initial equation is common between the two steps may make the second step easier.

We have two classes of explanation for the sawtooth learning curve. One type of explanation says that students exhibiting that pattern see the x - and y -intercepts as being different kinds of things and so understanding how to calculate one does not help in calculating the other. A second type of explanation says that students see those concepts as related, but problem-specific knowledge carried over from the first application to the second makes the second more likely to succeed.

This second type of explanation seems less likely to us, since it does not explain why the sawtooth pattern is not evident in students who are performing well. The top segment of students does not show the pattern at all, and the next two segments only show a sawtooth pattern in early opportunities. This type of pattern is what you might expect if some students (those in the top segment) start with unified conception of intercept (that applies to both x and y) and others (those in the next two segments) quickly acquire such a conception. We see no clear explanation for why problem-specific elements should facilitate transfer for poorer students but not for better ones.

To better understand how students might be thinking about intercept, we ran a pilot study in which we interviewed 19 seventh-graders at a local school. Eight of these students were given an equation in the form $ax + by = c$ and asked to calculate the x - and y -intercepts (four were asked to calculate the x intercept first; 4 were asked to calculate the y -intercept first). All 19 students were also given an equation in $ax + by = c$ form and asked to solve for y , given $x = 5$. For students with a unified concept of intercept, we might see this second task as a subset of the first. In the first, the student needs to translate understanding of the term *intercept* to a value for x or y and then substitute. The second task tests the substitution step. None of the eight students were able to correctly determine either the x - or y -intercept. In contrast, five of these same eight students were able to correctly do the substitution problem, and 11 of the full group of 19 were able to do this correctly.

Figure 1 shows another possible way to succeed in this task. Students might simply look at the coordinate notation [e.g., “(__, 0)”) and realize that the task requires them to substitute zero for y and calculate x . This understanding of the task would not require them to understand the terms “ x -intercept” or “ y -intercept.” For this reason, we also asked our pilot students to calculate a missing coordinate in a pair where one coordinate was zero and the given equation was of the form $ax + by = c$, again varying whether they were asked for x or y first. On this task, performance was 26% correct: four students were able to calculate both x - and y -intercepts; two got one of the two correct, and 13 got neither correct.

These results support the idea that these students have reasonable symbolic calculation skills but weak conceptual knowledge of intercepts. Interviews with the full group of students indicate that a likely disconnect is between graphical and symbolic notions of x and y . Many students failed on the “substitute a coordinate” task, despite correctly interpreting the coordinate notation as meaning that either x or y was zero. At least two of the 19 students proceeded to sketch a graph to find the intercepts, rather than simply substituting zero into the equation. Such students appear to act as if x and y have a meaning related to locating a point on a graph (and cued by coordinate notation and by the term “ y -intercept”) and only a weakly linked understanding of x and y as variables in a symbolic equation.

4 DISCUSSION

The theory behind KC modeling predicts that different KCs will be applicable to students with different levels of knowledge. With small and relatively homogeneous groups of students, this theoretical possibility might be of little practical importance. However, as we start to look at larger datasets comprising more heterogeneous student groups, we should expect to start to see cases where different subsets of students are best modeled with different sets of KCs. We believe that the example we present in this paper is one such case. Some students in our dataset have a unified conception of intercept; others seem to quickly reach that point, and still others persist in treating x - and y -intercept as different (and ill-understood) concepts.

It is also notable that there is little improvement (in either the x - or y -intercept) among students who exhibit the sawtooth learning curve. This might be taken as an indication that conceptual understanding is key to performing the task. Few students appear to reach mastery on calculating the y -intercept while still performing poorly on the x -intercept. An analysis of common errors supports this interpretation. Most errors are not incorrect calculations; they are students entering one of the coefficients, the absolute value of a coefficient or the constant as the intercept. This pattern is true for both x - and y -intercepts and indicates that student failures are predominantly failures to understand what the intercept means, rather than calculation errors. Our problem sets may inadvertently contribute to the misconception that one of the coefficients is the intercept. In approximately half of the problems, at least one of the intercepts is equal to a coefficient (or its absolute value). When the x -intercept was equal to one of the coefficients (or their absolute values), students were correct 48% of the time, compared to only 40% when the intercepts were different from coefficients. These numbers were 67% vs. 48% for the y -intercept. As a result of this finding, we have updated the problems in our most recent version of the software to eliminate those where the intercepts correspond to the coefficients.

The usual question with sawtooth learning curves is how to split the KC so that we see smoother learning curves for each new KC in the split. In this work we suggest a different way to use such patterns; they are an indicator of the student's conceptual knowledge. While practice in calculating the intercepts might be an appropriate activity for students who do not exhibit the sawtooth curve, remedial instruction in the meaning of intercept (and, particularly, instruction that unifies the graphical and algebraic notions of intercept) might be the correct instructional approach for students to exhibit a sawtooth curve. This insight suggests that adaptive learning systems might do well to pay attention not only to the level and learning rate of various KCs but also to patterns evident in learning curves.

REFERENCES

- Anderson, J.R., Conrad, F.G., & Corbett, A.T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science*, 13, 467-505.
- Anderson, J.R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Anderson, J.R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J.R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, 26, 85-112.
- Anderson, R.B. (2001). The power law as an emergent property. *Memory & Cognition*, 29, 1061-1068.
- Bloom, B.S. (1968). Learning for mastery. *Evaluation Comment*, 1(2).
- Cen, H., Koedinger, K.R., & Junker, B. (2005). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *8th international conference on intelligent tutoring systems* (pp. 164-175). Berlin: Springer.
- Corbett, A.T., & Anderson, J.R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User-Modeling and User-Adapted Interaction*, 4, 253-278.
- Fancsali, S.E., Sandbothe, M., & Ritter, S. (2019). Learning curves segmented by mastery. Manuscript submitted for publication at LAK19 Workshop on Sharing and Reusing Data and Analytic Methods with LearnSphere.
- Goldin, I., Pavlik Jr, P. I., & Ritter, S. (2016). Discovering domain models in learning curve data. In R.A. Sottolare, A.C. Graesser, X. Hu, A. Olney, B.D. Nye, & A. M. Sinatra (Eds.), *Design recommendations for intelligent tutoring systems: Volume 4 - domain modeling* (Vol. 4, pp. 115-126). Orlando, FL: U.S. Army Research Laboratory.
- Heathcote, A., Brown, S. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7, 185-207.
- Junker, B.W., Koedinger, K.R., & Trottini, M. (2000, July). Finding improvements in student models for intelligent tutoring systems via variable selection for a linear logistic test model. Paper presented at the 65th Annual Meeting of the Psychometric Society, Vancouver.
- Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2011). A data repository for the EDM community: The PSLC datashop. In S. Ventura, C. Romero, M. Pechenizkiy, & R.S.J.d. Baker (Eds.). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.
- Koedinger, K., McLaughlin, E., & Stamper, J. (2012). Automated student model improvement. In K. Yacef, O. Zaïane, A. HersHKovitz, M. Yudelson, & J. Stamper (Eds.). *Proceedings of the 5th*

International Conference on Educational Data Mining (EDM 2012) (pp. 17-24) International Educational Data Mining Society.

- Koedinger, K., Yudelson, M., & Pavlik, P. (2016). Testing theories of transfer using error rate learning curves. *Topics in Cognitive Science*, 8(3), 589-609.
- Martin, B., Mitrovic, A., & Koedinger, K.R. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adaptive Interaction*, 21, 249-283. <https://doi.org/10.1007/s11257-010-9084-2>
- Murray, R.C. et al. (2013). Revealing the learning in learning curves. In H.C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.). *Artificial intelligence in education (AIED 2013)*. (LNCS Vol. 7926, 473-482). Berlin: Springer.
- Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.). *Cognitive skills and their acquisition*, (pp. 1-55). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. (2007). Cognitive Tutor: applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249-255.
- Singley, M.K., & Anderson, J.R. (1989). *The transfer of cognitive skill*. Boston: Harvard University Press.
- Zerr, C.L., Berg, J.J., Nelson, S.M., Fishell, A.K., Savalia, N.K., & McDermott, K.B. (2018). Learning efficiency: Identifying individual differences in learning rate and retention in healthy adults. *Psychological Science*, 29(9), 1436-1450.

The Effects of Adaptive Learning in a Massive Open Online Course on Learners' Skill Development

Yigal Rosen^{1,2}, Glenn Lopez², Ilia Rushkin², Andrew Ang², Dustin Tingley²

¹ ACTNext by ACT, Inc.

² Harvard University

{yigal_rosen, glenn_lopez, ilia_rushkin, andrew_ang, dtingley}@gov.harvard.edu

Liberty Munson

Microsoft, Redmond, USA

liberty.munson@microsoft.com

Rob Rubin

Independent

Sharon, USA

rvrubin@gmail.com

Gregory Weber

Microsoft

Redmond, USA

gregory.weber@microsoft.com

ABSTRACT: We report an experimental implementation of adaptive learning functionality in a self-paced Microsoft MOOC (massive open online course) on edX. In a personalized adaptive system, the learner's progress toward clearly defined goals is continually assessed, the assessment occurs when a student is ready to demonstrate competency, and supporting materials are tailored to the needs of each learner. Despite the promise of adaptive personalized learning, there is a lack of evidence-based instructional design, transparency in many of the models and algorithms used to provide adaptive technology or a framework for rapid experimentation with different models. ALOSI (Adaptive Learning Open Source Initiative) provides open source adaptive learning technology and a common framework to measure learning gains and learner behavior. This study explored the effects of two different strategies for adaptive learning and assessment: Learners were randomly assigned to three groups. In the first adaptive group ALOSI prioritized a strategy of remediation – serving learners items on topics with the least evidence of mastery; in the second adaptive group ALOSI prioritized a strategy of continuity – that is learners would be more likely served items on similar topic in a sequence until mastery is demonstrated. The control group followed the pathways of the course as set out by the instructional designer, with no adaptive algorithms. We found that the implemented adaptivity in assessment, with emphasis on remediation is associated with a substantial increase in learning gains, while producing no big effect on the drop-out. Further research is needed to confirm these findings and explore additional possible effects and implications to course design.

Keywords: Assessment Analytics, Resource Recommendation, Diagnostics.

1 INTRODUCTION

Digital learning systems are considered adaptive when they can dynamically change the presentation of content to any user based on the user's individual record of interactions, as opposed to simply sending users into different versions of the course based on preexisting information such as user's demographic information, education level, or a test score. Conceptually, an adaptive learning system is a combination of two parts: an algorithm to dynamically assess each user's current profile (the current state of knowledge, but potentially also affective factors, such as frustration level), and, based on this, a recommendation engine to decide what the user should see next. In this way, the system seeks to optimize individual user experience, based on each user's prior actions, but also based on the actions of other users (e.g. to identify the course items that many others have found most useful in similar circumstances). Adaptive technologies build on decades of research in intelligent tutoring systems, psychometrics, cognitive learning theory and data science [2, 4, 10]. More specifically, Cognitive Tutors utilize knowledge tracing [9] to track knowledge acquisition and provide tailored instruction, by tracking performance on individual production rules in a cognitive model [3, 4, 11]. Extensions to this model have included estimating of the initial probability that the student knows a skill [5], estimating of the impact of help features on probability of acquisition [1], and integrating with models of item difficulty [6]. However, these approaches typically do not consider pacing and require significant content design workload in order to create learning and assessment content [2]. These limitations are critical in large-scale MOOC context. Pioneer studies on adaptive technologies in MOOCs indicated both technical feasibility and the educational promise [7, 8, 9]. Despite the promise of adaptive learning, there is a lack of evidence-based instructional design, transparency in many of the models and algorithms used to provide adaptive technology or a framework for rapid experimentation with different models. Harvard University partnered with Microsoft Learning to develop ALOSI (Adaptive Learning Open Source Initiative) provides open source adaptive learning technology and a common framework to measure learning gains and learner behavior. The key insights gained from the modeling and analysis work enable us to address the development of evidence-based guidelines for instructional design of future courses, and provides insights into our understanding of how people learn effectively. ALOSI uses Bayesian Knowledge tracing to both develop a predictive model of skills mastery for the learner, and improve the predictive attributes associated with the content. The key features in ALOSI's current adaptive framework include knowledge tracing and recommendation engine, while user modeling, feedback and recommendation of targeted learning materials are in development. The engine improves over time from the use of additional learner data and provides direct insights into the optimization processes (by contrast with commonly used commercial "black box" adaptive engines). Additionally, the architecture of the adaptive engine enables rapid experimentation with different recommendation strategies. This pilot study measured the effects of adaptive pathways on learning gains and dropout rates using different tuning parameters in the adaptive engine against the instructional design learning experience.

2 ALOSI ARCHITECTURE

In order to operationalize ALOSI framework, we developed the Bridge for Adaptivity and the adaptive engine, two open source applications supporting a modular framework for implementing adaptive learning and experimentation that integrates several components: the Bridge for Adaptivity, an Adaptive Engine (such as the ALOSI adaptive engine), a Learning Management System (Learning Tools Interoperability - LTI consumer such as Canvas or edX), and a Content Source (for example, an LTI provider like Open edX). The Bridge for Adaptivity

handles the integration of all system components to provide the adaptive learning experience, while the Adaptive Engine provides the adaptive strategy and is designed to be swapped in and out with compatible engines for experimentation and comparison. The diagram in Figure 1 describes the data passing in the system.

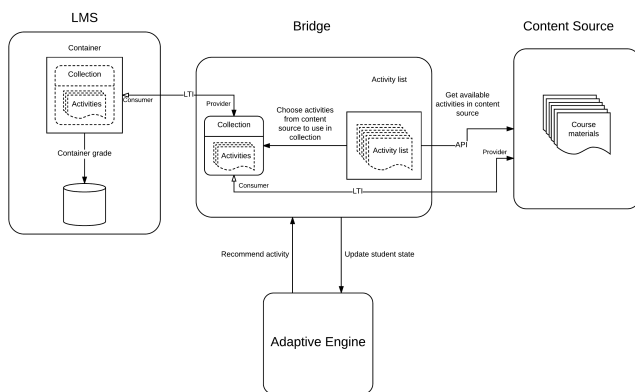


Figure 1. ALOSI Architecture

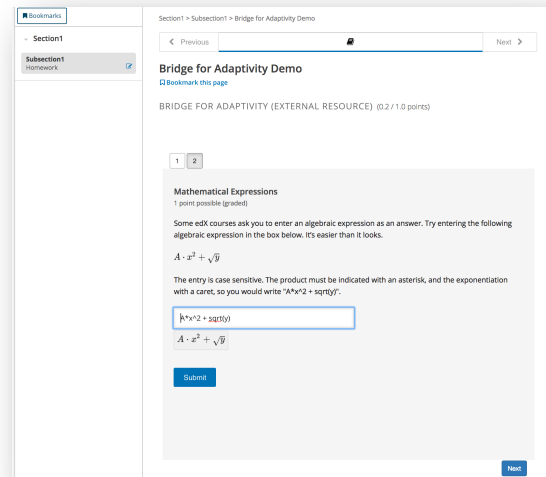


Figure 2. Adaptive assessment user interface

In this study, the Bridge for Adaptivity was used with the ALOSI adaptive engine to adaptively serve assessments from an Open edX platform instance in the Microsoft MOOC on edX.

The user interface seen by a learner when they encounter an installed tool instance is that shown in Figure 2. Assessment items (problems) from the edX course are displayed one at a time in a center activity window, with a surrounding toolbar that provides features such as navigation, and a score display. Every problem-checking event by the user sends the data to the adaptive engine, to update the mastery information real-time. Every “Next Question” event in an adaptive assessment sends to the engine a request for the next content item to be served to the user (this could be a learning or an assessment content). The engine sends back the recommendation, which is accessed as an edX XBlock and loaded.

2.1 Adaptive Engine Details

Our goal was to create a simple adaptive recommendation engine for an edX MOOC, capable of deciding what item to serve to a user next based on the user's history. We use a variety of Bayesian Knowledge Tracing (BKT) model to estimate the students' state. What makes our situation special is that, as we learned from the adaptive pilot and just generally from seeing MOOCs,

- 1) Questions in the course differ widely in nature, and in particular in difficulty. Thus, we cannot assign the same values of guess, slip and transit probabilities to them, even if they are all tagged with the same knowledge component.
- 2) Tagging is complicated: some of the questions are tagged with multiple knowledge components.

3) In a self-paced MOOC environment, there is a need for a causal structure in the knowledge components: we should not serve to a user items tagged with a knowledge component, if the user has shown lack of knowledge of other knowledge components that are pre-requisite to that one. In the simplest case, it can be dictated by a simple ordered list (the natural order of learning the content of the course), but it could also be a detailed graph of pre-requisite relationships among knowledge components.

4) In a MOOC, the number of students is high, so we can afford to define a model with a large number of parameters and optimize them based on the student interaction data.

Conceptually, our engine consists of two blocks: knowledge tracing and the recommendation engine, which uses the output of knowledge tracing as an input.

2.1.1 Knowledge Tracing

Let there be Q questions in the course ($q = 1, 2 \dots Q$), tagged with N knowledge components ($i = 1, 2 \dots N$), or KCs for short. We introduce matrices of guess, slip and transfer probabilities of the questions: $p_{qi}^{\text{guess}}, p_{qi}^{\text{slip}}, p_{qi}^{\text{trans}}$, which are the generalizations of the usual guess slip and transfer parameters of BKT [6]. We do not assume these parameters to be the same for all questions, due to the item diversity.

We assume that the mastery of each KC by each course user is a binary latent variable – the user either has learned it or not – and we update the mastery matrix p , where the element p_{ui} is the currently estimated probability that the user u has the mastery of the KC i . We define the mastery threshold $p^* \in [0, 1]$, and if $p_{ui} \geq p^*$, we say that the mastery of i by the user u is sufficiently certain and no longer needs verification. We initialize the mastery probability matrix $p = p^{(0)}$ (user's prior knowledge), after which, when a user submits an answer to the question, it gets a correctness value (score) $C_q^{(u)} \in [0, 1]$ and we update the mastery probability of each KC (i.e. this user's row of the matrix p).

The Bayesian updating is easier to write in terms of odds, or even logarithmic odds, rather than the probability p :

$$\mathcal{O}_{ui} = \frac{p_{ui}}{1-p_{ui}}, \quad L_{ui} = \log \mathcal{O}_{ui}, \quad L^* = \log \frac{p^*}{1-p^*} \quad (1.)$$

So we will translate the transit, guess and slip probabilities into odds as well: $o_{qi}^{\text{guess}} = p_{qi}^{\text{guess}} / (1 - p_{qi}^{\text{guess}})$ etc, and introduce the likelihood ratios for the case of incorrect (0) and correct (1) answer:

$$x_{qi}^0 = \frac{p_{qi}^{\text{slip}}}{1-p_{qi}^{\text{guess}}}, \quad x_{qi}^1 = \frac{1-p_{qi}^{\text{slip}}}{p_{qi}^{\text{guess}}} \quad (2.)$$

These matrices encode the *relevance* of a question q to a KC i . If the problem is irrelevant to a KC, the probability of correct or incorrect score should be independent of that KC. This will be the case if $p_{qi}^{\text{slip}} = 1 - p_{qi}^{\text{guess}}$, in which case $x_{qi}^0 = x_{qi}^1 = 1$. We propose to define the *relevance* matrix, which is essentially a generalization of tagging, as a sum of logarithmic odds of non-guessing and non-slipping:

$$k_{qi} = \log x_{qi}^1 - \log x_{qi}^0 = -\log o_{qi}^{\text{guess}} - \log o_{qi}^{\text{slip}} \quad (3.)$$

This can be viewed as a generalization of tagging items with KCs. While the tagging matrix is binary (a KC is either linked to a question or not), the relevance matrix shows the weight of each link: how much of an evidence for the KC mastery the question provides. The multiplicative factor earned by the mastery odds is:

$$x_{qi} = x_{qi}^0 \left(\frac{x_{qi}^1}{x_{qi}^0} \right)^{C_q^{(u)}} \quad (4.)$$

For binary (0 or 1) scores, this is just another way of saying that the factor should equal x_{qi}^0 or x_{qi}^1 . But we can also interpolate for fractional scores, and this is what Eq. 4 does. Exactly how we interpolate between these for fractional scores is a matter of choice. For instance, an alternative definition could be a linear interpolation $x_{qi} = x_{qi}^0 + C_q^{(u)}(x_{qi}^1 - x_{qi}^0)$. We settled on the multiplicative interpolation by looking at the location of the "borderline" score, for which $x_{qi} = 1$, representing the boundary between correctness and incorrectness. For instance, as a back-of-the-envelope estimate, let the guess and slip probabilities have equal values (typically, they are not too different). In Eq. 1, this sets the borderline score at a reasonable 0.5, whereas in case of linear interpolation the borderline score in such a situation equals the slip (= guess) probability, which is likely too low.

The posterior odds, with the evidence of the submitted problem, become $\mathcal{O}_{ui} \rightarrow \mathcal{O}_{ui} x_{qi}$. Additionally, we modify the mastery odds due to transfer of knowledge, so the full update procedure is:

$$\mathcal{O}_{ui} \rightarrow o_{qi}^{\text{trans}} + (o_{qi}^{\text{trans}} + 1)\mathcal{O}_{ui} x_{qi} \quad (5.)$$

This is a type of Bayesian Knowledge Tracing. The main modification is that we deliberately formulated it that we formulate it in such a way that there is no explicit requirement to tag each question with only one KC. If a problem is tagged with several KCs ($k_{qi} > 0$ for more than one value i). We essentially view the problem as a collection of sub-problems, each tagged with a single KC. This is our proposed the generalization of BKT to multiple tagging. The predicted odds of correct answer are found as

$$\mathcal{O}_{qu}^{\text{pred}} = \prod_i \frac{\mathcal{O}_{ui}(1-p_{qi}^{\text{slip}})+p_{qi}^{\text{guess}}}{\mathcal{O}_{ui}p_{qi}^{\text{slip}}+1-p_{qi}^{\text{guess}}} \quad (6.)$$

which is to say that we take the ratio of the probability that each sub-problem is answered correctly to the probability that each sub-problem is answered incorrectly (since we must remove from the ensemble the possibilities of correct answer on some but not all sub-problems).

The outlined procedure is multiplicative in nature. An obvious idea would be to replace it with an additive one by working with logarithmic odds L_{ui} (which we do, in fact, in the recommendation part of the engine). It would be clearly preferable from the computational point of view in the knowledge-tracing part as well, if it was not for the knowledge-transfer step: in the additive formulation this step would involve an exponentiation and a taking a logarithm.

For terminological simplicity we referred to the content items as questions. However, the model can accommodate instructional items as well, e.g. videos or text. We can adopt a rule that, if an item q is instructional, the outcome of user's interaction with it is always "correct". A way to think of it is to imagine that q includes an assessment part of trivial difficulty. The slip probabilities $p_{qi}^{\text{slip}} = 0$, the guess probabilities now have the meaning of the probability of not learning an KC from the item, and so we set them to $p_{qi}^{\text{guess}} = 1 - p_{qi}^{\text{trans}}$.

If the matrix p or other parameter matrices contain zeros or ones it is possible to encounter 0/0 indeterminacies. One way to preclude these is adopt a small cutoff, e.g. we can set $\epsilon = 10^{-10}$, and coerce all elements of the parameter matrices $p^{\text{slip}}, p^{\text{guess}}, p^{\text{trans}}$, as well as the initial mastery probability $p^{(0)}$, to the interval $[\epsilon, 1 - \epsilon]$.

2.1.2 Learning parameters of knowledge tracing

We will rely on a way to optimize our BKT parameters, inspired by the "empirical probabilities" method of [7]. At regular points in time, when we decide to run the optimization, suppose that the items submitted by a user u are $\{q_j^{(u)}\}$ ($j = 1, \dots, J^{(u)}$), indexed in chronological order, and let the correctness scores be $C_j^{(u)}$. We denote $K_{ij}^{(u)}$ this student's latent mastery of a KC i just before submitting the item $q_j^{(u)}$. Assuming that there is no forgetting, the knowledge is a non-decreasing function with values 0 and 1, so it is characterized simply by the position of the unit step: for j from 1 to some n_i knowledge is 0 and from there onward it is 1. We need to find which n_i gives the highest accuracy of predicting correctness from knowledge. Once this is done, the knowledge is not a latent variable anymore, and we can estimate guess, slip and transfer probabilities by frequencies of observations. The generalized number of errors on predicting the outcome based on mastery of a particular knowledge component are:

$$E_i^{(u)}(n) = -\sum_{j=1}^n C_j^{(u)} \log o_{qji}^{\text{guess}} - \sum_{j=n+1}^{J^{(u)}} (1 - C_j^{(u)}) \log o_{qji}^{\text{slip}} \quad (7.)$$

where $n \in [0, J^{(u)}]$ and we adopt the convention that if the lower limit of a sum is greater than the upper limit, the sum is 0. We set the knowledge step location for each KC: $n_i = \text{argmin}(E_i^{(u)})$, and construct the step-function $K_{ij}^{(u)}$ using it. If there are multiple equal minima, and hence multiple n_i , we take the average of the corresponding multiple step-functions (because of this, knowledge may now have fractional value). Note that, if user's problems are irrelevant for an KC, we will find a steadily growing knowledge of that KC. This is not bad, however, since for each KC we will average only over the users who experienced some relevant problems. Namely, we can define the sets of users

$$\mathcal{U}_i = \{\forall u: \sum_{j=1}^{J^{(u)}} k_{q_j^{(u)}i} > \eta\}$$

$$\mathcal{U}_{qi} = \{\forall u: \sum_{j=1}^{J^{(u)}} k_{q_j^{(u)}i} \mathbf{1}(q_j^{(u)} = q) > \eta\}$$

where $\eta \geq 0$ is a constant we set as a measure of how much total relevance of a KC is enough for the user to be included into the ensemble for estimating the parameters of that KC. As the simplest choice, in this implementation we set $\eta = 0$.

Now we can estimate the BKT parameter matrices from the user data:

$$p_{u'i}^{(0)} = \frac{\sum_{u \in \mathcal{U}_i} K_{i1}^{(u)}}{\sum_{u \in \mathcal{U}_i} 1} \quad (8.)$$

(same prior knowledge for all users u').

$$p_{qi}^{\text{guess}} = \frac{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}} (1 - K_{ij}^{(u)}) C_j^{(u)} \mathbf{1}(q_j^{(u)} = q) \right)}{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}} (1 - K_{ij}^{(u)}) \mathbf{1}(q_j^{(u)} = q) \right)} \quad (9.)$$

$$p_{qi}^{\text{slip}} = \frac{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}} K_{ij}^{(u)} (1 - C_j^{(u)}) \mathbf{1}(q_j^{(u)} = q) \right)}{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}} K_{ij}^{(u)} \mathbf{1}(q_j^{(u)} = q) \right)} \quad (10.)$$

$$p_{qi}^{\text{trans}} = \frac{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}-1} (1 - K_{ij}^{(u)}) K_{i,j+1}^{(u)} \mathbf{1}(q_j^{(u)} = q) \right)}{\sum_{u \in \mathcal{U}_{qi}} \left(\sum_{j=1}^{J^{(u)}-1} (1 - K_{ij}^{(u)}) \mathbf{1}(q_j^{(u)} = q) \right)} \quad (11.)$$

Here again, we adopt the convention that if the lower limit of a sum is greater than the upper limit, the sum is 0 (this happens when $J^{(u)}$ is 0 or 1). The value of the denominator in each of these expressions is a measure of how much student information we have for estimating the probability. In case there is no data, the expression becomes a 0/0. We should not want to update a probability in this case. Moreover, we imposed a threshold $M = 20$ and did not update a particular matrix element if the denominator in the corresponding equation is less than M . Likewise, we did not update if the calculated value was degenerate, e.g. a guess probability and a slip probability add up to more than 1. The updated prior knowledge values $p^{(0)}$ will be used for all users yet to come to the course, but also for the existing users for those knowledge components that they have not yet been exposed to.

2.1.3 Recommendation engine

The strategy we use for recommending the next item is a weighted combination of a number of sub-strategies. Each sub-strategy comes in with an importance weight (the vector of these weights is a governing parameter of the adaptive engine).

Let us first define the matrix of pre-requisite readiness. The pre-requisite relationships among the KCs are naturally visualized as a directed acyclic graph, and are stored as an $N \times N$ matrix w of pre-requisite strengths, w_{ij} representing the strength of the graph edge (KC j is a pre-requisite for KC i). We define this strength to be on the scale from 0 to 1. If the SME provided no pre-requisite relations form the KCs, w a zero matrix.

The pre-requisite readiness is defined for each KC and for each user as a matrix:

$$r_{ui} = \sum_{j=1}^N w_{ij} \min(0, L_{uj} - L^*) \quad (12.)$$

An element r_{ui} has value 0 if the user has sufficiently mastered all KCs pre-requisite for the KC i , and less than 0 if the mastery probabilities for some pre-requisites are not yet set. If the pre-requisite strength w_{ij} is weaker, it enters r_{ui} with a smaller weight, allowing less certain mastery of less important pre-requisites. If all the pre-requisites are ascertained, $r_{ui} = 0$, otherwise it is negative. We can deviate from this slightly and introduce a forgiveness parameter $r^* \geq 0$, so that a user u is sufficiently ready for learning a KC i if $r_{ui} + r^* \geq 0$.

To recommend the next question for a student, we subset the relevance matrix k_{qi} to only those questions (matrix rows) that belong to the adaptive module where the user u is and that the user has not seen yet. Thus, we obtain a user specific matrix $k_{qi}^{(u)}$. We define the non-negative user-specific vectors of "remediation", "continuity", "difficulty matching", and "readiness" (in terms of difficulty level of the problem $d_q \in [\epsilon, 1 - \epsilon]$):

$$R_q^{(u)} = \sum_{i=1}^N k_{qi}^{(u)} \max(0, L^* - L_{ui}) \quad (13.)$$

$$C_q^{(u)} = \sqrt{\sum_{i=1}^N k_{qi}^{(u)} k_{q_{\text{last}},i}} \quad (14.)$$

$$D_q^{(u)} = -\sum_{i=1}^N k_{qi}^{(u)} \left| L - \log \frac{d_q}{1-d_q} \right| \quad (15.)$$

$$P_q^{(u)} = \sum_{i=1}^N k_{qi}^{(u)} \min(0, r_{ui} + r^*) \quad (16.)$$

where q_{last} is the last item the user saw.

These expressions formulate the four sub-strategies of our recommendation engine. The vectors are the ratings of all potential items by the sub-strategies. The first sub-strategy, "remediation", rates higher those items on whose KCs the user's mastery is currently low. The second, "continuity", rates higher items tagged most similarly to the last seen item. The third favors items with the difficulty level that matches the mastery level and the fourth tries to avoid serving a question if the user has not mastered the KCs that are pre-requisite to the KCs of that question.

More competing subs-strategies can be added to the list at will, but in this implementation we used these four. We introduce a vector of sub-strategy weights: $W = (W_r, W_c, W_d, W_p)$, defined up to normalization. So that the overall rating of the available items is the weighted sum:

$$S_q^{(u)} = W_r R_q^{(u)} + W_c C_q^{(u)} + W_d D_q^{(u)} + W_p P_q^{(u)} \quad (17.)$$

The item q that maximizes $S_q^{(u)}$ will be served to the user u .

The serving stops naturally when we exhausted the available questions (the matrix $k^{(u)}$ has no rows). Additionally, we may adopt a "stop on mastery" policy and stop serving if $R_q^{(u)} = 0$ for all q , which means that the user has reached the mastery threshold p^* on all KCs relevant for the available pool of items.

2.2 Method

Adaptive functionality has been deployed in Microsoft MOOC on edX “[Essential Statistics for Data Analysis Using Excel](#)”. The instructional design team significantly enhanced the assessment scope, and included over 35 knowledge components and 400 assessment items tagged to those knowledge components. Our experimental design randomly assigned learners in the course to three independent groups: in the first adaptive group ALOSI prioritized a strategy of remediation – serving learners items on topics with the least evidence of mastery (Group A); in the second adaptive group ALOSI prioritized a strategy of continuity – that is learners would be more likely served items on similar topic in a sequence until mastery is demonstrated (group B); the control group followed the pathways of the course as set out by the instructional designer, with no adaptive algorithms (Group C). Thus, groups A and B of the students experienced two varieties of the adaptive engine.

The difference was in the recommendation sub-strategy weights. For group A, the weight of remediation was set to 2, and that of continuity to 1. For group B these values were reversed. The weights of the remaining two sub-strategies were the same for both groups: 1 for pre-requisite readiness and 0.5 for difficulty matching. The mastery threshold L^* was set to 2.2 (corresponding to p^* about 0.9. The pre-requisite forgiveness r^* was set to 0. The serving policy “stop on mastery” was not used: as long as a user requested more adaptive questions, they were served until the available pool was exhausted.

Note that the continuity sub-strategy does not use the answer correctness. Therefore, Group B experienced less variability in serving order than Group A (And Group C experience none at all). Furthermore, at the request of the course team, we suspended adaptive serving in the beginning of two assessment modules: the pre-test and the post-test. In these, for Groups A and B, the first 34 or 35 (respectively) items were served in a fixed sequence (same for everyone), and only afterwards the serving order became adaptive.

It should be noted that the approach in the first adaptive group was the most different from the conventional non-adaptive learning experience of the third group, and the second adaptive group occupies the intermediate position. Moreover, in the adaptive groups the learners were working on one item at a time, while in the control group the items were presented in the conventional edX approach – several items at once.

From the course SME we obtained the information about the assessment items: a list of KCs, a list of pre-requisite relations among them, tagging of items with KCs, difficulty level of each item and basic estimates of the guess, slip and transfer probabilities. These were used as cold guesses at the start, and in the progress of the course these values were updated with those learned from the data. The numerical estimates (e.g. the difficulty level or the connection strength between two KCs) were estimated by the SME using a 3-level scale (weak/medium/strong), which we then converted to numbers for the use in the engine.

Although our engine is capable of operating with multiple tagging, in this course it did not happen: each item was tagged with only one KC.

3 FINDINGS

All students in the course were administered a pre-test and a post-test, allowing a comparison of learning gains across three groups of students. For the adaptive groups A and B, the first 34 problems in the pre-test and the

first 35 in the post-test were served non-adaptively: their sequence was fixed, and only the remainder of problems in both tests was served adaptively. Thus, we use the average problem score of only these fixed parts of the tests for the comparison, to ensure that all students are compared on equal footing. For Group C we simply use the entire pre-test and post-test that this group received.

We observe no substantial differences across the groups in the average problem score in the pre-test, confirming the assumption that initially the composition of the three groups is comparable¹. If anything, group A was at a slight disadvantage initially.

The learning gains are observed as the difference between the average problem score in the post-test and in the pre-test. It appears that group A experienced the greatest learning gain (ES=0.641). Group B, whose version of adaptivity was weaker (continuity was emphasized rather than remediation), has lower learning gains (ES=0.542), and the control Group C had still less (ES=0.535).

Table 1. Learning gains across the three groups

Pre-test	Group A	Group B	Group C
Pre-test mean score	0.491	0.520	0.510
Post-test mean score	0.782	0.768	0.758
Effect size of learning gains	0.641	0.542	0.535

We estimate standard error of the post-test participation rates with the help of binomial distribution as slightly over 1% in all three groups, which means that the differences between the post-test participation are insignificant.

In the learning gains analysis above we included all users who submitted at least one question in a pre-test and in a post-test, i.e. students who are both pre-testers and post-testers. So the question remains how many of the pre-testers dropped out without reaching the post-test.

We further investigate the effect of the experimental groups on learning gains: how much of it was due to the simple fact that experimental users had access to many more questions in the learning modules than the control users, and therefore had more chances to practice their knowledge? The number of questions in the fixed sequences in the pre-test and post-test for the experimental groups was 34 and 35, respectively. The number of questions in the pre-test and the post-test for the Control group was 29 and 30 respectively. We have 793 (Remediation/Continuity/Control=238/263/292) users who submitted at least one question in the pre-test and at least one question in the post-test, but restricting the analysis to those who submitted the minimum of 29 pre-test and 30 post-test questions (the numbers of questions from the Control group). As a result, the number of users left is 448 (Remediation/Continuity/Control=127/154/167). Defining the learning gain as the difference between a user's post-test mean score and pre-test mean score, we train on these users a linear model where

¹ Everywhere in this paper, by p value we mean the p-value from the two-tailed t-test, and by the effect size (ES) we mean Cohen's d.

the outcome is the learning gain and the explanatory variables are the pre-test mean score, the experimental group, and the number of questions submitted in the modules 1-5 of the course. The adjusted R-squared of the model is 0.24. As a result, belonging to group A (“remediation”) increases the gain by 0.057 ($p=0.03$) compared to the control group C; belonging to group B (“continuity”) has no significant effect ($p=0.54$). Furthermore, the number of problems turns out to have no statistically significant effect on the learning gain ($p=0.65$), suggesting that the benefit of remediation adaptivity is not explained as simply the benefit of practicing with more questions.

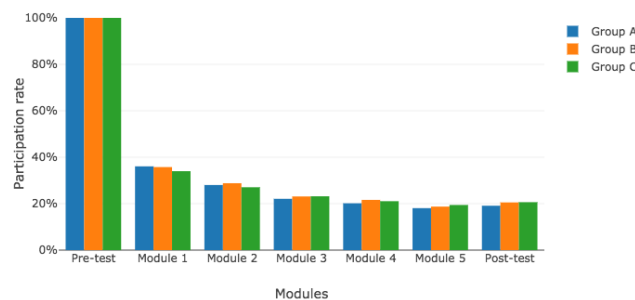


Figure 3. Participation rates by course module

In Figure 3, only the pre-testers are included, so the participation in the pre-test is by definition 100% in any group. The biggest drop-out occurs early on in the course, which is typical for any MOOC. Also, there is a small number of learners who skip the assessment in some modules but go to the post-test - this is manifest from the fact that the participation rate in the post-test is higher than in module 5. The numbers of learners in this graph are A/B/C=1245/1281/1415. The participation rates in the post-test are A/B/C=19.1/20.5/20.6%. We estimate standard error of the post-test participation rates with the help of binomial distribution as slightly over 1% in all three groups, which means that the differences between the post-test participation are insignificant.

We conclude that the implemented adaptivity in assessment with emphasis on remediation (Group A) is associated with a substantial increase in learning gains, while producing no big effect on the drop-out.

The knowledge tracing, which occurs in our engine, allows determining the demonstrated mastery probability for any knowledge component and for any learner after any submit event. This opens up the possibility of visualizing the learning curve, rather than simply relying on the difference between pre-test and post-test scores. Given that we have so many knowledge components, we prefer to aggregate them in groups for the purpose of visualization. Our approach is as follows. Within any assessment module, we average the mastery probabilities of any user across all the knowledge components that are represented in the tagging of the problems in that module. In this way, we create for each user an overall mastery level in a module. Then we can consider group averages of this overall mastery level. In the figure below we plot these group averages of mastery vs. the number of problems tried by a user in the module.

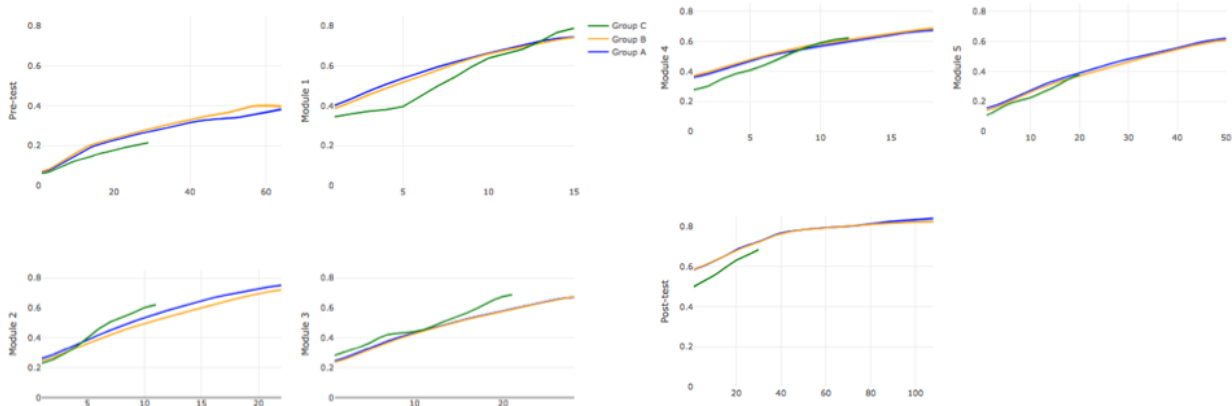


Figure 4. Learner curves by Group, by course module

One noticeable feature is that in many assessment modules the learning curves of adaptive groups are smoother. As the plots show, group C often had a smaller item bank than the adaptive groups, and with the exception of the pre-test, almost all users in this group submitted almost all problems (in the table below we show the mean percentages of submitted problems).

Table 1. Average percentage of problems submitted

	Average percentage of problems submitted						
	Pre-test	Module 1	Module 2	Module 3	Module 4	Module 5	Post-test
Group A	26%	86%	85%	86%	90%	72%	60%
Group B	26%	87%	88%	87%	90%	71%	63%
Group C	51%	93%	96%	97%	95%	93%	91%

Therefore, the sharp twists in the Group C learning curves are not explained away by population stratification. Adaptivity produces a smoother learning experience.

4 CONTRIBUTIONS

Our experimentation with adaptive assessments provided initial evidence on the effects of adaptivity in MOOCs on learning gains and dropout rates. Furthermore, the architecture of the Bridge for Adaptivity and the adaptive engine developed in this project enables rapid experimentation with different recommendation strategies in the future. In this study, adaptivity was implemented on Multiple-Choice assessment problems. There appear to be extensive opportunities to expand adaptive engine to a broad range of assessment item types and enable adaptivity in learning content (e.g., videos, readings, simulations) in MOOCs. Given the structure of many MOOCs, more integration between learning content and assessment could provide an adaptive experience that would guide learners to content that could improve their understanding based on how they perform on integrated assessments. Additional factors could be included to provide a more personalized learning experience. We can conceive an adaptive engine that decides what item to serve next based not just on the mastery but also on career interests and behavioral patterns interpreted as boredom or frustration.

In addition, we anticipate expanding this adaptive learning system to work with other LTI-compliant Learning Management Systems on a large scale.

5 ACKNOWLEDGMENTS

We are grateful for the support from the Office of the Vice Provost for Advances in Learning at Harvard University and Microsoft.

REFERENCES

- Beck, J., Chang, K-M., Mostow, J., and Corbett, A. 2008. Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In *International Conference on Intelligent Tutoring Systems*. Springer, 383–394
- Hawkins, W.J., Heffernan, N.T. and Baker, R.S., 2014. Learning Bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities. In *International Conference on Intelligent Tutoring Systems* (pp. 150-155). Springer, Cham.
- Koedinger, K., Anderson, J., Hadley, W., and Mark, M. 1997. *Intelligent tutoring goes to school in the big city*. (1997).
- Koedinger, K., and Stamper, J. 2010. A Data Driven Approach to the Discovery of Better Cognitive Models. In Baker, R.S.J.d., Merceron, A., Pavlik, P.I. Jr. (Eds.) *Proceedings of the 3rd International Conference on Educational Data Mining*. (EDM 2010), 325-326. Pittsburgh, PA.
- Pardos, Z., and Heffernan, N. 2010. Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Educational Data Mining 2010*.
- Pardos, Z., and Heffernan, N. 2011. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 243–254.
- Pardos, Z., Tang, S., Davis, D., and Vu Le C. 2017. Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework. *Proceedings of the Fourth ACM Conference on Learning @ Scale*.
- Rosen, Y., Rushkin, I., Ang, A., Federicks, C., Tingley, D., and Blink, M. - J. 2017. Designing adaptive assessments in MOOCs. *Proceedings of the Fourth ACM Conference on Learning @ Scale*.
- Rushkin, I., Rosen, Y., Ang, A., Fredericks, C., Tingley, D., Blink, M. J., and Lopez, G. 2017. Adaptive Assessment Experiment in a HarvardX MOOC. *Proceedings of the 10th International Conference on Educational Data Mining*.
- Stamper, J., Barnes, T., and Croy, M. 2011. Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. In Kay, J., Bull, S. and Biswas, G. eds. *Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011)*. 345-352. Berlin Germany: Springer.
- Corbett, A. T. and Anderson, J. R., 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4), 253-278.

Kidaptive's Journey Towards a Scalable Learning Analytics Solution

Josine Verhagen, David Hatfield, Dylan Arena

Kidaptive

{jverhagen, dhatfield, darena}@kidaptive.com

ABSTRACT: Since its founding in 2011, Kidaptive has built customized models that provide adaptivity and/or personalization in online learning environments. We began by enabling adaptive game-based learning through rule-based and then dynamic Bayesian psychometric models. Driven by strong demand, we also started developing models of learner behavior in more traditional online learning and online test preparation environments, based on learners' time management, answer behavior and test scores. These models produce personalized insights for learners and teachers to promote more effective study behavior. Because building a custom solution for every learning environment is not scalable, we have recently been working toward an abstracted version of the models we have built so far, which can be provided as an "out-of-the-box" product offering. One hurdle when onboarding new customers is getting their data into a form that is suitable for the types of (psychometric or behavioral) modeling we offer. Our new product provides a set of customer guidelines for mapping content to learning dimensions or skills and for sending learner responses, response times, and activity data as time-stamped events. Given those data, we provide a set of basic insights about learners' strengths and weaknesses, as well as the time learners take to answer questions and complete tests. Our next goal is to identify under which conditions the more interesting psychometric and behavioral models we developed for previous customers are feasible and valid, and to offer those models to new customers whenever those conditions are met. As we continue to build custom models for customers, we will also expand the set of models we can offer in our out-of-the-box product. This paper will cover some of the models we have successfully implemented (as well as lessons learned in the process), the current status of our self-service product, and some initial explorations of conditions for advanced model offering.

Keywords: Scalable learning analytics, Score prediction, Psychometrics, Game-based learning, Personalized learning

1 SUPPORTING GAME-BASED LEARNING AND PRODUCTIVE STUDY BEHAVIOR

1.1 Adaptive game-based learning

One fascinating aspect of game-based learning environments is their potential for helping players learn valuable skills in contexts that closely simulate the kinds of real-world situations in which such skills might be used (Gee, 2003; Shaffer, 2006) (Figure 1). Proficiency-based adaptivity can support this type of learning by aligning the difficulty of challenges presented to players with the proficiency levels of those players—which necessarily requires a valid assessment of the learner's proficiency. Designing game-based environments to enable valid assessment for proficiency-based adaptivity involves overcoming a number of challenges that are often specific to the particular learning environment.

One such challenge, particularly in contexts designed for young learners, is ensuring that the learner is actually making a skill-based choice. Tutorials that introduce new concepts or game mechanics make sense as guided interactions without assessable outcomes. However, proficiency-based adaptive games need learners to have the freedom to make consequential decisions and to get things wrong. Only when the game can capture both when learners are doing well and when they are struggling will it become possible to adapt to that demonstrated proficiency. A more substantial challenge is ensuring that what makes an adaptive game easy or difficult is as closely aligned with the target skill(s) as possible. When learners struggle or succeed with game challenges for reasons that don't have anything to do with the target skill(s), interpreting performance in terms of the target skill(s) becomes difficult if not impossible (so-called construct-irrelevant variance; Messick, 1994). Designing game-based learning environments so that the range of performance reflects the range in learners' proficiencies on the target skill(s) requires careful attention to the relationships between those skills and the choices and actions available to learners. A related challenge shared by all low-stakes learning environments is that learners in an online environment are not always performing to the best of their ability. Recognizing and accounting for disengaged learners (Baker & Ocumpaugh, 2014), learners who are trying to game the system (Baker, Corbett, & Koedinger, 2008), or learners who are simply trying out every single possibility for the sake of curiosity is essential for valid assessment and therefore for proficiency-based adaptivity.

Designing for proficiency-based adaptivity requires ensuring frequent measurements or other pieces of evidence to update a dynamic learner model. Learner proficiencies are expected to change over time. Sometimes this is simply because the learner is active in the world, and sometimes this change is facilitated by the learning environment. Regardless, adjusting to where the learner is each time he or she engages with a proficiency-based adaptive game requires dynamic and continuous updates to a persistent proficiency model. To allow flexibility and change in the estimate of learner proficiency, we chose early on to use psychometric models within a Bayesian framework. Our core mechanism for assessing proficiency over time is combining prior probability of proficiency or mastery with one observed piece of evidence (typically an item response or similar in-game observation), resulting in a posterior probability of proficiency or mastery (e.g. Bock & Mislevy, 1982). When, for a set of items, the posterior distribution after an observed item response is used as the prior distribution for a subsequent item response, the resulting proficiency estimates are equivalent to those from an assessment with traditional psychometric methods (e.g., a computer adaptive test; Van der Linden & Glas, 2000).

However, within the Bayesian framework, there is flexibility to give less weight to the posterior after the previous observation when determining the prior for the next observation—if, for example, a lot of time has passed between observations, or if there is evidence of an abrupt increase in proficiency (an “aha” moment). It is also possible to base this weight on the (assumed or estimated) probability of transitioning from one mastery state to another (similar to Bayesian Knowledge Tracing (BKT) models; Corbett & Anderson, 1995) or on a growth model. In addition, it is possible to incorporate information from outside the learning environment into a prior probability distribution, such as the learner's age or related assessment information from a different source.

For this proficiency estimation to work, item characteristics such as the difficulty of the items presented to a learner must be known. This is an additional challenge, both because the nature of

educational games encourages growth in learner proficiency and because adaptivity leads to items being presented only to subsets of learners with similar proficiency, which limits the use of traditional item calibration methods. As it is often not feasible for customers to have a pilot sample of learners to calibrate (a random subset of) game challenges, we have developed calibration methods using a combination of initial “guesstimates” and creative empirical calibration and equation methods, many of which use a version of the ELO-based algorithm (e.g. Pelánek, 2016) .

In summary, while the core mechanism of our proficiency-based adaptivity is the same across learning environments, in each new adaptive game the way in which we acquire evidence for proficiency and the way in which we can use additional information to improve our assessments is somewhat different and requires a significant and complex modeling effort.

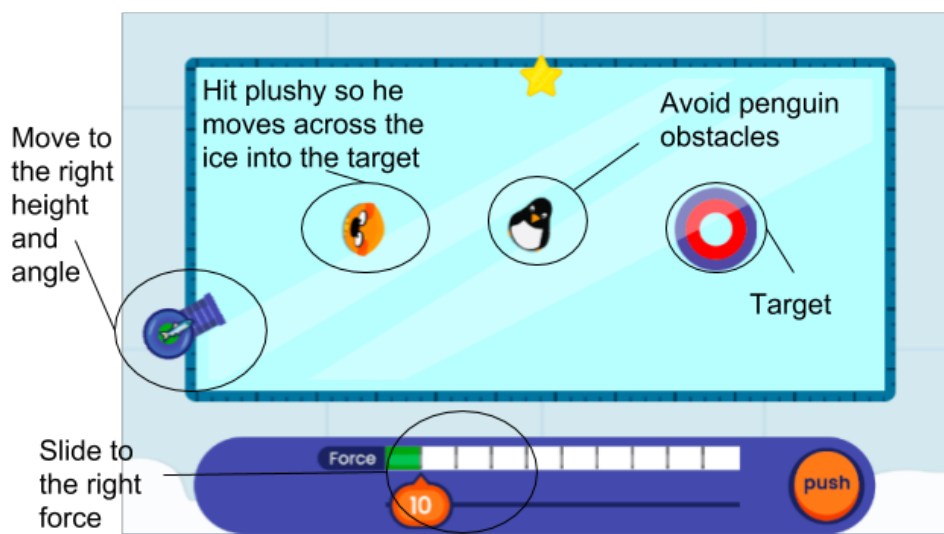


Figure 1. Fish Force game for young children to gain proficiency in scientific inquiry supporting knowledge about forces and motion by PBS kids.

1.2 Supporting productive study behavior

Advances in accessibility of digital content have led to a move from offline to online learning curricula. University courses are turned into MOOCs, publishers are converting their textbooks to interactive online courses, and preparation for standardized tests is moving from workbooks and tutors to online programs. Although it is scalable to replace teacher instruction with videos, the motivational component of learner-teacher interaction proves harder to scale. Dropout rates are very high in MOOCs (e.g. Andres, Baker, Gašević, Siemens, Crossley, & Joksimović, 2018) and learners find creative ways to game online learning environments that are still in their infancy (e.g. Baker et al., 2006). Many online learning environments lack a good representation of how learners are doing beyond the number of questions they get correct, and very few such environments show learners how their work contributes to achieving their learning goals, be it content mastery or a target standardized test score.

For the past few years we have supported a Korean publisher of educational textbooks for K–6 learners in Math, Korean, Social Studies and Science. This publisher historically offered printed

textbooks in combination with weekly in-person tutoring sessions. After the publisher transformed its content to interactive online material, we entered into a partnership to provide data-driven insights for tutors to use on their weekly visits; we have also co-developed ways to motivate productive self-study of the curriculum. The study material is organized equivalent to book chapters, with concept explanations followed by sets of practice questions. Learners are expected to complete a chapter each week, concluding with a test of the studied material. Developing models to support learning in this environment began with getting a good sense of the learner experience. Even though the learners and tutors could see which questions the learner got right or wrong, there was no clear expectation about the difficulty or time-intensiveness of questions. In addition, we discovered that many learners were using unproductive study behaviors like speeding, guessing, or entering random answers to see the hints provided after mistakes and then using those hints to produce a serious answer.

The first set of models we developed are based on (Bayesian) models for working speed and learner ability relative to peers, taking relative item difficulty and question or study duration into account. The results of these models are used to set personalized expectations about how much time to spend on a question (or question set) and the expected probability of getting a question correct. The models are also used to detect when learners are deviating from an expected pattern of behavior. Uncharacteristically long or extremely short answer times can indicate disengagement or struggle with a particular topic, while fast wrong answers to relatively easy questions can indicate skipping or random answer behavior. We have also developed models targeting specific problematic behaviors: e.g., for guessing, we developed a model that estimates the minimum time required to respond correctly to specific open-ended questions.

After implementing this first set of models in production, we then began distinguishing different learner types based on study behaviors, with the goal of personalizing study tips and content presentation for each group of learners. To achieve that goal, we performed a cluster analysis on study behaviors within chapters. No clearly separated clusters were found, but behaviors tended to cluster (a) learners who were taking relatively more time to master the material, (b) high-performing fast learners, (c) low-performing fast learners who seemed to skip or guess a lot, (d) learners who would not complete or skip lectures and questions, and (e), the largest group, learners doing essentially what they should be doing. To make the result of the cluster analysis actionable, however, we needed to make sure that each learner would fall into exactly one group that could receive tailored messages or content. The final “clusters” are inspired by but go beyond the initial cluster analysis; these “clusters” are defined by a set of thresholds delineating groups of learners with one or multiple significant behaviors. Learners classified into a group get messages to encourage or celebrate productive study behavior that are targeted to the characteristics of that group. The online environment also uses the categorization to eliminate some content for learners who take more time to master a topic, and to present more challenging content to fast high achievers.

Finally, an important goal was to help learners improve their test scores and set realistic expectations. This work is still in progress. We developed a model for test score prediction based on test scores in previous chapters of the same subject and test difficulty. The prediction gives the learners a realistic sense of an attainable test score given their previous performance and the

difficulty of the test. As we will describe in section 2, we are working towards an improved model that can inform learners how far they are from reaching their desired level of performance in the domain they are currently working on.

In summary, we learned a lot about building models to support online learning environments by diving deep into the data and patterns of this particular learning environment, as well as making the results of our models actionable in a product. Even though the way data are collected and the particular learner behaviors will differ in other learning environments, we believe that some of the core characteristics will be applicable to learning environments more widely.

2 TOWARDS SCALABLE LEARNING ANALYTICS

Because building a custom solution for every learning environment is not scalable, we have recently been working toward an abstracted version of the models we have built so far, which can be provided as an “out-of-the-box” product offering. Our goal is to provide a learning analytics solution that can support many different existing learning products. In the context of this goal, we define scalability not so much as capacity to process more data but instead as capacity to support a wider variety of learning contexts.

2.1 Basic learning analytics

One hurdle when onboarding new customers is getting their data into a form that is suitable for the types of (psychometric or behavioral) modeling we offer. Our new product provides a set of customer guidelines for mapping content to learning dimensions or skills and for sending learner responses, response times, and activity data as time-stamped events. Given those data, we provide a set of basic insights about learners’ strengths and weaknesses, as well as the time learners take to answer questions and complete tests. These insights can be used to generate learner-facing reports about their activity or to provide a dashboard giving the product owner an overview of what and how learners are doing in their product.

Although designers of learning products often care deeply that their products should *teach* particular concepts or skills, they often have not thought explicitly about how to *measure* when or even whether learners using their products are in fact learning. To do so means expanding these designs to map the things learners do, such as their responses to particular questions, to the skills intended to support achievement in those activities, in a way that supports inferences about learner proficiency and growth. This expansion involves identifying both which questions measure which skills and which other activities might provide additional evidence of proficiency, mindset, and/or engagement, and then generating event data when these measures occur during learner activity.

To make the mapping process more tractable, we advise our customers to use a predefined Core Skills Framework. Reference frameworks like the Common Core State Standards specification provide clear, research-based definitions for valuable skills that organizations can use when mapping their own learning products. For Kidaptive's customers, we have developed Core Skills Frameworks for the SAT (based on College Board documentation) and TOEIC standardized tests (based on ETS documentation), and a comprehensive early learning framework for preschoolers (based on state

and national level standards documents such as the California Department of Education, Child Development Division, 2010)

Once events have been defined and questions mapped, an important technical consideration involves how data are stored. Tracking and analyzing changes in behavior and proficiency over time require working with granular rather than aggregated data. It can be tempting to only track and store summary and aggregated data, such as final scores, sums of correct answers, or the latest attempts at questions, and it may seem efficient to avoid storing each individual event as opposed to simply updating an aggregate record. However, such choices can make detailed learning analytics extremely difficult if not impossible by removing the specific traces over time required. To report basic learning analytics, we ask our customers to collect time stamped data on interactions with the learning environment, as well as detailed information about question responses such as response times and attempt counts. Based on those data, we provide basic statistics about study behavior relative to other learners in the product, which can be shared with a student or teacher (Table 1).

Table 1: Basic learning analytics provided

	Statistics
Activity	Time the learner spent on each topic and distribution of study time across different modalities (e.g. lectures, questions).
Outcome data	Learner performance compared to other learners working on the same questions (first and second attempt), per topic or learning dimension.
Response time	Learner question response times relative to other learners. Distribution of response times across a set of test questions.

2.2 Insights from more complex models

Our next goal is to identify under which conditions the more complex psychometric and behavioral models we developed for previous customers are feasible and valid, and to offer them to new customers whenever those conditions are met. The first two models that we are preparing for scalability are a model for test-score prediction (based on our study behavior models) and a psychometric model for proficiency estimation and subsequent content recommendation (based on the models we developed for game-based learning).

2.2.1 Test score prediction

In many learning environments, test scores are used as a proxy for learner proficiency. Machine learning models are not ideal in these situations, because inferences about and feedback on the learning process are more important than predicting the test scores themselves. An exception is preparation for standardized tests, where learners typically strive to achieve a target test score. Being able to show a learner at any time how close he or she is to that score is therefore a valuable addition to any test preparation learning environment. We are investigating the performance of several machine learning models to enable real time test score prediction for students in test preparation environments, which are scalable across different products. Even though models customized to specific learning environments will necessarily result in better accuracy, we are aiming for a standardized model that will do a good enough job to be useful in practice.

Recent studies point to a combination of performance and behavior related features as optimal for predicting standardized test scores based on interaction data from online learning environments (e.g. Ritter, Joshi, Fancsali, & Nixon, 2013; Pardos, Baker, San Pedro, Gowda, & Gowda, 2014; Feng & Roschelle, 2016; Kostyuk, Alameda & Baker, 2018; San Pedro, Snow, Baker, McNamara, & Hefferman, 2015). We investigate several feature sets that we expect to be available in most test preparation learning environments: a set of features tracking performance on questions related to the various subdomains of the standardized test, a set of features related to additional test taking behaviors that can influence test score (e.g. time management when answering questions) and a set of features related to online study behavior (e.g. time spent on lectures versus practice, strategies for revisiting difficult concepts, engagement related metrics).

To be able to make predictions in any learning environment, certain conditions need to be met. For a predictive model to be trained, access to scores on the final test and/or regular scores on preliminary mock tests are required. In addition, a good representation of the skills required to perform well on the final test in the learning environment is necessary. Finally, the way in which information about these skills is collected in the learning environment needs to be standardized. We aim to provide good guidelines around mapping questions to skills and about collecting events from a learning environment to achieve a basic generalizable model for standardized test score prediction that can be trained for specific learning environments.

2.2.2 *Psychometric models*

Any learning environment that aspires to go beyond a simple reporting of how many questions a student got correct will need a psychometric model to support inferences about how a student's performance relates to the student's proficiency and progress in targeted learning dimensions. Such models are only valid and reliable if various assumptions are met, which makes scaling them a challenge. We are planning to provide a basic psychometric model (equivalent to a Rasch model; Rasch, 1960) to estimate the difficulty of a set of items (which by itself can be valuable for a product owner) and then to use these item difficulties to make inferences about learner proficiencies. We aim to support three different uses of these proficiency levels in our partners' learning products:

- Reporting on proficiency relative to other learners studying similar content, and reporting on personalized item difficulty
- Detecting and reporting on changes in proficiency over time
- Adapting content based on learner proficiency and item difficulty levels

Expert knowledge has proven an unreliable source for the estimation of item difficulty (Bejar, 1983; Hambleton, Sireci, Swaminathan, Xing, & Rizavi, 2003). Although expert knowledge can be useful as a starting point for (Bayesian or iterative) estimation of item difficulty, some form of empirical calibration of the difficulty of items in a learning product is usually necessary to avoid inaccurate proficiency estimates. Because item difficulty and learner proficiency are defined relative to each other, empirical estimates of item difficulty are highly dependent on the proficiency of the specific set of learners used to perform this calibration. Therefore, to reliably calibrate items in a learning environment, it is important to understand how the question-based educational content is structured in terms of which learners encounter which questions at what point in time.

Learning environments typically contain three different types of question-based content:

Diagnostics: single-session assessments where learners can test themselves. No feedback is given during the assessment; thus, we can assume a stable proficiency during this time.

Curriculum-based practice questions: blocks of questions always presented as a set, often related to a chapter or explanation in a curriculum. Even though the selection of question sets could in principle be personalized for each learner, the curricular sequencing of most products means that in practice learners typically proceed through blocks of questions in roughly similar order. Incorrect responses typically lead to feedback, and proficiency is assumed to increase with practice.

Continuously adaptive practice questions: A large set of questions administered sequentially, where questions get easier or harder depending on learner performance. Incorrect responses typically lead to feedback, and proficiency is assumed to increase with practice.

Different forms of empirical calibration are feasible for different types of question-based educational content, and these forms of calibration support different uses of the resulting proficiency estimates.

If a diagnostic or set of curriculum-based practice questions contains mostly (e.g. > 50%) the same questions for all learners who reach that part of the curriculum, it is possible to calibrate items with traditional item calibration methods. The resulting proficiency estimate is representative of how the learner did on that set of questions relative to other learners; these estimated proficiency estimates are comparable to “grades” for different subtopics within a subject throughout a school year, and will look similar to the proportion of items answered correctly. In the case of moderately personalized questions, it is possible to report on relative proficiency taking the different question paths into account. Additionally, the model can be used to indicate which questions were expected to be specifically difficult or easy for a learner by reporting on the probability that a learner would get a question correct.

Tracking and reporting on growth in learner proficiency over time is an important goal of many learning products. Following change over time requires both repeated assessments of the same learning dimension over time as well as a way to link the difficulty of the items in the different assessments, so that the proficiency estimates at different time points can be compared. Most learning environments are not set up in a way that enables this linkage of items from natural interaction with the product. In traditional curriculum-based settings, items are usually tied to one assessment that occurs at one point in the curriculum. Learner proficiency will increase throughout the curriculum, but without reference points IRT models cannot distinguish overall growth in learner proficiency from overall decrease in item difficulty. In adaptive settings or personalized curricula, (sets of) items are often presented only to learners with high, medium or low performance, which prevents calibration of all items relative to each other.

We aim to provide a set of recommendations for (approximate) linking of items across curricula and levels of adaptive learning environments. A reliable way of linking items in different parts of the curriculum is to repeat items or equivalent versions of items (e.g., by automated item generation; Gierl & Haladyna, 2012) over time; another is to augment an existing set of items with a set of pre-calibrated items. Correspondingly, a reliable way of linking items in an adaptive practice environment is to pilot questions from across the difficulty spectrum with a representative sample of learners and then use the resulting calibrations to anchor estimates for questions that are frequently attempted by learners who attempt the piloted questions.

However, the above linkage methods require changes to the learning environment not every customer is in a position to make. There are some alternatives that can lead to approximate calibration of items, which can be sufficient in a low stakes environment like a learning product. Examples are:

- Rely on initial difficulty estimates based on expert knowledge
- Make the assumption of equal test difficulty of supposedly equivalent (mock) tests in a standardized test preparation environment
- Use ability estimates from an initial diagnostic to link subsequent practice items to each other and the diagnostic.

Once approximate estimates of item difficulty and/or learner ability been established, we have found that in many cases fine-tuning the calibration of items based on an iterative approximation of the Rasch model following ELO-based heuristic equations (e.g. Brinkhuis, Savi, Hofman, Coomans, & van der Maas, 2018; Klinkenberg, Straatemeier, & van der Maas, 2011; Pelánek, 2016) is very efficient and provides mostly accurate item difficulties, while it can account for changing learner abilities over time. Another way to finetune these parameters is to use them as priors for a Bayesian parameter estimation method.

3 CONCLUSION

This paper presented an overview of the work Kidaptive has done over the past seven years on learning analytics solutions for first, second, and third party learning products. Kidaptive is currently in the process of using lessons learned from early work on adaptive game-based learning and supporting productive study behavior to provide a scalable learning analytics solution that can work “out-of-the-box” with new learning environments. As we keep building custom models for customers, we are aiming to simultaneously expand the set of models we can offer.

REFERENCES

- Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., & Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 71-78). ACM. <https://doi.org/10.1145/3170358.3170369>
- Baker, R. S., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D.J., & Beck, J. E. (2006). Adapting to when students game an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems* (pp. 392-401). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11774303_39
- Baker, R.S.J., Corbett, A.T., Roll, I., & Koedinger, K.R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18, 287-314. <https://doi.org/10.1007/s11257-007-9045-6>
- Baker, R. S., & Ocumpaugh, J. (2014). Cost-Effective, actionable engagement detection at scale. In *EDM* (pp. 345-346).
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303-310. <https://doi.org/10.1002/j.2333-8504.1981.tb01274.x>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer

- environment. *Applied psychological measurement*, 6, 431-444.
<https://doi.org/10.1177/014662168200600405>
- Brinkhuis, M. J., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics*, 5(2), 29-46. <https://doi.org/10.31234/osf.io/g4z85>
- California Department of Education, Child Development Division. (2008). California Preschool Learning Foundations (Volume 1). Retrieved from <http://www.cde.ca.gov/sp/cd/re/documents/preschoollf.pdf>
- Corbett, A. T.; Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278.
<https://doi.org/10.1007/bf01099821>
- Feng, M., & Roschelle, J. (2016). Predicting students' standardized test scores using online homework. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 213-216). ACM. <https://doi.org/10.1145/2876034.2893417>
- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan. <https://doi.org/10.1108/et.2004.00446dae.002>
- Gierl MJ, Haladyna TM. (2012). *Automatic item generation: Theory and practice*. New York: Routledge. <https://doi.org/10.4324/9780203803912>
- Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). Anchor-based methods for judgmentally estimating item difficulty parameters. *LSAC Research Report Series*.
<https://doi.org/10.4324/9780203874776.ch18>
- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57, 1813-1824. <https://doi.org/10.1016/j.compedu.2011.02.003>
- Kostyuk, V., Almeda, M. V., & Baker, R. S. (2018). Correlating affect and behavior in reasoning mind with state test achievement. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 26-30). ACM. <https://doi.org/10.1145/3170358.3170378>
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–23. <https://doi.org/10.2307/1176219>
- Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, 1, 107-128.
<https://doi.org/10.18608/jla.2014.11.6>
- Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98, 169-179. <https://doi.org/10.1016/j.compedu.2016.03.017>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Educational Research Institute, 1960. <https://doi.org/10.4135/9781412961288.n335>
- Ritter, S., Joshi, A., Fancsali, S., & Nixon, T. (2013). Predicting standardized test scores from Cognitive Tutor interactions. In *Educational Data Mining 2013*.
- San Pedro, M. O. Z., Snow, E. L., Baker, R. S., McNamara, D. S., & Heffernan, N. T. (2015). Exploring dynamical assessments of affect, behavior, and cognition and math state test achievement. *International Educational Data Mining Society*. <https://doi.org/10.1057/9780230601994>
- Shaffer, D. W. (2006). *How computer games help children learn*. New York, NY: Palgrave Macmillan.
- Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

PERSEUS – a Personalization Services Engine for Online Learning

Michael Yudelson

ACTNext by ACT, Inc
michael.yudelson@act.org

Peter Brusilovsky

University of Pittsburgh
peterb@pitt.edu

ABSTRACT: PERSEUS – a Personalization Service Engine – first launched almost a decade ago in 2009 as an attempt to promote the reuse of a suite of adaptation and personalization techniques previously piloted as separate "shell" apps providing adaptive access to educational resources. The design of PERSEUS targeted the ease of development and deployment of personalization and adaptation services in a generic Learning Management System. Since its first deployment, PERSEUS has been used in a several dozens of unique courses and nearly a hundred course offerings. PERSEUS features several classes of services supporting skill-based student knowledge tracking, topic-based modeling, and social navigation. It relies on an external user modeling server that supplies various forms of student activity statistics. PERSEUS requires the linear map of a learning space being personalized and outputs a personalized version of the map. Personalization could be in the form of modifying the elements of the learning space (adding, removing, reordering) or annotating the space and adding interactive elements to it.

Keywords: Adaptation service, personalization, component-based architecture.

1 INTRODUCTION

Modularization, component reuse, and interoperability has long been one of the promising trends in the field of adaptive education. The trend towards modularization started even before adaptive educational systems came to existence. Early user modeling shells separated from the user-adaptive systems in the 1980s. In the late 1990s, open-corpus hypermedia allowed content to be added at the run-time rather than at the design time. It's been a matter of time before adaptation would become a component or even a service of its own. First architectural solution – AHA! (De Bra & Calvi, 1998) – was proposed 20 years ago. Then, adaptive architecture decomposition was an idea ahead of its time. Fast forward six years later, a similar class of an architecture, Knowledge Tree (Brusilovsky, 2004), is proposed. The vision of the latter over the next few years leads to the design, creation, and validation of PERSEUS – a service-based engine for adaptation in online learning systems (Yudelson, 2010). Despite the fact that above mentioned works were introduced long time ago, it is only recently that adaptation and personalization floated up on the agenda of the computer-assisted learning and measurement fields.

In this work we are describing PERSEUS, a personalization server that started as a proof-of-concept of a component-based modular adaptive learning architecture. PERSEUS offers an abstraction of a fleet of commonly used adaptation techniques offered to content developers of content within

ADAPT2¹ architecture developed at the Personalized Adaptive Web Systems (PAWS) lab at the University of Pittsburgh. PERSEUS relies on the user-modeling server CUMULATE (Yudelson et al., 2007) and produces adaptive annotation and visualization.

PERSEUS allowed adaptation value to be requested as a service rather than programmed and built into the content. Multiple adaptation services were reused in many contexts and could be swapped without changing the structure of the courses. The problem of designing and building of the adaptation was reduced to the problem of configuration.

2 PERSEUS – PERSONALIZATION SERVICES ENGINE

The conceptual idea of how PERSEUS works is shown in Figure 1. A content management system (here ADAPT2's Knowledge Tree portal) provides access to a pre-constructed hyperspace. To render a personalized view of a particular page, the portal consults the personalization service engine. To do that, the portal sends the structure of the currently viewed page (as an RDF XML document) and context information (user/group id, personalization algorithm code, etc.). PERSEUS queries user modeling server(s) (and/or other data sources known to it) and performs the adaptation that was requested. The returned result is an RDF/XML document with personalizing updates. The new feed may have original links reordered or removed, new links inserted, annotations added to links. The portal parses the feed and renders a personalized page for the user.

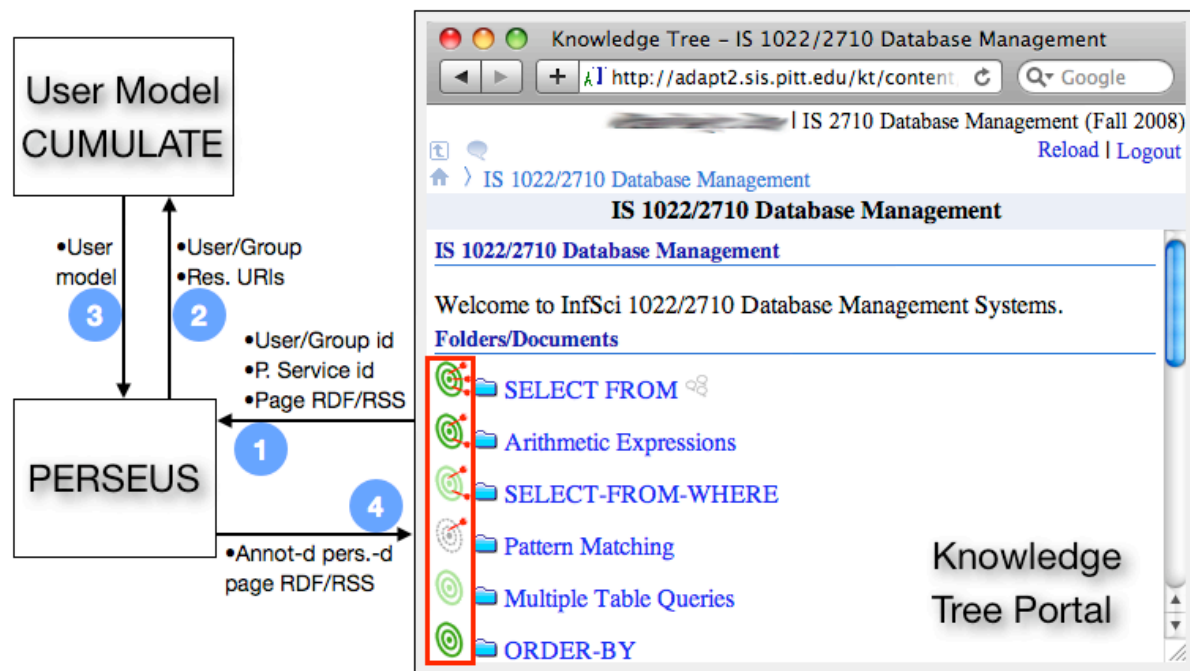


Figure 1: Example of PERSEUS's topic-based adaptive navigation for Knowledge Tree portal (Brusilovsky et al., 2005). Adaptive annotations (targets with darts) produced by PERSEUS are enclosed in a square. CUMULATE (Brusilovsky et al., 2010) is utilized as the user modeling server.

¹ ADAPT2 architecture <http://adapt2.sis.pitt.edu/wiki/ADAPT2>

The compliance requirement for the consumer system to be able to use PERSEUS is minimal. Every personalization method implemented in PERSEUS is exposed as a RESTful web service. The system has to be able to package its pages' link structures as a simple RDF/XML document and send it as one of the parameters to the selected service's URL. PERSEUS's response – modifications to the link structure, including annotations, potentially with interactive JavaScript code– should be parsed back from RDF/XML.

PERSEUS is built on the basis of the following elements.

- **PService** (personalization service) – is the algorithm, the computational approach to adaptation that is being implemented. The nature of the computation is often to extract student resource- or concept-based progress and aggregate it for posterior visualization. PERSEUS supports the following PServices.
 - Topic-based adaptive annotation. Student's learning is aggregated by topic, where each topic unites a set of interactive learning resources indexed with a concept taxonomy. See Figure 1 for an example of a result of such adaptation.
 - Concept-based adaptive annotation. Student learning is aggregated by resource (problem, example, or tutorial) and across concepts that it is indexed with.
 - Social adaptive annotation. In this case, no concept indexing is necessary; aggregation is performed on access/attempt count or success rate for a particular student and the average access/attempt count or success rate for their class/group. Contrasting the two values is the basis of the social annotation that promotes social motivation by instigating soft competition.
 - Resource recommendation. This PService queries a pool of resources additional to the one in the request and recommends them for a student to consider. One of the instances of this PService recommended programming examples for a lesson-full of problems and examples selected by the teacher.
- **Context** is a set of static and dynamic data sources that a PService relies upon. These data sources could be entry points to user models representing student knowledge and learning, static configuration data, and other data-bases. One context could be used by multiple PServices and vice-versa.
- **Visualizer** is a mapping function that takes raw aggregated values a PService generates and maps them to visualization cues to be displayed. In Figure 1, the cues are topic progress summaries showing 0, 1, 2, or 3 darts. Also, if topic schedule data is part of the context, target color is faded if the topic is not the focus of the class. Multiple visualizers could be applicable to one PService and vice versa.

Every call to PERSEUS has to contain the three elements described above: a PService id, a context id, and a visualizer id. In addition, a call for personalization usually contains user id, group id (class id), and, often, a pointer to a RDF/XML specification of the page to be personalized. PERSEUS PServices were built to provide adaptive annotations to named lists of links to educational resources.

Traditionally, every lesson of a course offered via ADAPT2 LMS called Knowledge Tree had a textual description and a list of resources students were offered.

3 PERSEUS EMBEDLETS – ADAPTATION MADE EASY

Adding adaptive annotation and personalization to a simple LMS although easy, requires a lot of preliminary work to create the course structure and add links to the learning resources. An extension to PERSEUS called embedlets, allows course authors to skip all that in its entirety and, instead, work with plain HTML code. An embedlet is a stationary configured call to one of the PERSEUS's adaptation techniques. It is comprised of an RDF/XML document – a flat link list with all of the course resources and a pointer to the desired personalization service. Each embedlet is exposed as quasi personalization service.

To invoke an embedlet one should insert an <object> HTML tag into the web page with data attribute pointing to the embedlet's URL. To only show part of the links from the embedlet's exhaustive list one must use an additional parameter and enumerate a subset of links. User and/or group identity need to be present in the embedlet as well. However, in the case of group-based navigation (see Figure 2), no individual users are distinguished, and group identity could also be statically specified, which allows adaptive navigation embedlets to be successfully used in static HTML pages. Embedlets are equivalent to regular PERSEUS services in terms of adaptation functionality offered. However, in terms of authoring they are a lot simpler.

4 DATA

Since its introduction, PERSEUS and ADAPT2 architecture it is part of have seen a significant use. 320 class sections and study groups were created containing 6m,700 students. Over 4,000,000 student actions were processed by the user-modeling server. These actions spanned from portal navigation to engaging with various types of content across subjects like C and Java programming, Databases and SQL, Interactive Systems Design, etc. PERSEUS processed 144,000 requests for personalization. Interactive adaptive cues generated by PERSEUS were viewed 15,000 times.

One of the types of data that PERSEUS collects are snapshots of the pages it personalizes for students. For each of the 144,000 requests PERSEUS processed, it is possible to restore the list of the educational resources that were listed on the page as well as the adaptive annotations and visualizations that were generated. Every adaptive view is tagged with a token that is passed to all subsequent interactions the student undertakes. Thus, it is possible to track the effect each adaptation has on further student path including learning.

5 CONCLUSIONS

PERSEUS is a service-based adaptation engine designed and developed nearly a decade ago. It was actively used maintenance free for over eight years. Although technology made a leap, PERSEUS is still highly capable as it was developed to sustain up to several thousand concurrent users (Yudelison, 2010). To the best of our knowledge, the design of and the feature set of PERSEUS has not been surpassed yet.

New adaptation engines, many of which are presented in this workshop, started to appear over the last few years. Computer-assisted learning and assessment industries are demanding adaptive solutions to be developed and deployed to handle larger volumes of traffic with strict latency requirements.

Having appeared at the dawn of Adaptive Educational Hypermedia era, PERSEUS would soon have to be replaced or reborn as a new system answering challenges of the day.

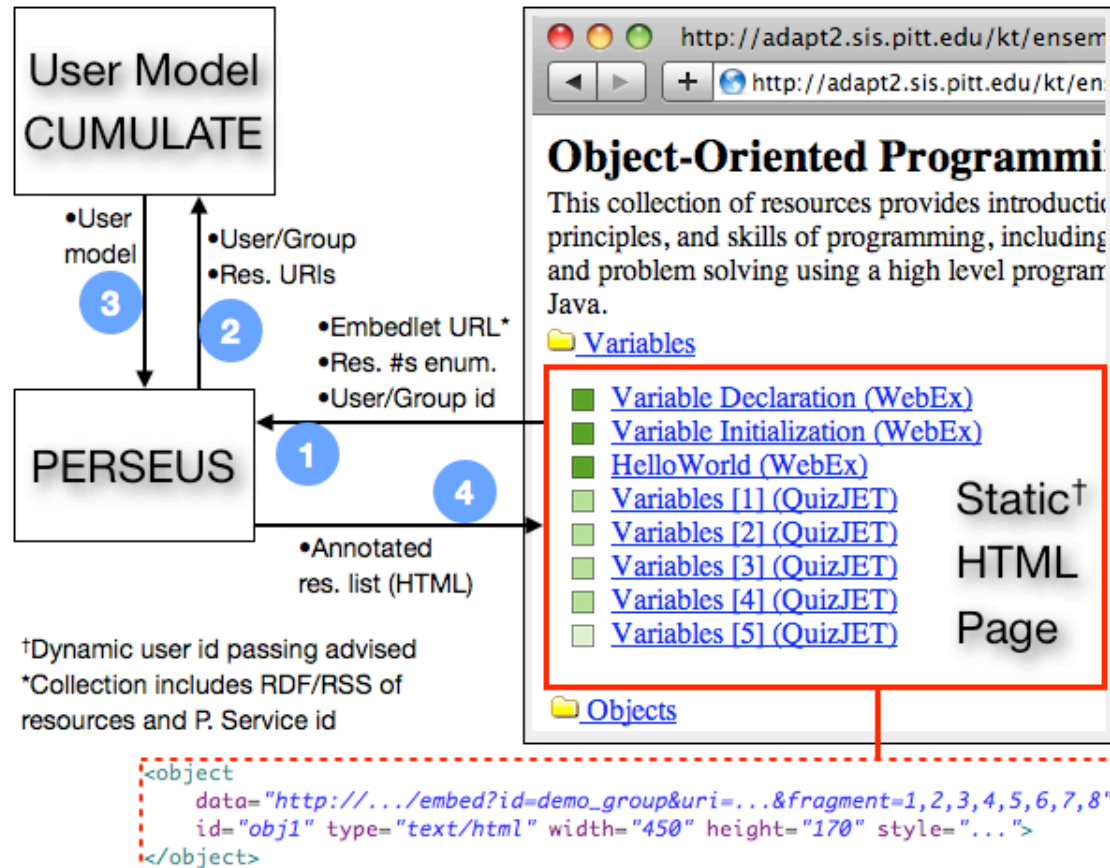


Figure 2: PERSEUS's group-based social navigation support as an embedlet.

REFERENCES

- Brusilovsky, P. (2004) KnowledgeTree: A distributed architecture for adaptive e-learning. In: Proceedings of The Thirteenth International World Wide Web Conference, WWW 2004, New York, NY, 17-22 May, 2004, ACM Press, pp. 104-113.
- Brusilovsky, P., Sosnovsky, S. A., Yudelso, M., Lee, D. H., Zadorozhny, V., and Zhou, X. (2010). Learning SQL Programming with Interactive Tools: From Integration to Personalization. ACM Transactions on Computing Education, 9(4), 1-15. <http://dx.doi.org/10.1145.1656255.1656257>
- Brusilovsky, P., Sosnovsky, S. A., and Shcherbinina, O. (2005). User Modeling in a Distributed E-Learning Architecture. In L. Ardisson, P. Brna, and A. Mitrovic (Eds.), 10th International Conference on User Modeling (UM 2005), (pp. 387-391). http://dx.doi.org/10.1007/11527886_50

- De Bra, P. and Calvi, L. (1998) AHA! An open Adaptive Hypermedia Architecture. *The New Review of Hypermedia and Multimedia*, vol. 4, pp. 115-139, Taylor Graham Publishers.
- Yudelson, M., Brusilovsky, P., and Zadorozhny, V. (2007) A User Modeling Server for Contemporary Adaptive Hypermedia: An Evaluation of Push Approach to Evidence Propagation. In: C. Conati, K. McCoy and G. Paliouras (eds.) *Proceedings of 11th International Conference on User Modeling, UM 2007*, Corfu, Greece, 25-29 June, 2007, Springer Verlag, pp. 27-36. http://dx.doi.org/10.1007/978-3-540-73078-1_6
- Yudelson, M. (2010) *Providing Service-based Personalization in an Adaptive Hypermedia System*. Doctoral Dissertation, University of Pittsburgh.