

Seiji Isotani · Eva Millán ·  
Amy Ogan · Peter Hastings ·  
Bruce McLaren · Rose Luckin (Eds.)

LNAI 11626

# Artificial Intelligence in Education

20th International Conference, AIED 2019  
Chicago, IL, USA, June 25–29, 2019  
Proceedings, Part II

2  
Part II

AIED 2019



Springer

# Lecture Notes in Artificial Intelligence 11626

Subseries of Lecture Notes in Computer Science

## Series Editors

Randy Goebel

*University of Alberta, Edmonton, Canada*

Yuzuru Tanaka

*Hokkaido University, Sapporo, Japan*

Wolfgang Wahlster

*DFKI and Saarland University, Saarbrücken, Germany*

## Founding Editor

Jörg Siekmann

*DFKI and Saarland University, Saarbrücken, Germany*

More information about this series at <http://www.springer.com/series/1244>

Seiji Isotani · Eva Millán ·  
Amy Ogan · Peter Hastings ·  
Bruce McLaren · Rose Luckin (Eds.)

# Artificial Intelligence in Education

20th International Conference, AIED 2019  
Chicago, IL, USA, June 25–29, 2019  
Proceedings, Part II

*Editors*

Seiji Isotani 

University of Sao Paulo  
Sao Paulo, Brazil

Amy Ogan

Carnegie Mellon University  
Pittsburgh, PA, USA

Bruce McLaren

Carnegie Mellon University  
Pittsburgh, PA, USA

Eva Millán 

University of Malaga  
Málaga, Spain

Peter Hastings 

DePaul University  
Chicago, IL, USA

Rose Luckin

University College London  
London, UK

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Artificial Intelligence

ISBN 978-3-030-23206-1

ISBN 978-3-030-23207-8 (eBook)

<https://doi.org/10.1007/978-3-030-23207-8>

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

## Preface

The 20th International Conference on Artificial Intelligence in Education (AIED 2019) was held during June 25–29, 2019, in Chicago, USA. AIED 2019 was the latest in a longstanding series of now yearly international conferences for high-quality research in intelligent systems and cognitive science for educational applications.

The theme for the AIED 2019 conference was “Education for All in the XXI Century.” Inequity within and between countries continues to grow in the industrial age. Education that enables new economic opportunities plays a central role in addressing this problem. Support by intelligent information technologies have been proposed as a key mechanism for improving learning processes and outcomes, but may instead increase the digital divide if applied without reflection. The collective intelligence of the AIED community was convened to discuss critical questions, such as what the main barriers are to providing educational opportunities to underserved teachers and learners, how AI and advanced technologies can help overcome these difficulties, and how this work can be done ethically.

As in several previous years, the AIED 2019 events were co-located with a related community, the Learning at Scale (L@S 2019) conference. Both conferences shared a reception and a plenary invited talk by Candace Thille (Stanford University, USA). Also, three distinguished speakers gave plenary invited talks illustrating prospective directions for the field with an emphasis on accessibility, equity, and personalization: Jutta Treviranus (Ontario College of Art and Design University, Canada); Nancy Law (University of Hong Kong, SAR China); and Luis von Ahn (Carnegie Mellon University, USA).

There were 177 submissions as full papers to AIED 2019, of which 45 were accepted as long papers (ten pages) with oral presentation at the conference (for an acceptance rate of 25%), and 43 were accepted as short papers (four pages) with poster presentation at the conference. Of the 41 papers directly submitted as short papers, 15 were accepted. Apart from a few exceptions, each submission was reviewed by three Program Committee (PC) members. In addition, submissions underwent a discussion period (led by a leading reviewer) to ensure that all reviewers’ opinions would be considered and leveraged to generate a group recommendation to the program chairs. The program chairs checked the reviews and meta-reviews for quality and, where necessary, requested for reviewers to elaborate their review more constructively. Final decisions were made by carefully considering both meta-reviews (weighed more heavily) scores and the discussions. Our goal was to conduct a fair process and encourage substantive and constructive reviews without interfering with the reviewers’ judgment. We also took the constraints of the program into account, seeking to keep the acceptance rate within the typical range for this conference.

Beyond paper presentations and keynotes, the conference also included:

- A Doctoral Consortium Track that provided doctoral students with the opportunity to present their emerging and ongoing doctoral research at the conference and receive invaluable feedback from the research community.
- An Interactive Events session during which AIED attendees could experience first-hand new and emerging intelligent learning environments via interactive demonstrations.
- An Industry and Innovation Track, intended to support connections between industry (both for-profit and non-profit) and the research community.

The AIED 2019 conference also hosted ten half-day workshops with topics across a broad spectrum of societal issues, such as: life-long learning; educational data mining; multi-modal multi-channel data for self-regulated learning; ethics; informal learning; human-centered AI products design; standardization opportunities; team tutoring; intelligent textbooks and using AI to teach AI in K12 settings.

We especially wish to acknowledge the great efforts by our colleagues at DePaul University for hosting this year's conference.

Special thanks goes to Springer for sponsoring the AIED 2019 Best Paper Award and the AIED 2019 Best Student Paper Award. We also want to acknowledge the amazing work of the AIED 2019 Organizing Committee, the PC members, and the reviewers (listed herein), who with their enthusiastic contributions gave us invaluable support in putting this conference together.

May 2019

Seiji Isotani  
Eva Millán  
Amy Ogan  
Peter Hastings  
Bruce McLaren  
Rose Luckin

# Organization

## **General Conference Chairs**

Bruce M. McLaren  
Rose Luckin

Carnegie Mellon University, USA  
London Knowledge Lab, UK

## Program Chairs

Amy Ogan  
Eva Millán  
Seiji Isotani

Carnegie Mellon University, USA  
Universidad de Málaga, Spain  
University of São Paulo, Brazil

# Local Organization Chair

Peter Hastings

DePaul University, USA

## **Workshop and Tutorial Chairs**

Mingyu Feng  
Ma. Mercedes T. Rodrigo

WestEd, USA  
Ateneo de Manila University, Philippines

## **Industry and Innovation Track Chairs**

Elle Yuan Wang  
Ig Ibert Bittencourt

ASU EdPlus, USA  
Federal University of Alagoas, Brazil

## Doctoral Consortium Chairs

Janice Gobert  
Mutlu Cukurova

Rutgers Graduate School of Education, USA  
University College London Knowledge Lab, UK

## Poster Chairs

# Alexandra Cristea Natalia Stash

Durham University, UK  
Eindhoven University, The Netherlands

## Awards Chairs

Tanja Mitrovic  
Julita Vassileva

# **International Artificial Intelligence in Education Society Organization**

## President

Bruce M. McLaren Carnegie Mellon University, USA

## **Secretary/Treasurer**

Benedict du Boulay  
(Emeritus) University of Sussex, UK

## **Journal Editors**

Vincent Aleven Carnegie Mellon University, USA  
Judy Kay University of Sydney, Australia

## **Membership Chair**

Benjamin D. Nye University of Southern California, USA

## **Publicity Chair**

Erin Walker Arizona State University, USA

## **Finance Chair**

Vania Dimitrova University of Leeds, UK

## **Executive Committee**

Ryan S. J. d. Baker	University of Pennsylvania, USA
Tiffany Barnes	North Carolina State University, USA
Min Chi	North Carolina State University, USA
Cristina Conati	University of British Columbia, Canada
Ricardo Conejo	Universidad de Málaga, Spain
Sidney D'Mello	University of Notre Dame, USA
Vania Dimitrova	University of Leeds, UK
Neil Heffernan	Worcester Polytechnic Institute, USA
Diane Litman	University of Pittsburgh, USA
Rose Luckin	University College London, UK
Noboru Matsuda	Texas A&M University, USA
Manolis Mavrikis	University College London Knowledge Lab, UK

Tanja Mitrovic	University of Canterbury, New Zealand
Amy Ogan	Carnegie Mellon University, USA
Zachary Pardos	University of California, Berkeley, USA
Kaska Porayska-Pomsta	University College London, UK
Ido Roll	University of British Columbia, Canada
Carolyn Penstein Rosé	Carnegie Mellon University, USA
Julita Vassileva	University of Saskatchewan, Canada
Erin Walker	Arizona State University, USA
Kalina Yacef	University of Sydney, Australia

## Program Committee

Esma Aimeur	University of Montreal, Canada
Patricia Albacete	University of Pittsburgh, USA
Vincent Aleven	Human-Computer Interaction Institute, Carnegie Mellon University, USA
Ivon Arroyo	Worcester Polytechnic Institute, USA
Nilufar Baghaei	OPAIC, New Zealand
Ryan Baker	University of Pennsylvania, USA
Gautam Biswas	Vanderbilt University, USA
Ig Ibert Bittencourt	Federal University of Alagoas, Brazil
Emmanuel Blanchard	IDU Interactive Inc., Canada
Nigel Bosch	University of Illinois Urbana-Champaign, USA
Jesus G. Boticario	UNED, Spain
Kristy Elizabeth Boyer	University of Florida, USA
Bert Bredeweg	University of Amsterdam, Netherlands
Christopher Brooks	University of Michigan, USA
Geiser Chalco Challco	ICMC/USP, Brasil
Maiga Chang	Athabasca University, Canada
Mohamed Amine Chatti	University of Duisburg-Essen, Germany
Min Chi	NC State University, USA
Andrew Clayphan	The University of Sydney, Australia
Cristina Conati	The University of British Columbia, Canada
Mark G. Core	University of Southern California, USA
Scotty Craig	Arizona State University, Polytechnic, USA
Mutlu Cukurova	University College London, UK
Ben Daniel	University of Otago, New Zealand
Diego Dermeval	Federal University of Alagoas, Brazil
Tejas Dhamecha	IBM, India
Barbara Di Eugenio	University of Illinois at Chicago, USA
Daniele Di Mitri	Open Universiteit, Netherlands
Vania Dimitrova	University of Leeds, UK
Peter Dolog	Aalborg University, Denmark
Fabiano Dorça	Universidade Federal de Uberlândia, Brazil
Mingyu Feng	SRI International, USA
Rafael Ferreira	Federal Rural University of Pernambuco, Brazil

Carol Forsyth	Educational Testing Service, USA
Davide Fossati	Emory University, USA
Reva Freedman	Northern Illinois University, USA
Dragan Gasevic	Monash University, Australia
Isabela Gasparini	UDESC, Brazil
Elena Gaudioso	UNED, Spain
Janice Gobert	Rutgers University, USA
Ashok Goel	Georgia Institute of Technology, USA
Ilya Goldin	2U, Inc., USA
Alex Sandro Gomes	Universidade Federal de Pernambuco, Brazil
Art Graesser	University of Memphis, USA
Monique Grandbastien	LORIA, Université de Lorraine, France
Gahgene Gweon	Seoul National University, South Korea
Jason Harley	University of Alberta, Canada
Andreas Harrer	University of Applied Sciences and Arts Dortmund, Germany
Peter Hastings	DePaul University, USA
Yuki Hayashi	Osaka Prefecture University, Japan
Tobias Hecking	University of Duisburg-Essen, Germany
Neil Heffernan	Worcester Polytechnic Institute, USA
Tsukasa Hirashima	Hiroshima University, Japan
Ulrich Hoppe	University Duisburg-Essen, Germany
Sharon Hsiao	Arizona State University, USA
Paul Salvador Inventado	California State University Fullerton, USA
Seiji Isotani	University of São Paulo, Brazil
Sridhar Iyer	IIT Bombay, India
G. Tanner Jackson	Educational Testing Service, USA
Patricia Jaques	UNISINOS, Brazil
Srecko Joksimovic	Teaching Innovation Unit and School of Education, University of South Australia, Australia
Pamela Jordan	University of Pittsburgh, USA
Sandra Katz	University of Pittsburgh, USA
Judy Kay	The University of Sydney, Australia
Fazel Keshtkar	St. John's University, USA
Simon Knight	University of Technology Sydney, Australia
Tomoko Kojiri	Kansai University, Japan
Amruth Kumar	Ramapo College of New Jersey, USA
Rohit Kumar	Raytheon BBN Technologies, Cambridge, MA, USA
Jean-Marc Labat	Université Paris 6, France
Sébastien Lallé	The University of British Columbia, Canada
H. Chad Lane	University of Illinois at Urbana-Champaign, USA
Nguyen-Thinh Le	Humboldt Universität zu Berlin, Germany
Blair Lehman	Educational Testing Service, USA
James Lester	North Carolina State University, USA
Chee-Kit Looi	National Institute of Education, Singapore
Yu Lu	Beijing Normal University, China

Vanda Luengo	LIP6, Sorbonne Université, France
Collin Lynch	North Carolina State University, USA
Leonardo Brandão Marques	University of São Paulo, Brazil
Roberto	University of Technology Sydney, Australia
Martinez-Maldonado	
Smit Marvaniya	IBM, India
Eleandro Maschio	Universidade Tecnológica Federal do Paraná, Brazil
Noboru Matsuda	North Carolina State University, USA
Manolis Mavrikis	London Knowledge Lab, UK
Gordon McCalla	University of Saskatchewan, Canada
Agathe Merceron	Beuth University of Applied Sciences Berlin, Germany
Eva Millán	Universidad de Málaga, Spain
Marcelo Milrad	Linnaeus University, Sweden
Ritayan Mitra	IIT Bombay, India
Tanja Mitrovic	University of Canterbury, Christchurch, New Zealand
Kazuhisa Miwa	Nagoya University, Japan
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology, Japan
Kasia Muldner	Carleton University, Canada
Roger Nkambou	Université du Québec À Montréal (UQAM), Canada
Amy Ogan	Carnegie Mellon University, USA
Hiroaki Ogata	Kyoto University, Japan
Andrew Olney	University of Memphis, USA
Jennifer Olsen	Ecole Polytechnique Fédérale de Lausanne, Switzerland
Helen Pain	The University of Edinburgh, UK
Ranilson Paiva	Universidade Federal de Alagoas, Brazil
Luc Paquette	University of Illinois at Urbana-Champaign, USA
Abelardo Pardo	University of South Australia, Australia
Zach Pardos	University of California, Berkeley, USA
Philip I. Pavlik Jr.	University of Memphis, USA
Radek Pelánek	Masaryk University Brno, Czechia
Niels Pinkwart	Humboldt-Universität zu Berlin, Germany
Elvira Popescu	University of Craiova, Romania
Kaska Porayska-Pomsta	UCL Knowledge Lab, UK
Anna Rafferty	Carleton College, USA
Martina Rau	University of Wisconsin - Madison, USA
Ma. Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Ido Roll	The University of British Columbia, Canada
Rod Roscoe	Arizona State University, USA
Jonathan Rowe	North Carolina State University, USA
José A. Ruipérez Valiente	Massachusetts Institute of Technology, USA
Nikol Rummel	Ruhr-Universität Bochum, Germany
Vasile Rus	The University of Memphis, USA
Demetrios Sampson	Curtin University, Australia
Olga C. Santos	UNED, Spain

Kazuhisa Seta	Osaka Prefecture University, Japan
Lei Shi	University of Liverpool, UK
Sergey Sosnovsky	Utrecht University, Netherlands
Pierre Tchounikine	University of Grenoble, France
Maomi Ueno	The University of Electro-Communications DFKI, Japan
Carsten Ullrich	GmbH, Germany
Kurt Vanlehn	Arizona State University, USA
Julita Vassileva	University of Saskatchewan, Canada
Felisa Verdejo	UNED, Spain
Rosa Vicari	Universidade Federal do Rio Grande do Sul, Brazil
Erin Walker	Arizona State University, USA
Elle Wang	Arizona State University, USA
John Whitmer	Blackboard, Inc., USA
Beverly Park Woolf	University of Massachusetts, USA
Marcelo Worsley	Northwestern University, USA
Kalina Yacef	The University of Sydney, Australia
Elle Yuan Wang	Arizona State University, USA
Diego Zapata-Rivera	Educational Testing Service, USA
Jingjing Zhang	Beijing Normal University, China
Gustavo Zurita	Universidad de Chile, Chile

## Additional Reviewers

Afzal, Shazia	Gitinabard, Niki
Anaya, Antonio R.	Harrison, Avery
Andrews-Todd, Jessica	Hartmann, Christian
Arevalillo-Herráez, Miguel	Hayashi, Yusuke
Arroyo, Ivon	Herder, Tiffany
Botelho, Anthony F.	Horiguchi, Tomoya
Chavan, Pankaj	Hulse, Taylyn
Chen, Chen	Hutchins, Nicole
Chen, Penghe	Hutt, Stephen
Choi, Heeryung	Ishola, Oluwabukola
Cochran, Keith	Ju, Song
D'Mello, Sidney	Kay, Judy
Deep, Anurag	Kent, Carmel
Deitelhoff, Fabian	Kojima, Kazuaki
Doberstein, Dorian	Landers, Richard
du Boulay, Benedict	Lawson, Marylynne
Erickson, John	Lelei, David Edgar
Galafassi, Cristiano	Lin, Tao Roa
Gaweda, Adam	Madaio, Michael
Gerritsen, David	Maehigashi, Akihiro

- Malkiewich, Laura  
Mao, Ye  
Matsumuro, Miki  
Mavrikis, Manolis  
McBroom, Jessica  
McNamara, Danielle  
Memon, Muhammad Qasim  
Minn, Sein  
Mishra, Shitanshu  
Mittal, Anant  
Molenaar, Inge  
Morita, Junya  
Munshi, Anabil  
Negi, Shivsevak  
Nikolayeva, Iryna  
Oertel, Catharine  
Okoilu, Ruth  
Patikorn, Thanaporn  
Praharaj, Sambit  
Rajendran, Ramkumar  
Rodriguez, Fernando  
Saha, Swarnadeep  
Shahriar, Tasmia  
Shen, Shitian  
Shimmei, Machi  
Smith, Hannah  
Smith, Karl  
Snyder, Caitlin  
Stewart, Angela  
Strauss, Sebastian  
Sánchez-Elvira Paniagua, Angeles  
Tan, Hongye  
Thompson, Craig  
Toda, Armando  
Tomoto, Takahito  
Tsan, Jennifer  
Vanlehn, Kurt  
Wang, April  
Wiggins, Joseph  
Yamamoto, Sho  
Yang, Xi  
Yett, Bernard  
Yi, Sherry  
Ying, Kimberly  
Yokoyama, Mai  
Zhang, Ningyu  
Zhou, Guojing

## Contents – Part II

### Short Papers (Posters)

Model-Based Characterization of Text Discourse Content to Evaluate Online Group Collaboration . . . . .	3
<i>Adetunji Adeniran, Judith Masthoff, and Nigel Beacham</i>	
Identifying Editor Roles in Argumentative Writing from Student Revision Histories . . . . .	9
<i>Tazin Afrin and Diane Litman</i>	
Degree Curriculum Contraction: A Vector Space Approach . . . . .	14
<i>Mohamed Alkaoud and Zachary A. Pardos</i>	
L2 Learners' Preferences of Dialogue Agents: A Key to Achieve Adaptive Motivational Support? . . . . .	19
<i>Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta</i>	
Eye Gaze Sequence Analysis to Model Memory in E-education . . . . .	24
<i>Maël Beugut, Sylvain Castagnos, Christophe Luxembourger, and Anne Boyer</i>	
What Inquiry with Virtual Labs Can Learn from Productive Failure: A Theory-Driven Study of Students' Reflections . . . . .	30
<i>Charleen Brand, Jonathan Massey-Allard, Sarah Perez, Nikol Rummel, and Ido Roll</i>	
The Role of Achievement Goal Orientation on Metacognitive Process Use in Game-Based Learning . . . . .	36
<i>Elizabeth B. Cloude, Michelle Taub, James Lester, and Roger Azevedo</i>	
Autoencoders for Educational Assessment . . . . .	41
<i>Geoffrey Converse, Mariana Curi, and Suely Oliveira</i>	
The Value of Multimodal Data in Classification of Social and Emotional Aspects of Tutoring . . . . .	46
<i>Mutlu Cukurova, Carmel Kent, and Rosemary Luckin</i>	
Conscientiousness, Honesty-Humility, and Analogical/Creative Reasoning: Implications for Instructional Designs in Intelligent Tutoring Systems . . . . .	52
<i>Jeanine A. DeFalco, Anne M. Sinatra, Elizabeth Rodriguez, and R. Stan Hum</i>	

Learners' Gaze Behaviors and Metacognitive Judgments with an Agent-Based Multimedia Environment . . . . .	58
<i>Daryn A. Dever, Megan Wiedbusch, and Roger Azevedo</i>	
Online Assessment of Belief Biases and Their Impact on the Acceptance of Fallacious Reasoning . . . . .	62
<i>Nicholas Diana, John Stamper, and Kenneth Koedinger</i>	
Early Dropout Prediction for Programming Courses Supported by Online Judges . . . . .	67
<i>Filipe D. Pereira, Elaine Oliveira, Alexandra Cristea, David Fernandes, Luciano Silva, Gene Aguiar, Ahmed Alamri, and Mohammad Alshehri</i>	
Developing a Deep Learning-Based Affect Recognition System for Young Children . . . . .	73
<i>Amir Hossein Farzaneh, Yanghee Kim, Mengxi Zhou, and Xiaojun Qi</i>	
Using Exploratory Data Analysis to Support Implementation and Improvement of Education Technology Product . . . . .	79
<i>Mingyu Feng, Daniel Brenner, and Andrew Coulson</i>	
Bayesian Diagnosis Tracing: Application of Procedural Misconceptions in Knowledge Tracing . . . . .	84
<i>Junchen Feng, Bo Zhang, Yuchen Li, and Qiushi Xu</i>	
Analysis of Gamification Elements. A Case Study in a Computer Science Course . . . . .	89
<i>Miguel García Iruela, Manuel J. Fonseca, Raquel Hijón Neira, and Teresa Chambel</i>	
Towards Adaptive Worked-Out Examples in an Intelligent Tutoring System . . . . .	94
<i>Nicholas Green, Barbara Di Eugenio, and Davide Fossati</i>	
Orchestrating Class Discussion with Collaborative Kit-Build Concept Mapping . . . . .	100
<i>Yusuke Hayashi, Toshihiro Nomura, and Tsukasa Hirashima</i>	
Automating the Categorization of Learning Activities, to Help Improve Learning Design . . . . .	105
<i>Wayne Holmes and Juliette Culver</i>	
Identifying the Structure of Students' Explanatory Essays . . . . .	110
<i>Simon Hughes, Peter Hastings, and M. Anne Britt</i>	

A Systematic Approach for Analyzing Students' Computational Modeling Processes in C2STEM . . . . .	116
<i>Nicole Hutchins, Gautam Biswas, Shuchi Grover, Satabdi Basu, and Caitlin Snyder</i>	
Intelligent Tutoring System for Negotiation Skills Training . . . . .	122
<i>Emmanuel Johnson, Gale Lucas, Peter Kim, and Jonathan Gratch</i>	
Robot Lecture for Enhancing Non-verbal Behavior in Lecture . . . . .	128
<i>Akihiro Kashihara, Tatsuya Ishino, and Mitsuhiro Goto</i>	
Design Prompts for Virtual Reality in Education. . . . .	133
<i>Lawrence Kizilkaya, David Vince, and Wayne Holmes</i>	
Assessing and Improving Learning Outcomes for Power Management Experiments Using Cognitive Graph . . . . .	138
<i>Yi Kuang, Bin Duan, Shuyang Zhong, and Mengping Lv</i>	
Does Choosing the Concept on Which to Solve Each Practice Problem in an Adaptive Tutor Affect Learning? . . . . .	143
<i>Amruth N. Kumar</i>	
Measuring Content Complexity of Technical Texts: Machine Learning Experiments . . . . .	148
<i>M. Zakaria Kurdi</i>	
Should Students Use Digital Scratchpads? Impact of Using a Digital Assistive Tool on Arithmetic Problem-Solving . . . . .	153
<i>Minji Kwak and Gahgene Gweon</i>	
What Does Time Tell? Tracing the Forgetting Curve Using Deep Knowledge Tracing . . . . .	158
<i>Amar Lalwani and Sweety Agrawal</i>	
Evaluating the Transfer of Scaffolded Inquiry: What Sticks and Does It Last? . . . . .	163
<i>Haiying Li, Janice Gobert, and Rachel Dickler</i>	
Automatic Short Answer Grading via Multiway Attention Networks . . . . .	169
<i>Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu</i>	
Automatic Classification of Error Types in Solutions to Programming Assignments at Online Learning Platform. . . . .	174
<i>Artyom Lobanov, Timofey Bryksin, and Alexey Shpilman</i>	

Using Recurrent Neural Networks to Build a Stopping Algorithm for an Adaptive Assessment . . . . .	179
<i>Jeffrey Matayoshi, Eric Cosyn, and Hasan Uzun</i>	
Participatory Design to Lower the Threshold for Intelligent Support Authoring . . . . .	185
<i>Manolis Mavrikis, Sokratis Karkalias, Mutlu Cukurova, and Emmanouela Papapetrou</i>	
Finding Relevant e-Learning Materials . . . . .	190
<i>Blessing Mbipom</i>	
Predicting Dialogue Breakdown in Conversational Pedagogical Agents with Multimodal LSTMs . . . . .	195
<i>Wookhee Min, Kyungjin Park, Joseph Wiggins, Bradford Mott, Eric Wiebe, Kristy Elizabeth Boyer, and James Lester</i>	
Pique: Recommending a Personalized Sequence of Research Papers to Engage Student Curiosity . . . . .	201
<i>Maryam Mohseni, Mary Lou Maher, Kazjon Grace, Nadia Najjar, Fakhri Abbas, and Omar Eltayeby</i>	
Group Formation for Collaborative Learning: A Systematic Literature Review . . . . .	206
<i>Chinasa Odo, Judith Masthoff, and Nigel Beacham</i>	
AI Meets Austen: Towards Human-Robot Discussions of Literary Metaphor . . . . .	213
<i>Natalie Parde and Rodney D. Nielsen</i>	
Discovery of Study Patterns that Impacts Students' Discussion Performance in Forum Assignments . . . . .	220
<i>Bruno Elias Penteado, Seiji Isotani, Paula Maria Pereira Paiva, Marina Morettin-Zupelari, and Deborah Viviane Ferrari</i>	
Automatic Construction of a Phonics Curriculum for Reading Education Using the Transformer Neural Network . . . . .	226
<i>Cassandra Potier Watkins, Olivier Dehaene, and Stanislas Dehaene</i>	
An Annotation Protocol for Collecting User-Generated Counter-Arguments Using Crowdsourcing . . . . .	232
<i>Paul Reisert, Gisela Vallejo, Naoya Inoue, Iryna Gurevych, and Kentaro Inui</i>	
Towards an Automatic Q&A Generation for Online Courses - A Pipeline Based Approach . . . . .	237
<i>Sylvio Rüdian and Niels Pinkwart</i>	

Semantic Matching of Open Texts to Pre-scripted Answers in Dialogue-Based Learning . . . . .	242
<i>Ştefan Ruşeti, Raja Lala, Gabriel Guțu-Robu, Mihai Dascălu, Johan Jeuring, and Marcell van Geest</i>	
Developing Game-Based Models of Cooperation, Persistence and Problem Solving from Collaborative Gameplay . . . . .	247
<i>Maria Ofelia Z. San Pedro, Ruitao Liu, and Tamera L. McKinniss</i>	
An Intelligent-Agent Facilitated Scaffold for Fostering Reflection in a Team-Based Project Course . . . . .	252
<i>Sreecharan Sankaranarayanan, Xu Wang, Cameron Dashti, Marshall An, Clarence Ngoh, Michael Hilton, Majd Sakr, and Carolyn Rosé</i>	
I Wanna Talk Like You: Speaker Adaptation to Dialogue Style in L2 Practice Conversation . . . . .	257
<i>Arabella J. Sinclair, Rafael Ferreira, Dragan Gašević, Christopher G. Lucas, and Adam Lopez</i>	
Understanding Students' Model Building Strategies Through Discourse Analysis . . . . .	263
<i>Caitlin Snyder, Nicole Hutchins, Gautam Biswas, and Shuchi Grover</i>	
Exploring Teachable Humans and Teachable Agents: Human Strategies Versus Agent Policies and the Basis of Expertise . . . . .	269
<i>John Stamper and Steven Moore</i>	
Learning from Videos Showing a Dialog Fosters More Positive Affect Than Learning from a Monolog . . . . .	275
<i>Samantha Stranc and Kasia Muldner</i>	
Automated Feedback on the Structure of Hypothesis Tests . . . . .	281
<i>Sietske Tacoma, Bastiaan Heeren, Johan Jeuring, and Paul Drijvers</i>	
Informing the Utility of Learning Interventions: Investigating Factors Related to Students' Academic Achievement in Classroom and Online Courses . . . . .	286
<i>Anna-Lena Theus and Kasia Muldner</i>	
Auto-Sending Messages in an Intelligent Orchestration System: A Pilot Study . . . . .	292
<i>Kurt VanLehn, Salman Cheema, Seokmin Kang, and Jon Wetzel</i>	
Adaptive Learning Material Recommendation in Online Language Education . . . . .	298
<i>Shuhan Wang, Hao Wu, Ji Hun Kim, and Erik Andersen</i>	

Deep Knowledge Tracing with Side Information . . . . .	303
<i>Zhiwei Wang, Xiaoqin Feng, Jiliang Tang, Gale Yan Huang,     and Zitao Liu</i>	
Analysis of Holistic Interactions Between Lecturers and Students in Lectures . . . . .	309
<i>Eiji Watanabe, Takashi Ozeki, and Takeshi Kohama</i>	
Take the Initiative: Mixed Initiative Dialogue Policies for Pedagogical Agents in Game-Based Learning Environments . . . . .	314
<i>Joseph B. Wiggins, Mayank Kulkarni, Wookhee Min,     Kristy Elizabeth Boyer, Bradford Mott, Eric Wiebe, and James Lester</i>	
Investigating on Discussion for Sharing Understanding by Using Reciprocal Kit-Build Concept Map . . . . .	319
<i>Warunya Wunnasri, Jaruwat Pailai, Yusuke Hayashi,     and Tsukasa Hirashima</i>	
<b>Doctoral Consortium</b>	
Detection of Collaboration: Relationship Between Log and Speech-Based Classification . . . . .	327
<i>Sree Aurovindh Viswanathan and Kurt Vanlehn</i>	
An Intelligent Tutoring System and Teacher Dashboard to Support Mathematizing During Science Inquiry . . . . .	332
<i>Rachel Dickler</i>	
Towards Adaptive Hour of Code . . . . .	339
<i>Tomáš Effenberger</i>	
Leaving No One Behind: Educating Those Most Impacted by Artificial Intelligence . . . . .	344
<i>Laura Gemmell, Lucy Wenham, and Sabine Hauert</i>	
Modeling Students' Behavior Using Sequential Patterns to Predict Their Performance . . . . .	350
<i>Mehrdad Mirzaei and Shaghayegh Sahebi</i>	
Personalization in OELEs: Developing a Data-Driven Framework to Model and Scaffold SRL Processes . . . . .	354
<i>Anabil Munshi and Gautam Biswas</i>	
Analyzing Engagement in an On-Line Session . . . . .	359
<i>Vandana Naik and Venkatesh Kamat</i>	
A Machine Learning Grading System Using Chatbots . . . . .	365
<i>Ifeanyi G. Ndukwue, Ben K. Daniel, and Chukwudi E. Amadi</i>	

Evidence-Based Recommendation for Content Improvement Using Reinforcement Learning . . . . .	369
<i>Machi Shimmei and Noboru Matsuda</i>	
A Virtual Counselor for Genetic Risk Communication. . . . .	374
<i>Shuo Zhou and Timothy Bickmore</i>	
<b>Industry Papers</b>	
A Multimodal Alerting System for Online Class Quality Assurance. . . . .	381
<i>Jiahao Chen, Hang Li, Wenxin Wang, Wenbiao Ding, Gale Yan Huang, and Zitao Liu</i>	
Leveraging Cognitive Science and Artificial Intelligence to Save Lives . . . . .	386
<i>Matthew Jensen Hays, Aaron Richard Glick, and H. Chad Lane</i>	
A Task-Oriented Dialogue System for Moral Education . . . . .	392
<i>Yan Peng, Penghe Chen, Yu Lu, Qinggang Meng, Qi Xu, and Shengquan Yu</i>	
Leveraging Student Self-reports to Predict Learning Outcomes . . . . .	398
<i>Shaveen Singh</i>	
Toward a Scalable Learning Analytics Solution . . . . .	404
<i>Josine Verhagen, David Hatfield, and Dylan Arena</i>	
Motivating Students to Ask More Questions. . . . .	409
<i>Yuan Wang, Turner Bohlen, Linda Elkins-Tanton, and James Tanton</i>	
Towards Helping Teachers Select Optimal Content for Students . . . . .	413
<i>Xiaotian Zou, Wei Ma, Zhenjun Ma, and Ryan S. Baker</i>	
<b>Workshop Papers</b>	
Supporting Lifelong Learning . . . . .	421
<i>Oluwabunmi (Adewoyin) Olakanmi, Oluwabukola Mayowa Ishola, Gord McCalla, Ifeoma Adaji, and Francisco J. Gutierrez</i>	
Educational Data Mining in Computer Science Education (CSEDM) . . . . .	422
<i>David Azcona, Yancy Vance Paredes, Thomas W. Price, and Sharon I-Han Hsiao</i>	
Measuring, Analyzing, and Modeling Multimodal Multichannel Data for Supporting Self-regulated Learning by Making Systems More Intelligent for All in the 21st Century . . . . .	423
<i>Roger Azevedo and Gautam Biswas</i>	

Ethics in AJED: Who Cares? . . . . .	424
<i>Wayne Holmes, Duygu Bektik, Maria Di Gennaro, Beverly Park Woolf,     and Rose Luckin</i>	
Adaptive and Intelligent Technologies for Informal Learning . . . . .	426
<i>H. Chad Lane, Jonathan Rowe, Stephen Blessing, and Nesra Yannier</i>	
Designing Human-Centered AI Products . . . . .	428
<i>Kristen Olson, Maysam Moussalem, Di Dang, Kristie J. Fisher,     Jess Holbrook, and Rebecca Salois</i>	
Standardization Opportunities for AI in Education. . . . .	429
<i>Robby Robson, Richard Tong, Robert Sottilare, and K. P. Thai</i>	
Approaches and Challenges in Team Tutoring Workshop. . . . .	430
<i>Anne M. Sinatra and Jeanine A. DeFalco</i>	
Intelligent Textbooks . . . . .	431
<i>Sergey Sosnovsky, Peter Brusilovsky, Rakesh Agrawal,     Richard G. Baraniuk, and Andrew S. Lan</i>	
K12 Artificial Intelligence Education . . . . .	433
<i>Ning Wang and James Lester</i>	
<b>Author Index . . . . .</b>	<b>435</b>

## Contents – Part I

Towards the Identification of Propaedeutic Relations in Textbooks . . . . .	1
<i>Giovanni Adorni, Chiara Alzetta, Frosina Koceva,     Samuele Passalacqua, and Ilaria Torre</i>	
Investigating Help-Giving Behavior in a Cross-Platform Learning Environment . . . . .	14
<i>Ishrat Ahmed, Areej Mawasi, Shang Wang, Ruth Wylie, Yoav Bergner,     Amanda Whitehurst, and Erin Walker</i>	
Predicting Academic Performance: A Bootstrapping Approach for Learning Dynamic Bayesian Networks . . . . .	26
<i>Mashael Al-Luhaybi, Leila Yousefi, Stephen Swift, Steve Counsell,     and Allan Tucker</i>	
The Impact of Student Model Updates on Contingent Scaffolding in a Natural-Language Tutoring System . . . . .	37
<i>Patricia Albacete, Pamela Jordan, Sandra Katz,     Irene-Angelica Chounta, and Bruce M. McLaren</i>	
Item Ordering Biases in Educational Data . . . . .	48
<i>Jaroslav Čechák and Radek Pelánek</i>	
A Comparative Study on Question-Worthy Sentence Selection Strategies for Educational Question Generation . . . . .	59
<i>Guanliang Chen, Jie Yang, and Dragan Gasevic</i>	
Effect of Discrete and Continuous Parameter Variation on Difficulty in Automatic Item Generation . . . . .	71
<i>Binglin Chen, Craig Zilles, Matthew West, and Timothy Bretl</i>	
Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools . . . . .	84
<i>Scott A. Crossley, Minkyung Kim, Laura Allen, and Danielle McNamara</i>	
The Importance of Automated Real-Time Performance Feedback in Virtual Reality Temporal Bone Surgery Training . . . . .	96
<i>Myles Davaris, Sudanthi Wijewickrema, Yun Zhou, Patorn Piromchai,     James Bailey, Gregor Kennedy, and Stephen O'Leary</i>	
Autonomy and Types of Informational Text Presentations in Game-Based Learning Environments . . . . .	110
<i>Daryn A. Dever and Roger Azevedo</i>	

Examining Gaze Behaviors and Metacognitive Judgments of Informational Text Within Game-Based Learning Environments . . . . .	121
<i>Daryn A. Dever and Roger Azevedo</i>	
Using “Idealized Peers” for Automated Evaluation of Student Understanding in an Introductory Psychology Course . . . . .	133
<i>Tricia A. Guerrero and Jennifer Wiley</i>	
4D Affect Detection: Improving Frustration Detection in Game-Based Learning with Posture-Based Temporal Data Fusion . . . . .	144
<i>Nathan L. Henderson, Jonathan P. Rowe, Bradford W. Mott, Keith Brawner, Ryan Baker, and James C. Lester</i>	
Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms . . . . .	157
<i>Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven</i>	
The Case of Self-transitions in Affective Dynamics . . . . .	172
<i>Shamya Karumbaiah, Ryan S. Baker, and Jaclyn Ocumpaugh</i>	
How Many Times Should a Pedagogical Agent Simulation Model Be Run? . . . . .	182
<i>David Edgar Kiprop Lelei and Gordon McCalla</i>	
A Survey of the General Public’s Views on the Ethics of Using AI in Education . . . . .	194
<i>Annabel Latham and Sean Goltz</i>	
Promoting Inclusivity Through Time-Dynamic Discourse Analysis in Digitally-Mediated Collaborative Learning . . . . .	207
<i>Nia Dowell, Yiwen Lin, Andrew Godfrey, and Christopher Brooks</i>	
Evaluating Machine Learning Approaches to Classify Pharmacy Students’ Reflective Statements. . . . .	220
<i>Ming Liu, Simon Buckingham Shum, Efi Mantzourani, and Cherie Lucas</i>	
Comfort with Robots Influences Rapport with a Social, Entrainig Teachable Robot. . . . .	231
<i>Nichola Lubold, Erin Walker, Heather Pon-Barry, and Amy Ogan</i>	
A Concept Map Based Assessment of Free Student Answers in Tutorial Dialogues. . . . .	244
<i>Nabin Maharjan and Vasile Rus</i>	
Deep (Un)Learning: Using Neural Networks to Model Retention and Forgetting in an Adaptive Learning System . . . . .	258
<i>Jeffrey Matayoshi, Hasan Uzun, and Eric Cosyn</i>	

Checking It Twice: Does Adding Spelling and Grammar Checkers Improve Essay Quality in an Automated Writing Tutor? . . . . .	270
<i>Kathryn S. McCarthy, Rod D. Roscoe, Aaron D. Likens, and Danielle S. McNamara</i>	
What's Most Broken? Design and Evaluation of a Tool to Guide Improvement of an Intelligent Tutor . . . . .	283
<i>Shiven Mian, Mononito Goswami, and Jack Mostow</i>	
Reducing Mind-Wandering During Vicarious Learning from an Intelligent Tutoring System . . . . .	296
<i>Caitlin Mills, Nigel Bosch, Kristina Krasich, and Sidney K. D'Mello</i>	
Annotated Examples and Parameterized Exercises: Analyzing Students' Behavior Patterns . . . . .	308
<i>Mehrdad Mirzaei, Shaghayegh Sahebi, and Peter Brusilovsky</i>	
Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching . . . . .	320
<i>Antonija Mitrovic, Matthew Gordon, Alicja Piotrkowicz, and Vania Dimitrova</i>	
Behavioural Cloning of Teachers for Automatic Homework Selection . . . . .	333
<i>Russell Moore, Andrew Caines, Andrew Rice, and Paula Buttery</i>	
Integrating Students' Behavioral Signals and Academic Profiles in Early Warning System . . . . .	345
<i>SungJin Nam and Perry Samson</i>	
<i>Predicting Multi-document Comprehension: Cohesion Network Analysis . . . . .</i>	358
<i>Bogdan Nicula, Cecile A. Perret, Mihai Dascalu, and Danielle S. McNamara</i>	
Student Network Analysis: A Novel Way to Predict Delayed Graduation in Higher Education . . . . .	370
<i>Nasheen Nur, Noseong Park, Mohsen Dorodchi, Wenwen Dou, Mohammad Javad Mahzoon, Xi Niu, and Mary Lou Maher</i>	
Automatic Generation of Problems and Explanations for an Intelligent Algebra Tutor . . . . .	383
<i>Eleanor O'Rourke, Eric Butler, Armando Díaz Tolentino, and Zoran Popović</i>	
Generalizability of Methods for Imputing Mathematical Skills Needed to Solve Problems from Texts. . . . .	396
<i>Thanaporn Patikorn, David Deisadze, Leo Grande, Ziyang Yu, and Neil Heffernan</i>	

Using Machine Learning to Overcome the Expert Blind Spot for Perceptual Fluency Trainings . . . . .	406
<i>Martina A. Rau, Ayon Sen, and Xiaojin Zhu</i>	
Disentangling Conceptual and Embodied Mechanisms for Learning with Virtual and Physical Representations . . . . .	419
<i>Martina A. Rau and Tara A. Schmidt</i>	
Adaptive Support for Representation Skills in a Chemistry ITS Is More Effective Than Static Support . . . . .	432
<i>Martina A. Rau, Miranda Zahn, Edward Misback, and Judith Burstyn</i>	
Confusion in Learning from Erroneous Examples: Does Type of Prompted Self-explanation Make a Difference? . . . . .	445
<i>J. Elizabeth Richey, Bruce M. McLaren, Miguel Andres-Bray, Michael Mogessie, Richard Scruggs, Ryan Baker, and Jon Star</i>	
Modeling Collaboration in Online Conversations Using Time Series Analysis and Dialogism . . . . .	458
<i>Robert-Florian Samoilescu, Mihai Dascalu, Maria-Dorinela Sirbu, Stefan Trausan-Matu, and Scott A. Crossley</i>	
Improving Short Answer Grading Using Transformer-Based Pre-training . . . . .	469
<i>Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi</i>	
Uniform Adaptive Testing Using Maximum Clique Algorithm . . . . .	482
<i>Maomi Ueno and Yoshimitsu Miyazawa</i>	
Rater-Effect IRT Model Integrating Supervised LDA for Accurate Measurement of Essay Writing Ability . . . . .	494
<i>Masaki Uto</i>	
Collaboration Detection that Preserves Privacy of Students' Speech . . . . .	507
<i>Sree Aurovindh Viswanathan and Kurt VanLehn</i>	
How Does Order of Gameplay Impact Learning and Enjoyment in a Digital Learning Game? . . . . .	518
<i>Yeyu Wang, Huy Nguyen, Erik Harpstead, John Stamper, and Bruce M. McLaren</i>	
Analyzing Students' Design Solutions in an NGSS-Aligned Earth Sciences Curriculum . . . . .	532
<i>Ningyu Zhang, Gautam Biswas, Jennifer L. Chiu, and Kevin W. McElhaney</i>	

Hierarchical Reinforcement Learning for Pedagogical Policy Induction . . . . .	544
<i>Guojing Zhou, Hamoon Azizsoltani, Markel Sanz Ausin, Tiffany Barnes,     and Min Chi</i>	
<b>Author Index</b> . . . . .	557

## **Short Papers (Posters)**



# Model-Based Characterization of Text Discourse Content to Evaluate Online Group Collaboration

Adetunji Adeniran<sup>1</sup>(✉), Judith Masthoff<sup>2</sup>, and Nigel Beacham<sup>1</sup>

<sup>1</sup> University of Aberdeen, Aberdeen, UK  
r01aba17@abdn.ac.uk

<sup>2</sup> Utrecht University, Utrecht, The Netherlands  
j.f.m.masthoff@uu.nl

**Abstract.** This paper presents a model that characterizes textual discourse contents of online groups and provides a visualization of the level of collaboration within groups. This approach is envisioned to provide an insight into a real-time intervention to scaffold collaboration within online learning groups.

**Keywords:** Joint problem-solving · Discourse content · Online groups

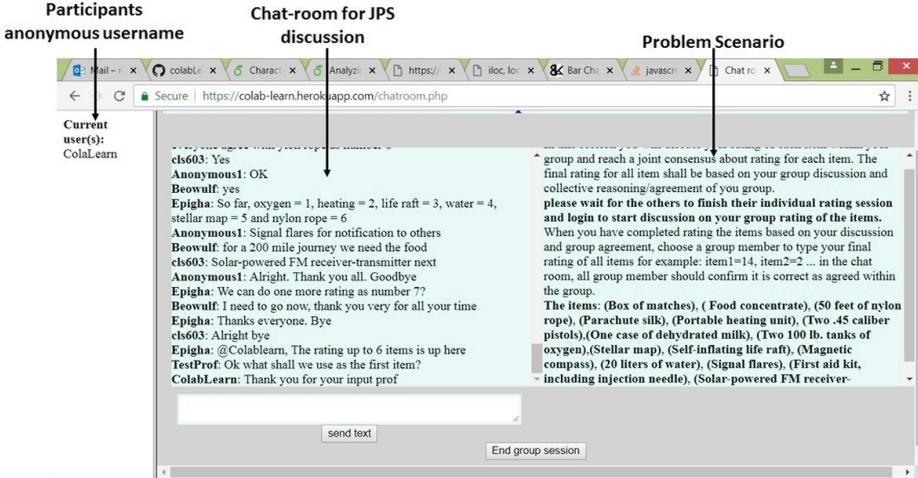
## 1 Introduction and Related Work

Online group learning involves virtual access to education without limitation of geographical location and a collaborative environment that provides cognitive benefits attributed to group learning as established in literature [10, 15, 17, 21, 22]. However, all *learning groups* do not automatically collaborate well [19], thus the rationale to support groups for optimal collaboration.

In this context, groups interact either through verbal or text-based discourse; both have been posited in existing work to be similar in collaborative effect during joint problem solving (JPS) and that they can be juxtaposed in context [4, 6–8, 14, 16, 18, 20]. This paper improves upon the work by Schwarz & Asternhan [18] to provide a simpler computational mechanism to visualize (1) group collaboration compared to their social network based evaluation of group collaboration, and (2) individual participation compared to their many bars representing each individual's *variables of participation*, and individual participation measures, which is cumbersome and hard to base a real-time intervention on.

## 2 Study Design and Procedure

**Demographics of Participants:** A convenience sample of twenty students participated in this study, randomly grouped into teams of 4 members (Group G1: 3 male, 1 female, all aged 18–25; G2: 3 male, 1 non-disclosed; all 18–25; G3:



**Fig. 1.** Chat-room for groups' JPS discussion

2 male, 2 female; all 18–25; G4: 4 male, all 26–35; G5: 4 male, 3 aged 26–35, 1 aged 36–45).

The learning task and context provides each group with a joint task to solve; we adopted the “NASA man on the moon task” [1] for this study; a scenario of a space crew on the moon that needs to vacate a faulty spaceship to another one 200 miles away, with the group needing to rate 15 items in order of priority to take along [1]. The task meets Cohen’s recommendation [5] of a group task with respect to complexity and being open ended.

**System Design for Data Collection:** We designed a JPS-discourse (JPSD) chat-room shown in Fig. 1, an environment for online groups similar to *AcademicTalk* [14], *specialized work space* [20], *discussion tool for education* [16], *web interface* [9] and *e-argumentation* [18]. JPSD chat-room collects text-based interaction data as input to our model, to provide a simplified visualization of the individual participation and group collaboration level.

## 2.1 Data Model

**Gini-Coefficient Measure of Symmetry (GCMS)** used in Adeniran et al. [3, 13] is adapted to capture variables on interaction within online groups based on of their textual discourse content. The model adaptation is as follows:

A member  $i$ 's sequential text contribution at different time intervals is a collection of statements given by  $\vec{S}_1, \vec{S}_2, \dots, \vec{S}_m$ , which we call  $\vec{k}_i$ . So, member  $i$  contributes  $|\vec{k}_i|$ , to the group's discussion. GCMS of  $|\vec{k}_i|$  within groups represents a measure of group interaction quality [2, 11]; this is computed as follows: the mean of  $|\vec{k}_i|$  for a group is calculated as shown in Eq. 1:

$$k_{mean} = \frac{1}{n} \sum_{i=1}^n |\vec{k}_i| \quad (1)$$

The GCMS of contributions within a group is as shown in Eq. 2:

$$G_c = \frac{\sum_{i=1}^n \sum_{j=1}^n |k_i - k_j|}{2n^2 k_{mean}} \quad (2)$$

$G_c$  ranges from 0–1: 0 for perfect symmetry and 1 for perfect asymmetry. We assume that an indication of good collaboration is proportional to  $\frac{1}{G_c}$ .

**Word-count of contribution within a group** is considered for a more robust metric of collaboration. We found in literature that, “*the more collaborative groups had higher levels of verbal activity*” [12] and that elaborated discussion through explanation is an indicator of group collaboration and this results in the generation of volume of text in a textual discussion [6]. Evidence of collaborative skills [19] and its indicators during JPS [2, 3], all involves generating a volume of text when JPS discourse is text-based. Hence, we use the volume of text contributions to measure collaboration when a group discussion is text-based. The overall word-count of contributions by member  $i$  is derived from their text contributions,  $\vec{k}_i$ . Each statement  $\vec{S}_j \in \vec{k}_i$  is a sequence of words. The total word-count of all contributions by member  $i$  is:

$$w_{ct}^i = \sum_{j=1}^m |\vec{S}_j|, \text{ where } m = |\vec{k}_i| \quad (3)$$

However, a group may contain a highly extrovert member who contributes unnecessarily long texts or an extremely introvert member who contributes short texts. Therefore, we compute the median:

$$G(w_{ct}) = \text{median}(w_{ct}^1, w_{ct}^2, \dots, w_{ct}^n), \text{ where } n \text{ is the group size} \quad (4)$$

We combine  $G(w_{ct})$  and  $G_c$  to obtain a more sensitive measure of collaboration based on discourse called **WC/GCMS metric of collaboration within a group** as shown in Figure

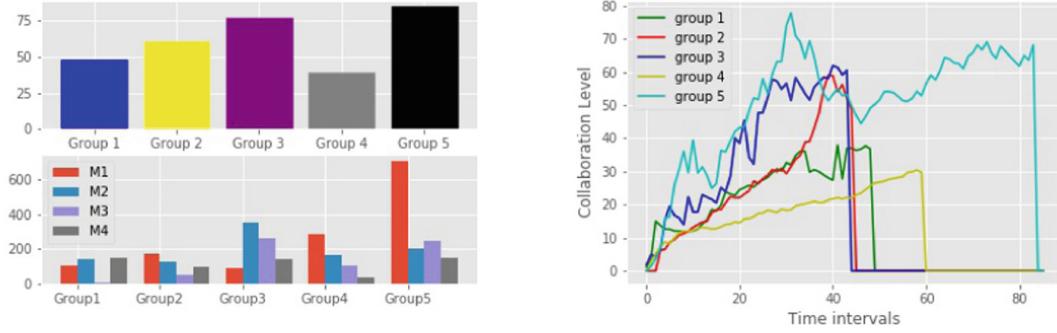
$$G_{cl} = \frac{G(w_{ct})}{G_c} \quad (5)$$

Figure 2a (*top*) shows the relative value of  $G_{cl}$  between the study groups.

## 2.2 Validating WC/GCMS Model and Visualization Output

**Real-Time Visualization of Group Collaboration Level:** Figure 2 shows the output of our collaboration metric model based on each group’s total discourse. Groups G3 and G5 collaborated more; this is corroborated by the real-time simulation (Fig. 2b) as G3 and G5 collaborated better, throughout JPS.

**Evaluation of Models’ Output and the Visualization:** We triangulate the output measures (as shown in Fig. 2), with qualitative data of the group discourse, considering the collaboration indicators/inhibitors identified in [2, 19]:



(a) Collaboration measure (top), and individual participation within groups (bottom).

(b) Simulated real-time collaboration level between groups with sequence of contributions at discrete time intervals

**Fig. 2.** Collaboration measure

**The discourse** in G3 & G5 shows evidence of collaborative skills [3, 19] with cognitive elaboration during JPS [22], whilst the other groups' discourse contains mainly *suggested solutions* which are mostly erroneous and blind agreements<sup>1</sup>. The latter groups' discourse is similar to what Webb [22] refers to as “giving and receiving non-elaborated help”, i.e. unexplained solutions to the JPS task. Such statements provide no cognitive benefit to the giver of the information nor to other members. In G1, 2, 4 many of these *unexplained solutions* are wrong.

**Individuals' participation level** influences the measure of group collaboration and there is evidence of non participating members in G1 and G4, members m3 in G1 and m4 in G4 respectively as shown in Fig. 2 (bottom) with “bar3” of G1 and “bar4” of G4, thus justifying the low collaboration measures for G1 and G4 shown in Fig. 2 (top).

**Quality of contribution and knowledge level** of context (in this case the environment of the moon) is evident in the discourse of groups G3 & G5 contrary to what we have in G1, 2 & 4. This justifies higher measures of collaboration in the former inline with the effect of knowledge level during JPS as presented in [2] and *Vygotskian perspective* mentioned in [22], which states that collaboration provides cognitive benefits when “a more expert member helps less-expert ones”.

### 3 Conclusions

Studies exist that have explored similar ideas as presented in this study; ours however adds to the existing knowledge to provide an easily interpretable visualization, based on a *scalable* and *generic* WC/GCMS model to evaluate the participation and collaboration level within online groups. Whilst the indicators of JPS collaboration exceed the characteristics of the text discourse content used in this paper, the WC/GCMS model is sensitive enough to serve as a

<sup>1</sup> For complete group discourse see [colab-learn.herokuapp.com/modelVS/groupX.php](http://colab-learn.herokuapp.com/modelVS/groupX.php) replacing X with the group number.

proxy-effective metric of collaboration and participation within online groups. However, whilst we gained valuable insights from our study, we would like to run a larger scale study to further investigate the indicators, factors and models presented. We will also investigate the use of our metrics and visualizations to provide real-time feedback to learners to scaffold collaboration, and measure both quantitatively and qualitatively the effect of such feedback on JPS. We further aim to develop algorithms for a computer agent (taking our models as input) to stimulate participation and consequently scaffold collaboration.

## References

1. Moon exercise. [humber.ca/centreforteachingandlearning/assets/files/pdfs/MoonExercise.pdf](http://humber.ca/centreforteachingandlearning/assets/files/pdfs/MoonExercise.pdf)
2. Adeniran, A.: Investigating feedback support to enhance collaboration within groups in computer supported collaborative learning. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 487–492. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_91](https://doi.org/10.1007/978-3-319-93846-2_91)
3. Adetunji, A., Masthoff, J., Beacham, N.: Analyzing groups problem-solving process to characterize collaboration within groups (2018)
4. Bromme, R., Hesse, F.W., Spada, H.: Barriers and Biases in Computer-Mediated Knowledge Communication: and How They May Be Overcome, vol. 5. Springer, Boston (2006). <https://doi.org/10.1007/b105100>
5. Cohen, E.G., Brody, C.M., Sapon-Shevin, M.: Teaching Cooperative Learning: The Challenge for Teacher Education. Suny Press, Albany (2004)
6. Curtis, D.D., Lawson, M.J.: Exploring collaborative online learning. *J. Asynchronous Learn. Netw.* **5**(1), 21–34 (2001)
7. Garcia, A., Jacobs, J.B.: The interactional organization of computer mediated communication in the college classroom. *Qual. Sociol.* **21**(3), 299–317 (1998)
8. Hancock, J.T., Landrigan, C., Silver, C.: Expressing emotion in text-based communication. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 929–932. ACM (2007)
9. Israel, J., Aiken, R.: Supporting collaborative learning with an intelligent web-based system. *Int. J. Artif. Intell. Educ.* **17**(1), 3–40 (2007)
10. Liu, S., Joy, M., Griffiths, N.: Incorporating learning styles in a computer-supported collaborative learning model (2008)
11. Martinez Maldonado, R.: Analysing, visualising and supporting collaborative learning using interactive tabletops (2013)
12. Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., Yacef, K.: Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *I. J. CSCL* **8**(4), 455–485 (2013)
13. Martinez-Maldonado, R., Kay, J., Yacef, K.: An automatic approach for mining patterns of collaboration around an interactive tabletop. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 101–110. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_11](https://doi.org/10.1007/978-3-642-39112-5_11)
14. Mcalister, S., Ravenscroft, A., Scanlon, E.: Combining interaction and context design to support collaborative argumentation using a tool for synchronous cmc. *J. Comput. Assist. Learn.* **20**(3), 194–204 (2004)

15. Newman, D.R., Webb, B., Cochrane, C.: A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpers. Comput. Technol.* **3**(2), 56–77 (1995)
16. Robertson, J., Good, J., Pain, H., et al.: Betterblether: the design and evaluation of a discussion tool for education. *Int. J. Artif. Intell. Educ.* **9**(3–4), 219–236 (1998)
17. Savery, J.R., Duffy, T.M.: Problem based learning: an instructional model and its constructivist framework. *Educ. Technol.* **35**(5), 31–38 (1995)
18. Schwarz, B.B., Asterhan, C.S.: E-moderation of synchronous discussions in educational settings: a nascent practice. *J. Learn. Sc.* **20**(3), 395–442 (2011)
19. Soller, A.: Supporting social interaction in an intelligent collaborative learning system. *Int. J. Artif. Intell. Educ.* **12**, 40–62 (2001)
20. Soller, A., Wiebe, J., Lesgold, A.: A machine learning approach to assessing knowledge sharing during collaborative learning activities. In: Computer Support for Collaborative Learning, pp. 128–137 (2002)
21. Suh, H., Lee, S.: Collaborative learning agent for promoting group interaction. *ETRI J.* **28**(4), 461–474 (2006)
22. Webb, N.M.: The teacher's role in promoting collaborative dialogue in the classroom. *Br. J. Educ. Psychol.* **79**(1), 1–28 (2009)



# Identifying Editor Roles in Argumentative Writing from Student Revision Histories

Tazin Afrin<sup>(✉)</sup> and Diane Litman

University of Pittsburgh, Pittsburgh, PA 15260, USA  
[{tazinafrin,litman}@cs.pitt.edu](mailto:{tazinafrin,litman}@cs.pitt.edu)

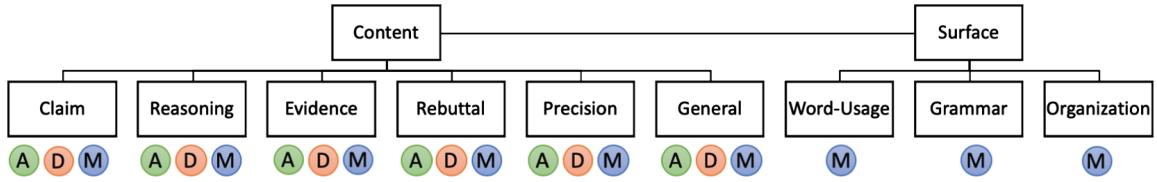
**Abstract.** We present a method for identifying editor roles from students' revision behaviors during argumentative writing. We first develop a method for applying a topic modeling algorithm to identify a set of editor roles from a vocabulary capturing three aspects of student revision behaviors: operation, purpose, and position. We validate the identified roles by showing that modeling the editor roles that students take when revising a paper not only accounts for the variance in revision purposes in our data, but also relates to writing improvement.

**Keywords:** Editor role · Argumentative writing · Revision

## 1 Introduction

Knowing that experienced and successful writers revise differently than inexperienced writers [4], various intelligent writing tools have been developed that provide localized feedback on text characteristics [3, 5, 6, 9]. These tools typically suggest edits to guide revision, rather than model the editing process after observing revisions. With the long term goal of developing an intelligent revision assistant, this paper presents an approach to modeling student editor roles.

Prior natural language processing (NLP) approaches to student revision analysis have focused on identifying revisions during argumentative writing and classifying their purposes and other properties [1, 7, 11, 12]. In contrast, editor roles have generally been studied in NLP using online collaborative writing applications such as Wikipedia [10]. Inspired by the use of Wikipedia revision histories [10], in this paper we similarly use topic modeling applied to revision histories to identify editor roles in the domain of student argumentative writing. To model student revision histories, between-draft essay revisions are extracted at a sentence-level and represented in terms of the following three aspects: operation (add, delete, or modify a sentence), purpose (e.g., correct grammar versus improve fluency), and position (revise at the beginning, middle or the end of an essay). To identify editor roles, a Latent Dirichlet Allocation (LDA) [2] graphical model is then applied to these revision histories. Finally, we show that the identified roles capture the variability in our data as well as correlate with writing improvement.



**Fig. 1.** The taxonomy of revision purposes [12] (A: Add, D: Delete, M: Modify).

**Table 1.** Example revision from aligned drafts of an essay from the Modeling Corpus.

Original draft	Revised draft	Operation	Purpose	Position
Self-driving vehicles pose many advantages and disadvantages	While self-driving vehicles pose many advantages and disadvantages, I am not on the bandwagon for them at this time	Modify	Claim	Beg.

## 2 Corpora

Our work takes advantage of several corpora of multiple drafts of argumentative essays written by both high-school and college students [11, 12], where all data has been annotated for revision using the framework of [12]. We divide our data into a Modeling Corpus (185 paired drafts, 3245 revisions) and an Evaluation Corpus (107 paired drafts, 2045 revisions), based on whether expert grades are available before (Score1) and after (Score2) essay revision. Although the grading rubrics for the college and high-school essays in the Evaluation Corpus are different, both are based upon common criteria of argumentative writing, e.g., clear thesis, convincing evidence, clear wording without grammatical errors, etc. We apply linear scaling<sup>1</sup> to bring the scores within the same range of [0,100]. After scaling, the average Score1 and Score2 are 64.41 and 73.59, respectively.

For all essays and prior to this study, subsequent drafts were manually aligned at the sentence-level based on semantic similarity. Nonidentical aligned sentences were extracted as the **revisions**, resulting in three types of revision **operations** - *Add*, *Delete*, *Modify*. Each extracted revision was manually annotated with a **purpose** following the revision schema shown in Fig. 1 (modified compared to [12] by adding the *Precision* category). For this study, each revision's **position** was in addition automatically tagged using its paragraph position in the revised essay. To maintain consistency across essays, instead of using paragraph number, we identify whether a revision is in the first (*beg*), last (*end*), or a middle (*mid*) paragraph. Table 1 shows a modified claim at the beginning of an essay from the Modeling Corpus.

## 3 Identifying Editor Roles

To create a vocabulary for topic modeling and to understand the repeating patterns of student editors, we represent each revision utilizing the three aspects

<sup>1</sup> Formula used to scale the scores =  $100 * (x-\min)/(max-\min)$ .

**Table 2.** Derived editor roles with top 10 revisions. (Blue: Surface, Orange: Content)

Proofreader	Copy editor	Descriptive editor	Analytical editor	Persuasive editor
Grammar_mid	Word-Usage_mid	+General_mid	Word-Usage_beg	+Reasoning_mid
Grammar_beg	Word-Usage_beg	Word-Usage_mid	+General_end	-Reasoning_mid
Word-Usage_mid	+Reasoning_mid	-General_mid	+Reasoning_end	+Claims_mid
Grammar_end	Word-Usage_end	General_mid	Word-Usage_end	+Evidence_mid
Word-Usage_end	Organization_mid	Evidence_mid	Organization_beg	+General_mid
Word-Usage_beg	-General_end	Precision_mid	-Reasoning_end	-General_mid
Precision_beg	General_end	-General_beg	+Claims_end	Reasoning_mid
General_mid	-Reasoning_mid	+General_beg	+Evidence_mid	-General_beg
General_end	Claims_mid	Reasoning_mid	+Rebuttal_end	-Claims_mid
Reasoning_beg	-General_mid	+Claims_beg	Organization_mid	+General_beg

described earlier: operation, purpose, and position. This yields a rich and informative vocabulary for modeling our data, consisting of 63 revision “words” (54 content, 9 surface). This is in contrast to the 24 word revision vocabulary used in the prior Wikipedia editor role extraction method [10], formed using a Wiki-specific revision taxonomy of operation and purpose. When describing our revision “words”, add and delete revisions are represented with ‘+’ and ‘-’ sign, and no sign for modification, e.g., *Claim\_beg* in Table 1. Editors are then represented by their history of revisions in terms of this revision vocabulary.

We trained the LDA model on the Modeling Corpus and experimented with 2 to 10 topics. After an extensive evaluation for topic interpretation based on top 10 revisions under each topic, we ended up with 5 topics where the revisions under each topic intuitively correspond to one of a set of potentially relevant editor roles for academic writing. We drew upon roles previously identified for writing domains such as newspaper editing (e.g., proofreader, copy editor), Wikipedia (e.g., technical editor, substantive expert), and academic writing<sup>2</sup> (i.e., descriptive, analytical, persuasive, and critical).

The final topics are shown in Table 2, labeled by us with the best-matching editor role from the anticipated set of potential roles, based on the vocabulary items in each topic. The defining characteristic of a **Proofreader** are surface-level error corrections. **Copy** editors ensure that the article is clear and concise as they revise for word-usage, clarity, and organization. **Descriptive** editors provide details and enhance clarity, with widespread development of general content. **Analytical** editors revise by adding information and better organizing thoughts, with top revision purposes being word-usage, content, reasoning, and rebuttal. **Persuasive** editors discuss ideas and facts with relevant examples and develop arguments with added information.

---

<sup>2</sup> <https://sydney.edu.au/students/writing/types-of-academic-writing.html>.

**Table 3.** Variance across editors for each revision purpose ( $p < .001$  :\*\*\*,  $N = 107$ ).

Purpose	Grammar	Word-usage	Organization	Claims	Reasoning	General	Evidence	Rebuttal
<b>R<sup>2</sup>-value</b>	0.573***	0.537***	0.043	0.240***	0.397***	0.459***	0.223***	0.025

**Table 4.** Partial correlations between role probabilities and Score2 controlling Score1.

Editor roles	Proofreader	Copy	Descriptive	Analytical	Persuasive
<b>Corr(p-value)</b>	-0.175(0.073)	-0.049(0.621)	-0.180(0.064)	-0.013(0.891)	0.205( <b>0.035</b> )

## 4 Validating Editor Roles

Using the trained topic model, we first calculate the probability of an editor belonging to each of the 5 roles, for each editor in the Evaluation Corpus. These probabilities represent each role’s contribution to the essay revision. Motivated by Wikipedia role validation [10], we first validate our editor roles by similarly using editor roles to explain the variance in revision purposes. We create 8 linear regression models, one for each revision purpose<sup>3</sup>. The models take as input a five dimensional vector indicating an editor’s contribution to each role and the output is the editor’s edit frequency for each revision purpose. The R-squared values in Table 3 show that our topic model can best explain the variance of Grammar, Word-Usage, General content, Claim, Reasoning, and Evidence edits.

A corpus study in [12] showed that content changes are correlated with argumentative writing improvement, reaffirming the statement of [4]. Using a similar method, we investigate if our editor roles are related to writing improvement. We calculate partial Pearson correlations between editor roles and Score2 while controlling for Score1 to regress out the effect of the correlation between Score1 and Score2 ( $\text{Corr.} = 0.692, p < 0.001$ ). Table 4 shows that the roles consisting of only surface edits or a mixture of edits are not correlated to writing improvement. However, Persuasive editor, which consists of content revisions, shows a positive significant correlation to writing improvement. Our results suggest that the Persuasive editor is the role of an experienced writer.

## 5 Conclusion and Future Work

Although editor roles have been studied for online collaborative writing [8, 10], our research investigates student revisions of argumentative essays. While our model follows previous methods [10], we introduce a unique vocabulary to model each editor’s revision history, with evaluation results suggesting that our identified roles capture salient features of writing. Future plans include using a Markov model to consider revision order, expanding the revision vocabulary, and using the predictions to provide feedback in an intelligent revision assistant.

**Acknowledgements.** This work is funded by NSF Award 1735752.

---

<sup>3</sup> The Evaluation Corpus does not have precision revisions.

## References

1. Afrin, T., Litman, D.: Annotation and classification of sentence-level revision improvement. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana, pp. 240–246, June 2018
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
3. Eli Review, T.: (2014). <https://elireview.com>. Accessed 02 June 2019
4. Faigley, L., Witte, S.: Analyzing revision. *Coll. Compos. Commun.* **32**(4), 400–414 (1981)
5. Grammarly (2016). <http://www.grammarly.com>. Accessed 02 June 2019
6. Roscoe, R.D., McNamara, D.S.: Writing pal: feasibility of an intelligent writing strategy tutor in the high school classroom. *J. Educ. Psychol.* **105**(4), 1010–1025 (2013)
7. Tan, C., Lee, L.: A corpus of sentence-level revisions in academic writing: a step towards understanding statement strength in communication. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2: Short Papers, Baltimore, MD, USA, pp. 403–408, June 2014
8. Welser, H.T., et al.: Finding social roles in Wikipedia. In: Proceedings of the 2011 iConference, iConference 2011, pp. 122–129. ACM, New York (2011)
9. The Writing Mentor: ETS writing mentor (2016). <https://mentormywriting.org/>. Accessed 02 June 2019
10. Yang, D., Halfaker, A., Kraut, R.E., Hovy, E.H.: Who did what: editor role identification in Wikipedia. In: Tenth International AAAI Conference on Web and Social Media, ICWSM, pp. 446–455. AAAI Press (2016)
11. Zhang, F., Hashemi, H., Hwa, R., Litman, D.: A corpus of annotated revisions for studying argumentative writing. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 1568–1578. Association for Computational Linguistics (2017)
12. Zhang, F., Litman, D.: Annotation and classification of argumentative writing revisions. In: Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications, Denver, Colorado, pp. 133–143. Association for Computational Linguistics, June 2015



# Degree Curriculum Contraction: A Vector Space Approach

Mohamed Alkaoud<sup>1</sup>(✉) and Zachary A. Pardos<sup>2</sup>

<sup>1</sup> University of California, Davis, Davis, CA, USA  
maalkaoud@ucdavis.edu

<sup>2</sup> University of California, Berkeley, Berkeley, CA, USA  
zp@berkeley.edu

**Abstract.** This paper introduces a curriculum contraction technique in the context of university degree programs using a vector space embedding approach. We propose a way to model degrees and majors and define a contraction that takes the curriculum of a degree program and defines a smaller set of courses to approximate it. For example, a computer science degree curriculum could be generated that takes three years to complete instead of four (a 75% contraction). We use seven years of student enrollment data from a public university to train our embedding model. The most popular majors at the university, and their corresponding minors, are used to evaluate the validity of this contraction approach where minors are treated as major contractions.

**Keywords:** Higher education · Vector space embedding · Curriculum

## 1 Introduction

Online and on-campus, educational platforms and institutions are looking to “right-size” the curricular experience of learners. This ranges from offering traditional four-year Bachelor’s degrees to six course “micro degree” credentials. In this work, we attempt to automate the process of contracting the length of a university degree program while retaining as much of the core value as possible (i.e., which courses should be chosen in a 1-year version of a 4-year program?). This problem does not arise only in traditional academic settings; MOOC providers can benefit from having a way to automatically group their courses to offer course sequence credentials<sup>1</sup>.

Related work [3] looked at algorithmic approaches to sequencing arbitrary curricular components for theoretical optimal retention. Also related are course recommendation systems [5] which help students navigate their chosen degree program. We do not know of any previous work that has used algorithmic approaches for curriculum contraction.

---

<sup>1</sup> <https://www.edx.org/micromasters>.

In this paper, we introduce an automated way to contract degrees. In Sect. 2, we talk about the enrollment data that we use in our approach. Section 3 highlights our degree modeling and contraction techniques. In Sect. 4, we show the results of our technique and measure the quality of its performance.

## 2 Data

We use anonymized student enrollment data to train our embeddings. The dataset, provided by UC Berkeley, contains all student enrollments (over 140,000 student) from 2008–2015 across all departments and divisions. Table 1 shows the structure of the dataset. We preprocess the data by removing graduate students and filtering out graduate courses from undergraduates who have taken them, in order to focus the models on only the undergraduate curriculum. Students who have been enrolled for less than eight semesters or more than twelve are also removed.

**Table 1.** Berkeley’s enrollment dataset

Masked ID	Year	Semester	Course ID	Masked ID	Major
111	2010	Fall	Integrative biology 127	111	Bioengineering
222	2012	Spring	Mathematics 55	222	Computer science

## 3 Approach

### 3.1 Vector Space Embedding

Word embedding algorithms, such as word2vec [2], GloVe [7], and FastText [1], are powerful tools that allow us to represent a word by a high-dimensional vector while still capturing semantic information. Pardos et al. [5, 6] proposed a new way to represent courses by using word embedding techniques to learn a vector representation from students’ enrollments. They represent each student as a sequence of the courses they’ve taken and train a word embedding technique on the sequences generating a vector of real values for each course. We followed a similar approach to the one performed in [5, 6] in generating course embeddings. We start by representing each student as a sequence of the courses they’ve taken:

$$s_i = [c_{i_1}, c_{i_2}, c_{i_3}, \dots, c_{i_n}] \quad (1)$$

where  $s_i$  is student  $i$  and  $[c_{i_1}, \dots, c_{i_n}]$  is the sequence of all the courses student  $i$  has taken. Notice that  $c_i$  only refers to the symbol of the class (e.g. Physics 7A) without any auxiliary information such as the course description and syllabus. Now we run FastText [1] to get a vector representation for each course. We can now vectorize each course by its learned embedding:

$$\text{course2vec}(c_i) = [z_1, z_2, z_3, \dots, z_n] \quad (2)$$

where  $c_i$  is course  $i$  and  $z_i$ 's are real numbers that are computed using FastText. Our contraction technique requires a vector representation of academic degrees. We present a dynamic way to represent degrees as vectors learnt from students enrollments as follow:

$$\text{degree2vec}(d) = \frac{1}{|S_d||C_s|} \sum_{s \in S_d} \sum_{c \in C_s} \text{course2vec}(c) \quad (3)$$

where  $d$  is a degree,  $S_d$  is the set of all students majoring in  $d$ , and  $C_s$  is the set of all courses that student  $s$  has taken.

### 3.2 Contraction

Let  $D$  be a degree plan consisting of courses. The high-level description of our contraction technique is as follow: (1) embed the courses in a vector space, (2) calculate a degree representation vector, and (3) find the best set of classes of size  $k$  that approximates the degree representation. For steps one and two, we use  $\text{course2vec}$  and  $\text{degree2vec}$  respectively. For the third step, we want the best subset of courses that are closest to the degree  $D$ :

$$\text{contract}(D, k) = \arg \min_{d \in \mathcal{P}_k(D)} \sqrt{[\text{degree2vec}(D) - \frac{1}{|d|} \sum_{c \in d} \text{course2vec}(c)]^2} \quad (4)$$

where  $\mathcal{P}_k(D)$  is all subsets of  $D$  that are of size  $k$ . Finding all subsets is computationally expensive and is not feasible for typical classes sizes; picking 10 classes from a 100 will yield more than 10 trillion sets! One way to make it faster is to use a *greedy* approach to find the closest  $k$  courses instead of finding the closest set of size  $k$ . We want to avoid using a greedy solution since we do not want to lose semantic relationships; two courses may be far from the degree vector but their average may be closer than any other course vector. We instead use a *hybrid* approach where we find the closest set of size four instead of  $k$ . After that, we remove the best four courses from  $D$ , decrement  $k$  by four, and repeat the process. Eventually, when  $k$  is small, we take combinations of sets of sizes: three, two, and/or one. We choose four to be our subset size since it scales well and still captures interesting compositions.

## 4 Evaluation and Discussion

To evaluate our approach, we take a major, apply our contraction technique on it, and then compare it with its corresponding minor. We picked the ten most popular majors in Berkeley from 2010–11 to 2014–15 [4] that have corresponding minors to evaluate our approach. The problem with minors is that some departments do not design them to reflect the original full-length major. This is often reflected in the thin structure of the minor. To avoid this problem, we went through each minor individually and ensured that they have a well-defined structure, as expressed by their requirements.

**Table 2.** Performance of our approach for each major/minor pair

Recall@ $k$	Greedy	Hybrid	Recall@ $k$	Greedy	Hybrid
EECS	0%	57.14%	History	83.33%	83.33%
Mechanical engineering	71.43%	71.43%	Anthropology	62.5%	75%
Architecture	25%	37.5%	Chemical engineering	40%	40%
Statistics	44.44%	44.44%	Rhetoric	42.86%	57.14%
Environ econ & policy	37.5%	50%	Philosophy	66.67%	66.67%
Average Recall@ $k$				47.37%	58.27%

We apply *degree2vec* on each major  $m$  to get its embedding. After that, we create a set containing all department courses that students majoring in  $m$  have taken and remove courses from other departments. While some minors contain courses from other departments, we do not want to handpick these departments as it may add bias to the evaluation. We then run our contraction to get  $k$  courses. We pick  $k$  to be equal to the number of courses in the corresponding minor. The recall@ $k$  is then measured for each major. Recall@ $k$  gives us the proportion of the minor classes we found in the contraction, i.e the percentage of the minor requirements satisfied if one takes the classes proposed by the contraction. In the case of an elective course requirement, we say that the contraction satisfied the requirement if it contains one of the elective courses. Table 2 shows the performance of the two contraction approaches discussed in Subsect. 3.2. It is important to note that it is impossible for us to achieve perfect recall@ $k$  since we only take only department courses; the maximum recall@ $k$  we can achieve is 87.88%. Table 3 shows an example of applying our contraction method to the Mechanical Eng. major and its minor. One thing we notice in Table 2 is that the hybrid approach is always better or equal to the greedy approach which reinforces our assumption that the greedy approach will miss important relationships.

We showed in this paper that we can achieve good performance in curriculum contraction by using a simple vector space model that does not incorporate any textual information about the courses. There are many ways in which this work can be extended including a more sophisticated way of combining courses instead of averaging, and a more comprehensive technique to include possibly relevant courses from departments outside of the major by finding department vectors that are close to the main department vector.

**Table 3.** An example of contracting the mechanical engineering major and comparing it with its minor. Courses that map to a minor requirement are highlighted in bold.

Minor requirements		Hybrid contraction	
Physics 7A	Mec Eng upper division class 1	Mec Eng 98	<b>Mec Eng 106</b>
Mec Eng 40	Mec Eng upper division class 2	Mec Eng W85	<b>Mec Eng 120</b>
Mec Eng 104	Mec Eng upper division class 3	<b>Mec Eng 104</b>	<b>Mec Eng 190AC</b>
Mec Eng C85	-	<b>Mec Eng C85</b>	-

## References

1. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint [arXiv:1607.01759](https://arxiv.org/abs/1607.01759) (2016)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
3. Novikoff, T.P., Kleinberg, J.M., Strogatz, S.H.: Education of a model student. Proc. Natl. Acad. Sci. **109**(6), 1868–1873 (2012)
4. Office of Planning & Analysis, University of California, Berkeley: Majors and Minors of Degree Recipients, 2010–11 to 2014–15 (2016)
5. Pardos, Z.A., Fan, Z., Jiang, W.: Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. User Modeling and User-Adapted Interaction (in press). <https://arxiv.org/abs/1803.09535>
6. Pardos, Z.A., Nam, A.J.H.: A Map of Knowledge (2018)
7. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)



# L2 Learners’ Preferences of Dialogue Agents: A Key to Achieve Adaptive Motivational Support?

Emmanuel Ayedoun<sup>(✉)</sup> , Yuki Hayashi, and Kazuhisa Seta

Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai,

Osaka 599-8531, Japan

[eyedoun@ksm.kis.osakafu-u.ac.jp](mailto:eyedoun@ksm.kis.osakafu-u.ac.jp)

**Abstract.** This study reports on differences observed among learners’ preferences of two conversational strategies embedded in a dialogue agent dedicated to enhancing their willingness to communicate (WTC) in a second language. We found that the combination of both strategies is, in general, the most preferred by learners. However, perception, as well as effects of the support provided by these strategies seem to vary according to learners’ level of willingness to communicate. Lower WTC learners tended to prefer affective backchannels while their counterparts seem rather favor communication strategies. These results were also in line with posttest results which revealed that learners’ expected WTC tended to be higher after interacting with dialogue agents embedding their preferred strategies. In sum, these results can be viewed as preliminary evidence of the meaningfulness to account for second language learners’ preferences towards balancing adaptively the type of strategies employed by dialogue agents to motivate learners towards communication in the target second language.

**Keywords:** Adaptive · Language learning · Conversational agents · Willingness to communicate in L2 · Communication strategies · Affective backchannels

## 1 Introduction

The primary purpose of second language (L2) learning is to provide learners the ability to convey their intended meaning effectively in the target language and, by extension, facilitate exchanges between people coming from different countries. MacIntyre et al. [1] suggested that the key factor ensuring a sustained L2 use is the willingness to communicate (WTC) in L2, defined as a “readiness to enter into discourse at a particular time with a specific person or persons, using an L2”. WTC studies have shown that learners displaying high WTC are more likely to show more improvement in their communication skills [2] and to attain higher levels of language fluency [3], supporting the idea that increasing L2 learners’ WTC should be the ultimate goal of L2 learning [1].

In our previous works [4, 5], we implemented a conversational agent enhanced with two types of conversational strategies (i.e., communication strategies (CS) and affective backchannels (AB), see Table 1 for some examples) dedicated to carrying on WTC-

friendly conversations with learners in an English-as-a-foreign-language context. An experimental evaluation of the system hinted on the practical significance of using such conversational strategies enhanced dialogue agent to foster L2 learners' WTC [5].

**Table 1.** Examples of scaffolds (CS and AB) implemented in the conversational agent [5].

	Strategy	Description	Example
CS	Simplification	Use an alternative or a shorter term, which express the meaning of the target lexical term	Agent: May I have your order? Learner: ... (silent) Agent: Order please
AB	Encouraging AB	Used when the learner seems to hesitate to the extent that he/she remains silent	• Come on! Don't be shy... • You can do it...

**Table 2.** Overview of the experiment flow.

Steps	Group 1 (n = 10)	Group 2 (n = 10)	Group 3 (n = 10)	Group 4 (n = 10)	Group 5 (n = 10)	Group 6 (n = 10)
Step 0	First WTC questionnaire (Pretest)					
Step 1	Warm-up interaction with the system					
Step 2	CS + AB	CS + AB	CS	AB	CS	AB
Step 3	Second WTC questionnaire (Posttest)					
Step 4	CS	AB	CS + AB	CS + AB	AB	CS
Step 5	System preference survey					

In the present study, we take a closer look at learners' perceptions of the support provided by each of the strategies mentioned above or their combination and investigate differences in preferred strategies according to learners' WTC levels. We finally discuss the feasibility to achieve a tailored deployment of these conversational strategies accordingly with learners' preferences and actual level of WTC.

## 2 Experimental Study

### 2.1 Research Questions and Study Design

Our work investigates the following research questions:

**RQ1.** *What are the differences in L2 learners' perceptions or preferences of the WTC support provided by the system?*

**RQ2.** *How do WTC outcomes vary according to such L2 learners' differences?*

**RQ3.** *How can we tailor the WTC support provided by the system towards accounting for such differences in L2 learners' preferences?*

The flow of the experiments was designed according to six steps (Step 0 to Step 5) to compare learners' WTC results across different versions of the system on the one

hand and examine their preference after interacting with different versions of the system, on the other as shown in Table 2. We gauged learners' WTC by using a self-report survey [6] before (Step 0) and after (Step 3) their first interaction with the system (Step 2). The WTC surveys targeted three variables: *confidence*, *anxiety*, and *desire to communicate*, which are the immediate precursors of WTC [7].

A total of 60 male and female university students' data were gathered and used in this study. Experiment data for participants of groups 1 to 4 were obtained from our previous work [5]. We then just ran additional experiments to collect data for those of groups 5 and 6. To preserve the uniformity of conditions across the two studies, we also rigorously made sure to replicate the experimental settings as in [5]. Moreover, a one-way ANOVA was conducted and confirmed the homogeneity of initial WTC conditions (Step 0) among the six groups. In detail, the tests revealed that there were no statistically significant differences among the six groups in terms of initial *confidence* [ $F(5, 54) = 1.85, p = .12$ ], *anxiety* [ $F(5, 54) = 0.44, p = .81$ ], and *desire* [ $F(5, 54) = 1.36, p = .25$ ].

## 2.2 Results

### Differences in Learners' Preferences of Conversational Strategies

In order to investigate differences in participants' preferences of CS and or AB versions of the system, we analyzed the results of the system preference survey (Step 5) with respect to learners' initial WTC level. To that extent, all the participants were labeled as lower or higher WTC according to the results of their First WTC questionnaire (Step 0). More concretely, participants who had all their initial WTC precursors (*confidence*, *anxiety*, and *desire*) better than average scores were labeled as higher WTC, while their counterparts were categorized as lower WTC. The resulting distribution of participants according to their WTC level was relatively uniform across the 6 groups. A Barnard's test for independence was conducted indicating a relationship between learners' WTC level and their preference for CS or CS + AB ( $p = .04$ ), with a medium (Cramer's  $V = .47$ ) effect size according to Cohen's conventions for Cramer's V [8]. Similarly, Barnard's test for independence indicated relationships between learners' WTC level and their preference for AB or CS ( $p = .01, V = .6$ ). Finally, we found a trend towards relationship between learners' WTC level and their preference for AB or CS + AB ( $p = .09, V = .42$ ).

In sum, these results indicate that participants' preference tendencies of the different system versions seem to be related to their WTC level to some extent.

### Relationships between Learners' WTC Level and WTC Outcomes

In order to investigate whether and how WTC outcomes would vary according to learners' WTC level, we analyzed differences among lower and higher learners' WTC results after interactions with the three different versions of the system. A one-way ANCOVA was conducted to determine whether there were statistically significant differences between WTC posttest results whilst controlling for pretest results. Post-hoc Tukey Kramer tests were additionally run to further investigate the differences. There was a significant difference in lower WTC participants' expected *confidence*

[ $F(2, 28) = 3.55, p < .05$ ], *anxiety* [ $F(2, 28) = 3.40, p < .05$ ] and *desire* [ $F(2, 28) = 3.39, p < .05$ ] among the three versions (i.e., CS, AB and CS + AB). The post-hoc Tukey Kramer tests showed that as far as lower WTC participants are concerned, the CS + AB and AB versions are more promising than the CS version in enhancing their WTC.

Regarding higher WTC participants, the one-way ANCOVA tests revealed that there was a significant difference in their expected *confidence* [ $F(2,24) = 3.48, p < .05$ ], *desire* [ $F(2,24) = 4.97, p < .05$ ], and a trend towards significant difference for *anxiety* [ $F(2,24) = 2.88, p < .1$ ]. The post-hoc Tukey Kramer tests showed that as far as higher WTC participants are concerned, the CS + AB and CS versions are in most cases more promising than the AB version in enhancing learners' WTC.

To sum up, the analysis of WTC outcomes with respect to participants' WTC level suggests that: CS + AB and AB versions seem to work better for lower WTC participants, while for higher WTC participants, the most effective system versions seem to be the CS + AB and CS versions.

### 2.3 Discussion and Limitations

The above-described results allow us to draw a number of preliminary conclusions.

Firstly, we found that learners' perception of the system support, through the use of CS and or AB, tended to vary according to the stage of development of their WTC. Although the combination of both strategies (i.e., CS + AB) was the most preferred by learners, we observed that lower WTC learners tended to prefer AB over CS, whereas higher WTC learners on the contrary, tended to favor CS over AB (*RQ1*).

Secondly, we found that the effectiveness of CS and or AB toward increasing WTC is related to learners' WTC level, and consequently to their preferences. The CS + AB and CS versions tended to work better for higher WTC learners, while lower WTC learners tended to benefit more from their interactions with the CS + AB and AB versions. In other words, learners' preferred versions and effective system versions towards enhancing their WTC tended to be coherent, both for lower and higher WTC learners (*RQ2*).

Altogether, these results indicate that it would be reasonable to achieve a more tailored WTC support to L2 learners by accounting for their preferences. To this extent, we assume that a carefully balanced use of the combination of CS and AB according to learners' WTC level and preferences may be more beneficial for L2 learners. For example, lower WTC learners (i.e., who tended to prefer AB over CS) could be presented with a CS + AB version where AB are more frequently triggered than CS, while for higher WTC learners (i.e., who tended to prefer CS over AB), a version of the system where CS are more frequently used than AB may be employed (*RQ3*).

## 3 Conclusion and Future Works

In this paper, we shed light on differences in learners' preferences of motivational scaffolds used by a dialogue agent dedicated to enhancing their motivation towards communication. We have found that not only learners' preferences tend to vary

according to their WTC level, but also, such preferred strategies tended to be promising towards increasing their WTC. Such findings suggest the key role that L2 learners' preferences could play in achieving a personalized computer-based L2 WTC support.

Directions for future works include redesigning the using, as well as the fading balance of motivational scaffolds by the conversational agent according to learners' preferences, and carrying out additional evaluations with a larger sample size to validate the premises of this paper. We will also explore the feasibility to achieve a higher degree of personalization in the WTC support provided by the dialogue agent.

## References

1. MacIntyre, P.D., Clément, R., Dörnyei, Z., Noels, K.A.: Conceptualizing willingness to communicate in a L2: a situational model of L2 confidence and affiliation. *Mod. Lang. J.* **82** (4), 545–562 (1998)
2. Yashima, T., Zenuk-Nishide, L., Shimizu, K.: The influence of attitudes and affect on willingness to communicate and second language communication. *Lang. Learn.* **54**(1), 119–152 (2004)
3. Derwing, T.M., Munro, M.J., Thomson, R.I.: A longitudinal study of ESL learners' fluency and comprehensibility development. *Appl. Linguist.* **29**(3), 359–380 (2008)
4. Ayedoun, E., Hayashi, Y., Seta, K.: Web-services based conversational agent to encourage willingness to communicate in the EFL context. *J. Inf. Syst. Educ.* **15**(1), 15–27 (2016)
5. Ayedoun, E., Hayashi, Y., Seta, K.: Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *Int. J. Artif. Intell. Educ.* **29**(1), 29–57 (2019)
6. Matsuoka, R.: Willingness to communicate in English among Japanese college students. In: Proceedings of the 9th Conference of Pan-Pacific Association of Applied Linguistics, pp. 165–176 (2006)
7. MacIntyre, P.D., Charos, C.: Personality, attitudes, and affect as predictors of second language communication. *J. Lang. Soc. Psychol.* **15**(1), 3–26 (1996)
8. Aron, A., Coups, E., Aron, E.: Statistics for Psychology, 6th edn. Pearson Education Limited (2013)



# Eye Gaze Sequence Analysis to Model Memory in E-education

Maël Beuget<sup>1</sup>, Sylvain Castagnos<sup>1()</sup>, Christophe Luxembourger<sup>2</sup>,  
and Anne Boyer<sup>1</sup>

<sup>1</sup> CNRS-LORIA-University of Lorraine, Vandœuvre-lès-Nancy, France  
[sylvain.castagnos@loria.fr](mailto:sylvain.castagnos@loria.fr)

<sup>2</sup> 2LPN-University of Lorraine, Nancy, France

**Abstract.** Intelligent Tutoring Systems are now mature technologies that successfully help students to acquire new knowledge and competencies through various educational methods and in a personalized way. However, evaluating precisely what they recall at the end of the learning process remains a complex task. In this paper, we study if there are correlations between memory and gaze data in the context of e-education. Our long-term goal is to model the memory of students thank to an eye-tracker in a continuous and transparent way. These models could then be used to adapt recommendations of pedagogical resources to the students' learning rate. So as to address this research question, we designed an experiment where students were asked to learn a short lesson about Esperanto. Our results show that some gaze characteristics are correlated with recall in memory.

**Keywords:** Eye-tracking study · Memory · Gaze behavior · E-education

## 1 Introduction

The introduction of digital technologies in the society during the past decades impacts inevitably many aspects of our lives. Education is one of them, and the way numeric tools can improve the quality of learning is nowadays a full-fledged research field. Intelligent Tutoring Systems (ITS) are pieces of software that help students at mastering courses. They provide them with some lessons, incorporate knowledge assessment tools, allow instant and personalized feedback, and eventually propose students to train on their specific academic deficiencies. However, estimating accurately a user state of knowledge remains a difficult task. It is quite impossible to cover every aspect of the lesson through an exam, as the evaluation process should not be too much time consuming. In some cases, students may also provide randomly correct answers. An imprecise or incomplete user model may impact the feedback quality and the relevance of the recommendations provided by ITS. To overcome this problem, we are wondering if it is possible to infer more precisely what the user remembers by collecting

implicit traces of interactions. The latter are not consciously indicated by the user, and could be for example action logs, facial expressions or eye movements. Gaze data interests us more specifically as it has been reported that the gaze behavior could reflect some cognitive processes [3, 13]. Our goal is to analyze if it could exist correlations between gaze characteristics and the fact to remember some courses items. To do so, we designed a pilot study where 22 students had to learn a short lesson about Esperanto while wearing eye-tracking glasses. After the learning phase, they were asked to pass an exam. We then compared the gaze data with their grade on the exam. Results show that fixation durations, scanpath length and scanpath angles are good predictors of what have been recalled. As a perspective, eye-trackers are thus promising tools to model users' memory in real time when they are reading their lessons.

## 2 Related Work

### 2.1 A Brief History of the Memory Models

It exists not only one but several forms of memories, and the way we can model these forms of memories is not unanimous among psychologists. The most widely accepted models are those proposed by Atkinson-Shiffrin [12], Baddeley [2] and Miyake *et al.* [14]. In each of them, memory refers to the ability to encode, store and retrieve past experiences. It is composed of several memory modules. The sensory register allows the incoming information to be encoded for treatment. The short-term memory holds a small amount of information in mind in an active readily available state for a few seconds. The working memory manages and handles the information required to carry out complex tasks such as comprehension, reasoning, and learning. The differences between the models of Atkinson-Shiffrin, Baddeley and Miyake mainly focus on the distinction and overlap between short-term memory and working memory [1]. Finally, the long-term memory stores information for an extended period of time, consciously or not [17]. In our case, we take an interest in the working memory since it is the one involved in the learning process.

In parallel with this segmentation of the different forms of memory, researchers distinguish recognition and recall. The recognition process is the fact to remember something when the stimuli is present, whereas the recall involves to remember a stimulus which is not physically present [6]. In the context of e-education, we assume that learning a lesson correctly consists in storing the information in the recall memory.

### 2.2 Linking Cognitive Processes and Eye Movements

According to Just and Carpenter, what a person is looking at is assumed to indicate the “on top of the stack” thought of cognitive processes [9]. More recently, Steichen *et al.* aimed at identifying gaze patterns [18], while Bondareva *et al.* use the gaze data of users while they interact with an intelligent tutoring system to

predict in real time the efficiency of the users' learning process [3]. However, they did not investigate the possibility to infer memory. Let us note that memorized items are the results of the learning process. Predicting the quality of learning therefore amounts to predicting the amount of information stored, without however knowing how to distinguish the forgotten pieces of information from the remembered ones. Regarding the link between memory and gaze data, Hanula et al. have established the ability to predict recognition through a user study where the stimuli proposed to users were faces of people [7]. Several studies also tried to exploit gaze characteristics to predict memorability [4] and recall [13] of images. To our knowledge, no study has sought to establish this link between gaze and recall in a context closer to e-education, *i.e.* with multimedia content (text and images). This is what drives our research project. Our purpose is to verify if a link could exist between the recall process and some gaze metrics, to later potentially infer a user state of memories.

To analyze gaze data the first step is to transform the eye-tracker's sampled signal into a scanpath composed by fixations and saccades. Fixations are points of gaze where the fovea is concentrated for a short period of about 200ms, whereas saccades are the eye movement that link two consecutive fixations. Many algorithms are described in the literature to transform the sampled signal into a scanpath [10, 11, 15]. In our study we used the IV-T algorithm provided by the Tobii Pro lab software. The way we can extract metrics and information from scanpath has also been widely documented [5, 8, 13, 16].

### 3 User Study

We designed a pilot study where subjects were asked to learn a short lesson (5 printed static pages from an online course) and had to report their knowledge during an exam. Documents were not allowed during this exam. We have chosen to work in the context of language learning. So as to avoid the bias of languages already learned, we based our lesson on the Esperanto language. The latter has been created in 1887 with the ambition to become a vehicular language, and has been built by combining several existing languages. It is one of the less taught languages, so few people studied Esperanto. We had 22 participants (10 females/12 males) aged from 11 to 16 years old. Our eye-tracker was a Tobii Pro Glasses 2 (100 Hz frequency rate). The experiment is divided in 5 parts and lasts about 1 h (see Fig. 1): the first part is the vocabulary's WISC subtest, the second part is about to learn the Esperanto lesson, the third part is the code's WISC subtest, the fourth part is a quiz about the lesson, and the last part is the letter-number sequencing's WISC subtest. WISC stands for *Wechsler Intelligence Scale for Children* and is the most widely used IQ test for children (especially in the US, Canada, and Europe). It is composed by several subtests, each of them estimating precise cognitive abilities. The vocabulary subtest estimates the verbal comprehension, fluency and word-knowing, the code part estimates a learning factor and the non-verbal working memory speed, and the letter-number sequencing subtest estimates the verbal auditory recall (working



**Fig. 1.** The 5 parts of our experiment: blue ones are recorded with the eye-tracker.

memory). The second and forth part of the experiment (lesson + exam) are about to estimate a learning factor and the recall memory. The participants are evaluated on the translation of words and sentences, on the differences between the two alphabets, and on some elements of the grammar. We chose to use the WISC subtests at distracting tasks to limit the primacy-recency bias.

## 4 Analysis

Many variables have been collected through our experiment: the WISC subtests score, the gaze data, the results of the exam. We studied the distribution of the gaze data with the Shapiro-Wilk test, and found that some gaze characteristics do not follow the theoretical normal distribution. We thus chose to use an ANOVA permutation test as it is a non-parametric test. We considered 21 global gaze metrics including the normalized sum, mean and standard deviation of fixation duration, saccade horizontal amplitude, vertical amplitude, vectorial amplitude, absolute and relative angles, and pupil dilatation as defined in [8, 13]. We used the *aovp()* R function from the *lmpperm* package to perform our tests. We tried to explain the Esperanto test score or the WISC subtest scores by finding an interesting combination of factors (gaze metrics). Due to the combinatorial complexity we tested up to 6 factors at the same time and saved the models when all the factors were significant (*p*-value < 0.05). Figure 2 shows an example of gaze features that are highly correlated to the global exam score.

	Df	R Sum Sq	R Mean Sq	Iter	Pr(Prob)
MedianFixationDuration	1	351.17	351.17	5000	<2e-16 ***
ScanpathLengthNorm	1	204.33	204.33	5000	0.0026 **
SumRelativeAngles	1	282.22	282.22	5000	<2e-16 ***
StandardDeviationAbsoluteAngle	1	431.80	431.80	5000	<2e-16 ***
Residuals	17	320.76	18.87		

**Fig. 2.** One output of the *aovp()* function.

## 5 Conclusion

Regarding our preliminary results in Fig. 2, it seems that a link exists between some gaze metrics and the global score obtained at the Esperanto exam. These metrics are promising to predict the global learning quality. However, we found out that other combinations of gaze characteristics actually explain this variable. As a perspective, our ambition is to identify, map and order all gaze parameters based on their ability to predict the quality of the recall memory. In addition, beyond the global memorization score, we aim at developing a machine learning technique that can accurately distinguish what is learned from what is forgotten. We plan to extend this pilot study by passing our study to entire classes in our academy, so as to increase the statistic power of our tests. We hope that these preliminary study is a first step toward recommender systems based on the memory of users.

## References

1. Aben, B., Stapert, S., Blokland, A.: About the distinction between working memory and short-term memory. *Front. Psychol.* **3**, 301 (2012)
2. Baddeley, A.: The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* **4**(11), 417–423 (2000). [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2)
3. Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R., Bouchet, F.: Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 229–238. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_24](https://doi.org/10.1007/978-3-642-39112-5_24)
4. Borkin, M.A., et al.: Beyond memorability: visualization recognition and recall. *IEEE Trans. Vis. Comput. Graph.* **22**(1), 519–528 (2016)
5. Bylinskii, Z., Borkin, M.A., Kim, N.W., Pfister, H., Oliva, A.: Eye fixation metrics for large scale evaluation and comparison of information visualizations. In: Burch, M., Chuang, L., Fisher, B., Schmidt, A., Weiskopf, D. (eds.) ETVIS 2015. MV, pp. 235–255. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-47024-5\\_14](https://doi.org/10.1007/978-3-319-47024-5_14)
6. Clariana, R.B., Lee, D.: The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educ. Technol. Res. Dev.* **49**(3), 23–36 (2001)
7. Hannula, D.E., Ranganath, C.: The eyes have it: hippocampal activity predicts expression of memory in eye movements. *Neuron* **63**, 592–599 (2009)
8. Holland, C., Komogortsev, O.V.: Biometric identification via eye movement scanpaths in reading. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8, October 2011. <https://doi.org/10.1109/IJCB.2011.6117536>
9. Just, M.A., Carpenter, P.A.: Eye fixations and cognitive processes. *Cogn. Psychol.* **8**, 441–480 (1976)
10. Komogortsev, O., Gobert, D., Jayarathna, S., Koh, D., Gowda, S.: Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Trans. Biomed. Eng.* **57**, 2635–2645 (2010). <https://doi.org/10.1109/TBME.2010.2057429>

11. Komogortsev, O.V., Jayarathna, S., Koh, D.H., Gowda, S.M.: Qualitative and quantitative scoring and evaluation of the eye movement classification algorithms. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA 2010, pp. 65–68. ACM, New York (2010). <https://doi.org/10.1145/1743666.1743682>
12. Shiffrin, R.M., Atkinson, R.C.: Storage and retrieval processes in long-term memory. *Psychol. Rev.* **76**, 179–193 (1969). <https://doi.org/10.1037/h0027277>
13. Marchal, F., Castagnos, S., Boyer, A.: First attempt to predict user memory from gaze data. *Int. J. Artif. Intell. Tools* **27**(6), 1850029 (2018)
14. Miyake, A., Friedman, N., Emerson, J.M., Witzki, A., Howerter, A., Wager, T.: The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000). <https://doi.org/10.1006/cogp.1999.0734>
15. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, ETRA 2000, pp. 71–78. ACM, New York (2000). <https://doi.org/10.1145/355017.355028>
16. Sharafi, Z., Shaffer, T., Sharif, B., Guéhéneuc, Y.: Eye-tracking metrics in software engineering. In: 2015 Asia-Pacific Software Engineering Conference (APSEC), pp. 96–103, December 2015. <https://doi.org/10.1109/APSEC.2015.53>
17. Squire, L.R., Dede, A.J.O.: Conscious and unconscious memory systems. *Cold Spring Harbor Perspect. Biol.* **7**(3), a021667 (2015)
18. Steichen, B., Wu, M.M.A., Toker, D., Conati, C., Carenini, G.: Te,Te,Hi,Hi: eye gaze sequence analysis for informing user-adaptive information visualizations. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 183–194. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-08786-3\\_16](https://doi.org/10.1007/978-3-319-08786-3_16)



# What Inquiry with Virtual Labs Can Learn from Productive Failure: A Theory-Driven Study of Students' Reflections

Charleen Brand<sup>1</sup>✉, Jonathan Massey-Allard<sup>2</sup>, Sarah Perez<sup>2</sup>, Nikol Rummel<sup>1</sup>, and Ido Roll<sup>2</sup>

<sup>1</sup> Ruhr-Universität Bochum, Bochum, Germany

{charleen.brand, nikol.rummel}@rub.de

<sup>2</sup> University of British Columbia, Vancouver, BC, Canada

jmassall@phas.ubc.ca, {sarah.perez, ido.roll}@ubc.ca

**Abstract.** During inquiry learning with virtual labs students are invited to construct mathematical models that capture key features of the underlying structures. However, students typically fail to construct complete models. In order to identify ways to support learners without restricting them, we look at the literature of Productive Failure and Invention activities (often termed PS-I, Problem Solving before Instruction). PS-I activities are designed to facilitate specific cognitive mechanisms that aid learning. This paper seeks to (1) evaluate in what ways PS-I activities compare to inquiry learning, (2) whether students in inquiry learning report similar processes to PS-I, and (3) whether these are associated with better learning. We begin by synthesizing the two approaches in order to highlight their similarities. Following, we coded self-reported post-activity reflections by 139 students who worked with two virtual labs. Students reported processes that are typical to PS-I and, out of these, prior knowledge activation was associated with constructing more complete models. Based on this, we suggest ways to support students in learning from their inquiry.

**Keywords:** Inquiry learning · Invention activities · Productive failure · Virtual labs · Exploratory learning environments

## 1 Introduction

In inquiry learning with virtual labs, students learn about scientific phenomena by using interactive simulations to construct mathematical models of the underlying structures of a target topic [9, 10, 12]. While virtual labs have been shown to support students' inquiry processes [7, 10], many students struggle to construct complete models [6, 10]. The challenge is thus to provide an appropriate level of guidance without mitigating the benefits of authentic inquiry afforded by the virtual labs. For this, we turn towards a similar framework, which offers students support without limiting their exploration – Invention and Productive Failure (also termed PS-I, Problem Solving before Instruction) [3, 11, 14, 19]. Both approaches invite students to model unfamiliar target topics by generating mathematical representations [5, 13, 20]. While both also show differences (e.g. data collection in inquiry learning; a subsequent instruction in PS-I) [13] the

similarity in the activities suggests that they may also share cognitive processes important for learning. Thus, research on PS-I might provide insights into the challenges of understanding and supporting inquiry learning with virtual labs, which could yield valuable design implications such as intelligent agents (cf. [1] for exploratory learning environments in invention), which prompt relevant cognitive processes.

The goals of this paper are therefore to (1) establish a theoretical synthesis between PS-I and inquiry learning with virtual labs, (2) identify evidence for PS-I processes in inquiry with virtual labs, and (3) begin evaluating how these relate to the outcome of the inquiry process, i.e. the students' final model of the target concept. We approach these goals by first providing a mapping of PS-I and the inquiry phases and then applying this mapping to a study of students' self-reported reflections following an inquiry activity. Finally, we discuss potential options for supporting students in virtual labs.

## 2 Mapping PS-I to Inquiry

In the following, we consolidate the similarities of PS-I and inquiry with virtual labs with regards to their cognitive processes. Based on the overview of inquiry processes in the meta-study of Pedaste et al. [16], we align central processes in PS-I with their corresponding phases of inquiry with virtual labs. We outline processes that are key to PS-I (*P*), inquiry (*I*) or common to all modeling activities (*M*).

**Conceptualization.** In this phase, students in inquiry generate ideas and hypotheses about the topic [9, 16]. This lines up with the generation of intuitive ideas for inventing different representations and solution methods in PS-I [11, 18]. In both frameworks prior knowledge is used to generate ideas, however PS-I activities are specifically designed to activate prior knowledge, as we assume that *prior knowledge activation*(*P*) is a key learning mechanism that helps students to attend to the underlying structures of the target concept [11, 14]. While this is not the case for inquiry learning with virtual labs, prior knowledge activation could also function as key learning mechanism.

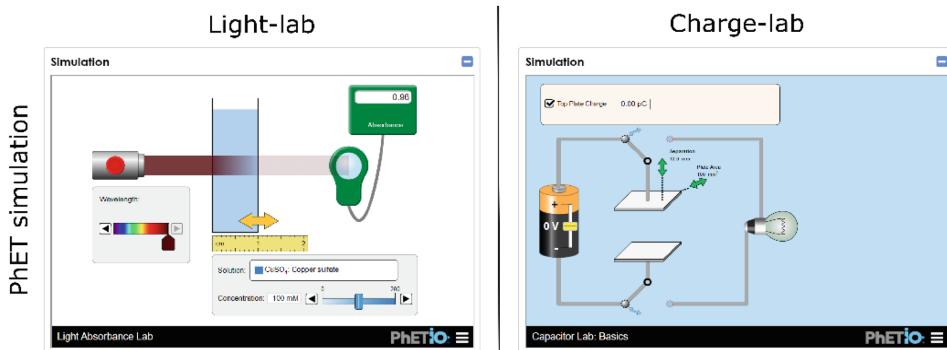
**Investigation.** For data analysis, students in inquiry first need to engage in *exploration* (*I*) and *experimentation* (*I*), i.e. collecting data by running hypothesis-driven experiments [16], and use strategies such as the *control of variables strategy* (CVS) (*I*), in which the effect of one variable is isolated by holding all other variables constant [2]. With the help of the virtual lab, they are expected to *interpret visual feedback* (*I*) (i.e. observing results of an experiment or plotted graph) [10, 16]. These processes are key to inquiry and different from PS-I, in which data is provided. In PS-I, students analyze the given data sets by inventing and contrasting different solutions. Presumably, this helps them to *identify deep features* (*P*) of the underlying structures and, thus, to construct more complete models [8, 11, 14, 18]. Given that students in inquiry might contrast and compare different data sets as well, we expect that students who explicitly notice deep features during the inquiry process show a more successful inquiry process. As in all modeling activities, data analysis and interpretation help students to *model* (*M*) patterns and relationships between variables.

**Conclusion.** In both inquiry, PS-I and other modeling activities, students are expected to *draw conclusions* (*M*) and *evaluate their domain knowledge* (*M*) by comparing present to prior knowledge [9, 16]. As students in PS-I mostly struggle to construct complete models, we assume that evaluating their process raises an *awareness of knowledge gaps* (*P*) [14], which is assumed to be key to learning as it might motivate students to close their gaps [14, 21]. As students in inquiry struggle to construct complete models as well [10], they likely become aware of gaps, which could support them in restructuring their models, facilitating learning.

In the above mapping, we substantiated that inquiry learning in virtual labs might share central learning mechanism with PS-I. In the following analyses, we aim to find evidence for these processes during inquiry learning with virtual labs. We hypothesize that students who report having shown *prior knowledge activation* (*P*), *identification of deep features* (*P*) and *awareness of knowledge gaps* (*P*) present more complete models at the end of the inquiry process, and by this exhibit a more productive inquiry.

### 3 Methods

The data for this paper was taken from a study by Perez et al. [17]. Participants were first and second year undergraduate students from a large Canadian university ( $N = 139$ ). Students were randomly assigned to one of two structurally similar virtual lab activities (see Fig. 1) on light absorbance or charge of parallel plate capacitors.



**Fig. 1.** Interface of the absorbance lab (left) and capacitor lab (right)

Before and during the virtual lab activity students were asked to predict which factors determined the dependent variable (i.e. light absorbance or charge on plate capacitor) and how. This allows us to compare students' models on the target concept prior to the activity (pre-model) and after inquiry (main-model). We analyzed students' reflection on their processes (i.e. how did they learn through the lab) and strategy-use (i.e. what worked well) with a binary coding scheme (1 for reported process, 0 for not reported) based on our mapping. It included ten PS-I, inquiry and general modeling activity processes (see *italics* above). A second rater coded 11.51% of the data with high levels of agreement for almost all ten variables (Cohen's Kappa .75 to 1; *modeling* and *identifying deep features* .59 to .63) [15]. Students' pre- and main-models were

analyzed based on correct qualitative and quantitative relationships with the dependent variables, with overall scores range from 0 to 1 for pre-model and 0 to 2 for main-model. Inter-rater reliability of 27% of the original data was satisfactory (Cohen's Kappa .75 to .86) [17].

## 4 Results

Descriptive analyses showed that PS-I processes indeed were common in students' reflections on the inquiry activity. Out of 139 students, 89 students reported at least one out of three PS-I processes, with an average of .82 ( $SD = .74$ ) processes, compared to an average of 1.96 ( $SD = .78$ ) out of four reported inquiry and an average of 0.99 ( $SD = .87$ ) out of three modeling activity processes. Students' model improved significantly after inquiry (Wilcoxon:  $Z = -9.268, p < .0001$ , *effect size* = .786), showing on average .30 points ( $SD = .39$ ; *min* = 0, *max* = 1) for pre-model and 1.19 points ( $SD = .57$ ; *min* = 0, *max* = 2) for main-model. We calculated two linear regression analyses between students' main model as dependent variable and (1) grouped PS-I, inquiry and modeling activity processes, and (2) single PS-I processes. Both controlled for pre-model scores, the virtual lab topic and ruled out multicollinearity (tolerance statistics .90 to .99), which means that independent and controlled variables (e.g. pre-model and PS-I processes) were not related to each other [4]. The reported, grouped PS-I processes did not reach significance,  $\beta = .147, p = .092$ , but showed a higher association with main-model than inquiry and modeling processes ( $\beta = -.009, p = .919$  and  $\beta = .042, p = .629$  respectively). For the single PS-I processes, only *prior knowledge activation* predicted main-model scores significantly,  $\beta = .225, p = .009$ .

## 5 Conclusion

In order to find new opportunities to support students' inquiry, this study emphasizes the similarities of two bodies of literature in theory and practice: PS-I and inquiry learning. Our hypotheses were partially confirmed. Students indeed reported PS-I processes in their post-activity reflections. Of all processes, only reported *prior knowledge activation* significantly predicted learning. Our findings are a first indicator that PS-I processes might also occur in inquiry with virtual labs and that *prior knowledge activation* could be associated with a more complete model at the end of the virtual lab activity. Also, our tests for multicollinearity (i.e. interdependence between independent and control variables) suggest that activating prior knowledge was unrelated to actual levels of prior knowledge. That is, students who activated their prior knowledge learned more, regardless of their level of knowledge. Our results implicate new design opportunities for virtual labs in inquiry learning. For instance, with the help of intelligent agents, students could be prompted to activate prior knowledge throughout the activity by asking them to generate multiple hypotheses to the task.

However, due to the use-of self-reported reflections of students' processes, which do not necessarily correspond to the processes students really showed, this study cannot

make claims on how frequent PS-I processes occur in inquiry with virtual labs, but rather provides first evidence that they occur at all. Thus, future studies need to substantiate our findings with the help of controlled experiments and process data.

## References

1. Chase, C., Marks, J., Bennett, D., Aleven, V.: The Design of an exploratory learning environment to support invention. In: AIED Workshops (2015)
2. Chen, Z., Klahr, D.: All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* **70**(5), 1098–1120 (1999). <https://doi.org/10.1111/1467-8624.00081>
3. Chowrira, S.G., Smith, K.M., Dubois, P.J., Roll, I.: DIY productive failure: boosting performance in a large undergraduate biology course. *NPJ Sci. Learn.* **4**, 1 (2019). <https://doi.org/10.1038/s41539-019-0040-6>
4. Farrar, D.E., Glauber, R.R.: Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* **49**(1), 92 (1967). <https://doi.org/10.2307/1937887>
5. Ford, K.M., Bradshaw, J.M., Adams-Webber, J.R., Agnew, N.M.: Knowledge acquisition as a constructive modeling activity. *Int. J. Intell. Syst.* **8**(1), 9–32 (1993). <https://doi.org/10.1002/int.4550080103>
6. Glaser, R., Schauble, L., Raghavan, K., Zeitz, C.: Scientific reasoning across different domains. In: De Corte, E., Linn, M.C., Mandl, H., Verschaffel, L. (eds.) Computer-Based Learning Environments and Problem Solving, pp. 345–371. Springer, Heidelberg (1992). [https://doi.org/10.1007/978-3-642-77228-3\\_16](https://doi.org/10.1007/978-3-642-77228-3_16)
7. Hmelo-Silver, C.E., Duncan, R.G., Chinn, C.A.: Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* **42**(2), 99–107 (2007). <https://doi.org/10.1080/00461520701263368>
8. Holmes, N.G., Day, J., Park, A.H.K., Bonn, D.A., Roll, I.: Making the failure more productive: scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instr. Sci.* **42**(4), 523–538 (2014). <https://doi.org/10.1007/s11251-013-9300-7>
9. de Jong, T.: Scaffolds for scientific discovery learning. In: Handling Complexity in Learning Environments: Theory and Research, pp. 107–128 (2006)
10. de Jong, T., van Joolingen, W.R.: Scientific discovery learning with computer simulations of conceptual domains. *Rev. Educ. Res.* **68**(2), 179 (1998). <https://doi.org/10.2307/1170753>
11. Kapur, M., Bielaczyc, K.: Designing for productive failure. *J. Learn. Sci.* **21**(1), 45–83 (2012). <https://doi.org/10.1080/10508406.2011.591717>
12. Lazender, A.W., Hagemans, M.G., de Jong, T.: Offering and discovering domain information in simulation-based inquiry learning. *Learn. Instr.* **20**(6), 511–520 (2010). <https://doi.org/10.1016/j.learninstruc.2009.08.001>
13. Loibl, K., Rummel, N.: The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instr. Sci.* **42**(3), 305–326 (2014). <https://doi.org/10.1007/s11251-013-9282-5>
14. Loibl, K., Roll, I., Rummel, N.: Towards a theory of when and how problem solving followed by instruction supports learning. *Educ. Psychol. Rev.* **29**(4), 693–715 (2017). <https://doi.org/10.1007/s10648-016-9379-x>
15. McHugh, M.L.: Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012). <https://doi.org/10.11613/bm.2012.031>
16. Pedaste, M.: Phases of inquiry-based learning: definitions and the inquiry cycle. *Educ. Res. Rev.* **14**, 47–61 (2015). <https://doi.org/10.1016/j.edurev.2015.02.003>

17. Perez, S., et al.: Control of variables strategy across phases of inquiry in virtual labs. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 271–275. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_50](https://doi.org/10.1007/978-3-319-93846-2_50)
18. Roll, I., Holmes, N.G., Day, J., Bonn, D.: Evaluating metacognitive scaffolding in guided invention activities. *Instr. Sci.* **40**(4), 691–710 (2012). <https://doi.org/10.1007/s11251-012-9208-7>
19. Schwartz, D.L., Martin, T.: Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. *Cogn. Instr.* **22**(2), 129–184 (2004). [https://doi.org/10.1207/s1532690xci2202\\_1](https://doi.org/10.1207/s1532690xci2202_1)
20. Schwartz, D.L., Chase, C.C., Oppezzo, M.A., Chin, D.B.: Practicing versus inventing with contrasting cases: the effects of telling first on learning and transfer. *J. Educ. Psychol.* **103**(4), 759–775 (2011). <https://doi.org/10.1037/a0025140>
21. VanLehn, K.: Toward a theory of impasse-driven learning. In: Mandl, H., Lesgold, A. (eds.) Learning Issues for Intelligent Tutoring Systems, pp. 19–41. Springer, New York (1988). [https://doi.org/10.1007/978-1-4684-6350-7\\_2](https://doi.org/10.1007/978-1-4684-6350-7_2)



# The Role of Achievement Goal Orientation on Metacognitive Process Use in Game-Based Learning

Elizabeth B. Cloude<sup>1</sup>(✉), Michelle Taub<sup>1</sup>, James Lester<sup>2</sup>,  
and Roger Azevedo<sup>1</sup>

<sup>1</sup> University of Central Florida, Orlando, FL 32816, USA

elizabeth.cloude@knights.ucf.edu,

{michelle.taub, roger.azevedo}@ucf.edu

<sup>2</sup> North Carolina State University, Raleigh, NC 27695, USA

lester@ncsu.edu

**Abstract.** To examine relations between achievement goal orientation—a construct of motivation, metacognition and learning, multiple data channels were collected from 58 students while problem solving in a game-based learning environment. Results suggest students with different goal orientations use metacognitive processes differently but found no differences in learning. Findings have implications for measuring motivation using multiple data channels to design adaptive game-based learning environments.

**Keywords:** Motivation · Metacognition · Game-based learning environments

## 1 Introduction

Students engage in self-regulation by monitoring and adjusting cognition, affect, metacognition, and motivation to attain learning goals [1]. Game-based learning environments (GBLEs) are effective tools for addressing the educational challenges of the 21st century and preparing the future workforce of the United States [2–4]. Research on GBLEs reveals students are more likely to achieve learning objectives and demonstrate more engagement while problem solving compared to classrooms [5, 6]. Research suggests students with different motivational states use SRL processes differently, revealing differences in learning outcomes [7, 8]. This study examined relationships between AGO and metacognitive process use by analyzing multiple data channels in conjunction with self-report and performance data before, during, and after problem solving with CRYSTAL ISLAND (CI).

## 2 Methods

### 2.1 Participants, Materials, and Experimental Procedure

58 undergraduates from a North American university participated in the study ( $M_{age} = 20.12$ ,  $SD = 1.57$ ), and students were compensated \$10/hr. Upon consent, students were

randomly assigned to one of three conditions, but the control condition was only analyzed. Self-report measures, demographics, and a 21-item, multiple-choice pretest and posttest ( $M_{pre} = .58$ ,  $SD = .13$ ;  $M_{post} = .68$ ,  $SD = .14$ ) were administered before and after problem solving with CI [5]. The Achievement Goal Questionnaire-Revised (AGQ-R) [9] was the only self-report data included in analyses ( $\alpha > .84$ ). CI is a narrative-based GBLE where students play the role of a scientist to identify a pathogen source by interacting with non-player characters, reading books and articles, and scanning food items. Students were given tools to foster SRL processes: (1) concept matrix and (2) diagnosis worksheet. Students had to submit a correct diagnosis worksheet to complete the game. Students sat in front of a computer where they completed pretest materials and problem solved with CI ( $M = 81$  min,  $SD = 23$ ) and then completed a posttest.

## 2.2 Coding and Scoring

A proportional learning gain formula that considers prior knowledge while calculating differences between pre and posttest scores was used ( $M = .22$ ,  $SD = .33$ ) [10]. Total metacognitive processes were extracted from log files of all student actions for analyses. AGQ-R scores were summed and separated into four scores: mastery, performance, approach and avoidance. Two grouping variables with three levels each: (1) mastery, performance, and combined mastery and performance and (2) approach, avoidance, and combined approach and avoidance were created, and students were assigned based on how high they scored compared to other levels, where if students scored less than a 2-pt difference, they were assigned to the combination group.

# 3 Results

## 3.1 RQ1: Are There Differences Between AGO Groups on Proportional Learning Gain (PLG) After Problem Solving with CI?

A one-way ANOVA was conducted to assess if there were significant differences in PLG between AGO groups after problem solving with CRYSTAL ISLAND. Our results found no significant differences in PLG between AGO groups ( $p > .05$ ).

## 3.2 RQ2: Are There Differences Between AGO Groups on the Frequency of Metacognitive Process Use While Problem Solving with CI?

A nonparametric Friedman test was conducted to examine differences between AGO groups on frequency of using metacognitive processes with CRYSTAL ISLAND. Our analysis revealed significant differences between AGO groups in frequency of metacognitive process use,  $\chi^2(5) = 207.52$ ,  $p = .000$ . These findings support our hypothesis where we expected to see differences in frequency of metacognitive processes between AGO groups. See Table 1 for mean ranks between groups. Follow up related-samples Wilcoxon signed rank tests revealed differences between AGO groups on the frequency of reading complex text (i.e., research articles and books combined),

between mastery, performance, and combined mastery and performance orientations ( $z = 6.627, p = .000, r = .87$ ) and approach, avoidance, and combined approach and avoidance orientations ( $z = 6.627, p = .000, r = .87$ ). There were also differences in frequency of using the concept matrix between mastery, performance, and combined mastery and performance orientations ( $z = 6.627, p = .000, r = .87$ ) as well as approach, avoidance, and combined approach and avoidance orientations ( $z = 6.627, p = .000, r = .87$ ). Analyses revealed differences in the frequency of scanning food items between mastery, performance, and combined mastery and performance orientations ( $z = 6.625, p = .000, r = .87$ ) and approach, avoidance, and combined approach and avoidance orientations ( $z = 6.624, p = .000, r = .87$ ). Additional analyses found differences in frequency of submitting diagnosis worksheets between mastery, performance, and combined mastery and performance orientations ( $z = 6.569, p = .000, r = .86$ ) and approach, avoidance, and combined approach and avoidance orientations ( $z = 6.568, p = .000, r = .86$ ).

**Table 1.** Mean ranks of metacognitive process use between AGO groups.

Groups	Metacognitive process use			
	Complex text	Concept matrix	Diagnosis worksheet	Food item scans
Mastery	18.21	18.21	3.36	15.93
Performance	13.71	13.71	2.79	16.00
Mastery/Performance combination	13.64	13.64	5.29	16.86
Approach	16.64	16.64	3.07	20.14
Avoidance	12.43	12.43	3.57	16.07
Approach/Avoidance combination	15.71	15.71	4.50	11.71

### 3.3 RQ3: Are There Differences Between AGO Groups on the Proportion of Time Engaging in Metacognitive Processes While Problem Solving with CI?

A nonparametric Friedman test was calculated to examine differences between AGO groups on the proportion of time engaging in metacognitive processes while problem solving with CRYSTAL ISLAND. Analysis revealed differences between AGO groups on proportion of time engaging in metacognitive processes,  $\chi^2(5) = 274.08, p = .000$ . See Table 2 for mean ranks between groups. Follow up related-samples Wilcoxon signed rank tests revealed differences in proportion of time in reading (e.g., research articles and books) between mastery, performance, and combined mastery and performance groups ( $z = -6.624, p = .000, r = -.87$ ) and approach, avoidance, and combined approach and avoidance groups ( $z = -6.624, p = .000, r = -.87$ ). There were differences in proportion of time using the concept matrix between mastery, performance and combined mastery and performance orientations ( $z = -6.624, p = .000, r = -.87$ ) and approach, avoidance and combined approach and avoidance groups ( $z = -6.624,$

$p = .000, r = -.87$ ). Analyses also found differences in proportion of time using the diagnosis worksheet between mastery, performance, and combined mastery and performance groups ( $z = -6.624, p = .000, r = -.87$ ) and approach, avoidance, and combined approach and avoidance groups ( $z = -6.624, p = .000, r = -.87$ ). There were also differences between mastery, performance and combined mastery and performance groups in proportion of time scanning food items ( $z = -6.624, p = .000, r = -.87$ ) and approach, avoidance, and combined approach and avoidance groups ( $z = -6.624, p = .000, r = -.87$ ).

**Table 2.** Mean ranks for proportional duration of metacognitive use between AGO groups.

Groups	Metacognitive process use			
	Complex text	Concept matrix	Diagnosis worksheet	Food item scans
Mastery	21.43	2.57	15.86	6.29
Performance	22.00	5.57	13.86	9.43
Mastery/Performance combination	20.79	7.64	16.36	7.50
Approach	21.21	4.36	15.50	8.21
Avoidance	22.07	7.07	15.64	8.79
Approach/Avoidance combination	21.36	5.64	14.64	6.21

### 3.4 RQ4: Do AGO Scores Predict Frequency and Proportion of Time Engaging in Metacognitive Processes While Problem Solving with CI?

Analyses revealed a significant linear regression where AGQ-R scores predicted proportion of time engaging in metacognitive processes,  $F(4, 54) = 7.202, p = .000$  with an  $R^2$  of .286. Specifically, the higher mastery-oriented students were, less time was used on the concept matrix ( $\beta = -.827, p = .000$ ), while the higher avoidance-oriented students were, more time was used on the concept matrix ( $\beta = .544, p = .005$ ).

## 4 Discussion

Examining how achievement goal orientation affects metacognition and learning is the first step to understanding how motivation affects SRL processes while problem solving with GBLEs. Understanding what personally motivates students to learn and factors which influence motivation could propel the development of adaptive GBLEs that consider the students' motivational needs to maximize metacognitive process use and learning. Future research should use multiple data channels instead of relying on self-report and performance data collected *before* and *after* problem solving as it does not capture changes in motivation. If GBLEs could detect motivation by analyzing eye-gaze behaviors, concurrent verbalizations, and facial expressions, the system could

detect motivational changes based on how students interact with features of the system and adapt features to meet motivational needs. However, the first step is operationalizing motivation as dynamic and complex states that are likely to change across tasks.

**Acknowledgements.** This research was funded by the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2011-1006). Authors would like to thank members of the SMART Lab and intelliMEDIA at NCSU for their assistance and contributions.

## References

1. Azevedo, R., Taub, M., Mudrick, N.: Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In: Alexander, P.A., Schunk, D.H., Greene, J.A. (eds.) *Handbook of Self-Regulation of Learning and Performance*, 2nd edn. Routledge, New York (2018)
2. The National Academies of Sciences, Engineering, and Medicine. <https://doi.org/10.17226/24783>. Accessed 08 Feb 2019
3. The National Academies of Sciences, Engineering, and Medicine. <https://doi.org/10.17226/13398>. Accessed 08 Feb 2019
4. The National Academies of Sciences, Engineering, and Medicine. <https://doi.org/10.17226/13078>. Accessed 08 Feb 2019
5. Rowe, J., Shores, L., Mott, B., Lester, J.: Integrating learning, problem solving, and engagement in narrative-centered learning environments. *Int. J. Artif. Intell. Educ.* **21**(1–2), 115–133 (2011)
6. Winne, P.: Cognition and metacognition within self-regulated learning. In: Alexander, P.A., Schunk, D.H., Greene, J.A. (eds.) *Handbook of Self-Regulation of Learning and Performance*, 2nd edn. Routledge, New York (2018)
7. Vaessen, B., Prins, F., Jeuring, J.: University students' achievement goals and help seeking strategies in an intelligent tutoring system. *Comput. Educ.* **72**(31), 196–208 (2014)
8. Cloude, E.B., Taub, M., Azevedo, R.: Investigating the role of goal orientation: metacognitive and cognitive strategy use and learning with intelligent tutoring systems. In: Nkambou, R., Azevedo, R., Vassileva, J. (eds.) *ITS 2018. LNCS*, vol. 10858, pp. 44–53. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91464-0\\_5](https://doi.org/10.1007/978-3-319-91464-0_5)
9. Elliot, A., Murayama, K.: On the measurement of achievement goals: critique, illustration, and application. *J. Educ. Psychol.* **100**(3), 613–628 (2008)
10. Witherspoon, A.M., Azevedo, R., D'Mello, S.: The dynamics of self-regulatory processes within self-and externally regulated learning episodes during complex science learning with hypermedia. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) *ITS 2008. LNCS*, vol. 5091, pp. 260–269. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69132-7\\_30](https://doi.org/10.1007/978-3-540-69132-7_30)



# Autoencoders for Educational Assessment

Geoffrey Converse<sup>1</sup>(✉), Mariana Curi<sup>2</sup>, and Suely Oliveira<sup>1</sup>

<sup>1</sup> University of Iowa, Iowa City, IA, USA

{geoffrey-converse,suely-oliveira}@uiowa.edu

<sup>2</sup> University of São Paulo, São Carlos, Brazil

mcuri@icmc.usp.br

**Abstract.** In educational assessment research, a common goal is to determine students' knowledge about some construct. This knowledge is latent and can be represented by continuous variables which influence the individual's performance on a test. Item response theory (IRT) models structure this relation, defining specific functions between the knowledge of the individual, and the probability of answering an item correctly. Previous research implies that neural networks can emulate these models, and, with a modification in its architecture, overcome some of the limitations concerned to "big data" analysis. In this work, we compare two different types of neural networks for this application: autoencoders (AE) and variational autoencoders (VAE). Not only can these neural networks be used as similar predictive models, but they can recover and interpret parameters in the same way as in the IRT approaches.

**Keywords:** Neural networks · Interpretability · Cognitive models · Item response theory

## 1 Introduction

Under the framework of autoencoders, Guo, Cutumisu and Cui [4] proposed a modification of AE methods [3] to estimate skill mastery in cognitive diagnostic assessment. Their method incorporates the *Q*-matrix (which defines the relationship between the latent skills and the assessment questions) to define connections between the nodes of the final hidden layer with the output layer.

Xue [8] also proposed the use of modified AE to partially avoid the necessity of defining the *Q*-matrices in evaluating cognitive diagnostics. Curi et al. [1] proposed the use of VAE incorporating the *Q*-matrix and a MIRT model with two item parameters [6] in the decoder. The objective of this work is to compare AE and VAE methods for simulated educational testing data using a *Q*-matrix to define the connection between the continuous latent variables and the items. In theory, VAE is better suited than AE to the structure of educational testing because of the stochastic assumption for the latent traits. This enables VAE to learn variance (of the latent trait estimates), a very important information indicative of the quality of the data analysis.

## 2 Background

### 2.1 Educational Assessment Models

A very popular model in IRT is the multidimensional logistic 2-parameter (ML2P) model [5], designed to test subject's performance in  $J$  latent traits on an assessment with  $I$  items. Denote  $\Theta_k = (\theta_{k1}, \dots, \theta_{kJ})$  as the set of latent skills of subject  $k$ . Larger values of  $\theta_{kj}$  represent a higher knowledge of that skill. Let the binary variable  $X_{ki}$  denotes the student  $k$  answer to item  $i$ . The ML2P model defines the probability of success as

$$P(X_{ki} = 1 | \Theta_k) = \frac{1}{1 + \exp\left(-\sum_{j=1}^J a_{ji}\theta_{kj} + b_i\right)} \quad (1)$$

with discrimination parameters  $a_{ji}$ , measuring the relation of latent skill  $j$  with assessment item  $i$ , and difficulty parameter  $b_i$ , for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . It is usually assumed that  $a_{ji} \geq 0$  for all  $i, j$ ; intuitively, this means that more skill in a latent area cannot decrease the probability of answering an item correctly. Further, if  $a_{ji} = 0$ , then item  $i$  does not require skill  $j$  to be answered correctly.

In order to keep track of which latent skills are related to each item, a  $Q$ -matrix can be defined [7]. We define  $Q \in \mathbb{R}^{J \times I}$  by  $Q_{ji} = 1$  if item  $i$  requires skill  $j$ , and  $Q_{ji} = 0$  otherwise.

### 2.2 AE and VAE

Autoencoders are a special class of neural networks in which the input and output layers are the same. An AE consists of two neural networks: an encoder and a decoder. The encoder takes the high-dimensional input  $x$ , and feeds forward through one or more hidden layers to some latent space. The decoder takes in this latent dimension, and feeds forward through hidden layers to output a reconstruction of the original input,  $\hat{x}$ .

A variational autoencoder is very similar to a regular autoencoder, with a few important differences. A VAE still consists of an encoder and decoder, but the latent space returned from the encoder is trained to learn a probability distribution for  $\Theta$  given  $x$ . For example, this allows the network to map the training data to a (latent) normal distribution. We then sample from this distribution, and feed forward that sample through the decoder [2].

## 3 Integrated ANN and IRT Models

Our AE and VAE architectures are combined with the ML2P model in the following way: no hidden layer in the decoder, a sigmoidal activation function on the output layer nodes (with non-negative weights), and a  $Q$ -matrix to determine the connections between the latent traits and the output items. Without the last assumption, the latent dimension of the neural network is difficult be interpreted as specific latent traits, and would remain abstract.

The sigmoidal activation function relates the activation of output nodes with the parameters in Eq. (1). Because of this, we can interpret the weights  $w_{ij}$  in the decoder as estimates to the discrimination parameters  $a_{ji}$ , and the biases in the output layer as estimates to the difficulty parameters  $b_i$ .

When training our neural network, we obtain estimates for the subjects' latent traits, as well as for the item's discrimination and difficulty parameters. This relationship between the decoder and model (1) is valuable and define a new approach to validate Q-matrix specification. Based on the decoder, different connections between the latent traits and the output items can be compared. The magnitude of decoder weights reflect the accuracy of the  $Q$ -matrix - if  $w_{ij}$  is close to 0, then item  $i$  may not require skill  $j$  as previously thought.

## 4 Results and Discussion

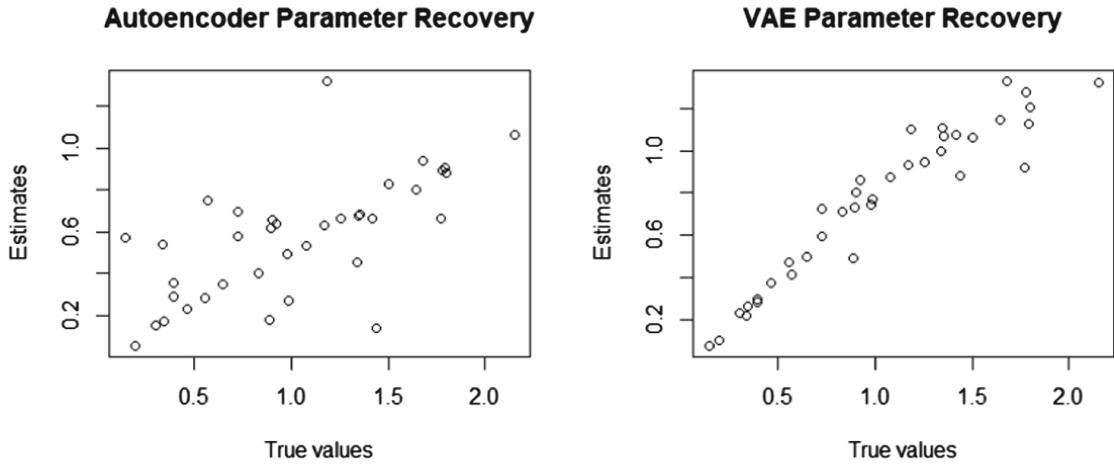
To compare the effectiveness of AE and VAE as a ML2P IRT model for educational assessment, we simulated three continuous latent traits, each generated independently from  $\mathcal{N}(0, 1)$ , for 10,000 students. We simulated 10 sets of answers to a 28 item assessment for each student, relating the items to latent traits using a  $Q$ -matrix specified based on the Examination for the Certificate of Proficiency in English for nonnative English speakers, studied in prior IRT literature. We train an AE and a VAE on each of these responses to obtain estimates  $\hat{\Theta}_k$ ,  $\hat{a}_{ji}$ , and  $\hat{b}_i$  after averaging across the ten replicates.

**Table 1.** Absolute value relative bias (AVRB), root mean square error (RMSE), and correlation (CORR) between the true values and parameter estimates.

Model	$a_1$	$a_2$	$a_3$	$b$	$\theta_1$	$\theta_2$	$\theta_3$	Statistic
AE	0.680	0.227	0.529	2.305	7.425	3.107	16.260	AVRB
VAE	0.284	0.159	0.264	1.894	1.844	1.713	4.009	
AE	0.585	0.481	0.534	1.651	1.788	1.523	1.746	RMSE
VAE	0.322	0.346	0.264	1.670	0.664	0.760	0.646	
AE	0.529	0.547	0.748	0.917	0.970	0.937	0.971	CORR
VAE	0.924	0.920	0.986	0.990	0.965	0.940	0.969	

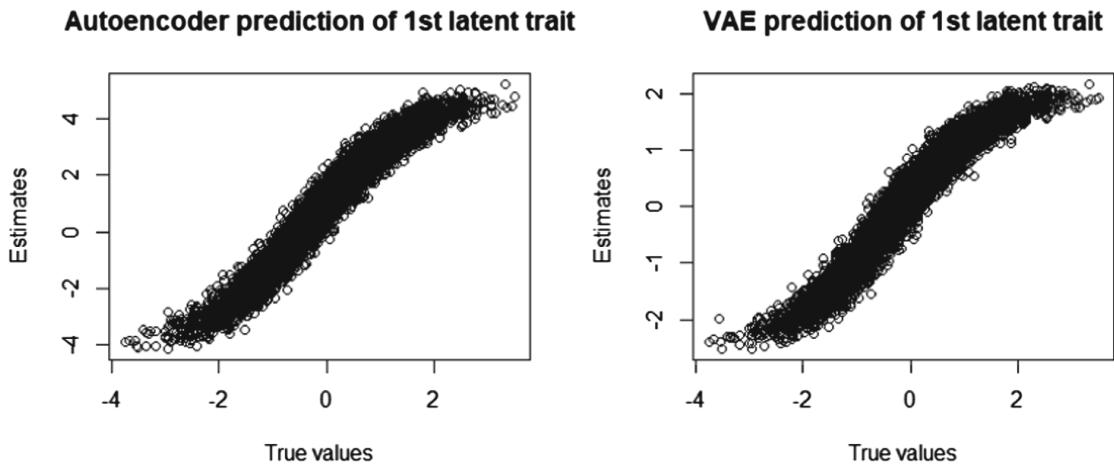
We can see in Table 1 that for parameter recovery, the VAE has much smaller error terms and much higher correlations. This result is corroborated by the correlation plots between the true discrimination parameters and the weights of the decoder, displayed in Fig. 1. There is a linear relationship between the ML2P parameters and the trained parameters in the decoder of both neural networks, though this relationship is much stronger in the VAE.

We also measure the predictive power of latent variables of both neural networks. As seen in Table 1, the error statistics, especially AVRB, are much worse



**Fig. 1.** Autoencoder and VAE discrimination parameter ( $a_{ji}$ ) recovery

in the AE, albeit with high correlation (Fig. 2). Notice that the scale of the estimates is much larger in the AE than the VAE, explaining the difference in error measures. The sigmoidal tendency in the correlation plots of Fig. 2 shows that neither AE or VAE are particularly good at predicting latent traits in the tails of the Normal distribution. A possible explanation is that the simulated difficulty parameters  $b_i$  were sampled uniformly from  $(-3, 3)$ . Therefore, the assessment was not designed to precisely estimate latent traits out of this range.



**Fig. 2.** Autoencoder and VAE predictions for  $\theta_1$

Both AE and VAE show promising results as predictive models for assessing abilities of subjects from test result data. Additionally, these models provide estimates for the discrimination and difficulty parameters in the ML2P model. In this purpose, a VAE provides much more accurate estimates, with lower error and better consistency, along with yielding the standard deviation.

The application of neural networks as educational assessment models also add interpretability to ANN. Typically, the hidden layers of autoencoders and VAE

don't represent concrete ideas. But by using a  $Q$ -matrix to determine weights in the decoder, we can interpret a hidden layer in the neural network as the latent traits of subjects. Further, the weights and biases in the decoder are interpreted as estimates to parameters from other popular models.

## References

1. Curi, M., Converse, G., Hajewski, J., Oliveira, S.: Interpretable variational autoencoders for cognitive models (2019, accepted for publication)
2. Doersch, C.: Tutorial on variational autoencoders. [arXiv:1606.05908](https://arxiv.org/abs/1606.05908) (2016)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2017)
4. Guo, Q., Cutumisu, M., Cui, Y.: A neural network approach to estimate student skill mastery in cognitive diagnostic assessments. In: 10th International Conference on Educational Data Mining (2017)
5. McKinley, R., Reckase, M.: The Use of the General Rasch Model with Multidimensional Item Response Data. American College Testing, Iowa City (1980)
6. Reckase, M.: Multidimensional Item Response Theory. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-89976-3>
7. Tatsuoka, K.K.: Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* **20**(4), 345–354 (1983)
8. Xue, K.: Non-model based attribute profile estimation with partial q-matrix information for cognitive diagnosis using artificial neural network. In: Proceedings of the 11th International Conference on Educational Data Mining (2018)



# The Value of Multimodal Data in Classification of Social and Emotional Aspects of Tutoring

Mutlu Cukurova<sup>(✉)</sup>, Carmel Kent, and Rosemary Luckin

University College London, London, UK  
m.cukurova@ucl.ac.uk

**Abstract.** There are many aspects of tutoring that are associated with social and emotional learning. These are complex processes that involve dynamic combinations of skills, abilities and knowledge. Here, we present the results of our investigation on the particular personal, emotional, and experience traits of tutors who are likely to be successful at social and emotional aspects of tutoring. In particular, we present our approach to measure the social and emotional aspects of tutoring through classification models of 47 candidates' multimodal data from audio and psychometric measures. Moreover, we compare the accuracy of models with unimodal and multimodal data, and show that multimodal data leads to more accurate classifications of the candidates. We argue that when evaluating the social and emotional aspects of tutoring, multimodal data might be more preferable.

**Keywords:** Social and emotional learning · Multimodal data · Tutoring

## 1 Introduction

Today, as educators and as learners, we are faced with many challenges. For example, an increasingly automated and AI augmented world in which children will experience a very different life to that of their parents [12]. We must prepare for the much-anticipated upheaval by ensuring that our education and training is tuned to the new demands of the workplace and society [15]. To achieve this ambitious goal, young learners across the globe should be educated on a broad array of social and emotional skills, attitudes, and values to succeed in school, careers, and in life [1]. Effective tutors are those who can support students on those social and emotional aspects as well as their academic capabilities. However, the evaluation of the social and emotional aspects of tutoring is a challenging task. In this paper, we are investigating the particular personal, emotional, and experience traits of tutors that are likely to make them successful at social and emotional aspects of tutoring. More specifically, we present our classification models of trainee tutors in terms of their success at social and emotional aspects

of tutoring based on a tailored personality questionnaire and their audio data analysis. We also compare the accuracy of models based on unimodal and multimodal data. Although there is an emerging research in multimodal machine learning [2, 13] and multimodal learning analytics [3, 14] to investigate social [4] and emotional [5] aspects of learning, we are not familiar with any previous work undertaken in the complex social context of debate tutoring.

## 2 Methodology

### 2.1 The Context of the Study

In order to create the baseline data, we asked expert tutors to score the social and emotional aspects of 47 tutor candidates with a performance-based activity. In the activity, the candidates were given a tutoring task and three expert tutors observed the candidates' activity and gave them a score from 1 to 5. In these scores, 1 and 2 represent an excellent candidate, who can generally be placed at any school to deliver tutoring, 3 is used for those who might need some further training on these aspects, and 4 and 5 are not desirable candidates. Following the independent scoring, in cases of a discrepancy among experts, the evaluators negotiated their scores and reached a consensus. In addition to these scores, we interviewed expert tutors to get insights on what kind of social and emotional aspects they were observing during their evaluations. Frequently emerging themes for the “expected candidates” were social interactivity, engagement, emotional intelligence, and appropriate encouragement/praise of others.

### 2.2 Data Collection Methods

There is a long-standing wealth of literature characterizing effective tutoring and many of them emphasise the value of social and emotional aspects of tutoring (i.e. [6, 10]). In order to be able collect meaningful and relevant data on social and emotional aspects of tutoring, we collected data on various psychometric measures. More specifically, temperament was represented by two dimensions—social closeness and social anger, that are assessed by the Adult Temperament Questionnaire (ATQ; [8]). Then, the empathy which we consider as a potential representative of the emotional intelligence, we used the Trait Emotional Intelligence Questionnaire (TEIQue-SF; [11]). These personal characteristics aim to reflect the candidates' social and emotional capabilities, specifically, testing their ability to develop social connections, be orientated to communicate and readiness to be exposed to a large volume of social interaction, along with abilities to empathy and emotion control that are all crucial in the context of tutoring. Furthermore, to identify the self-reflection of the candidates on their charismatic abilities, namely, be pervasive, confident and have abilities to make other people comfortable during mutual interactions, which are argued as significant for the effective debate tutoring [7], we used the General Charisma Inventory (GCI; [13]). Finally, the questionnaire capturing the reflection of the tutors on their

abilities to follow plans and commitments and assessed by items utilised from the Big Five Inventory (BFI; [9]). In addition to these, we added two items about candidates' previous experience in debating and tutoring. Further to the tailored questionnaire and experience data, we collected 90sec audio recordings of the candidates while they answer the question "why do you want to become a tutor?" to have input information on their *emotional traits*. To analyse the audio data we use OpenSMILE open source software package. To clean the data, we omitted windows that are smaller than 1600 ms, computed SD for each variable/candidate, omitted samples outliers, and computed mean values without outliers.

### 3 Results

In order to reduce the large set of variables into a smaller set of components, a principal components analysis (PCA) was run on the 22-question questionnaire created. Inspection of the correlation matrix showed that two variables (social closeness and rare social anger) had no correlation coefficient greater than 0.3, thus they were removed from the PCA. All the rest 20 variables had at least one correlation coefficient greater than 0.3 (KMO = .791, Barlett's sphericity was significant  $p < 0.0005$ ). PCA revealed five components that had eigenvalues greater than one and which explained 29.728%, 10.847%, 9.159%, 7.458% and 5.683% of the total variance, respectively. Visual inspection of the scree plot indicated that four components should be retained. The four-component solution explained 57.193% of the total variance. The rotated solution exhibited simple structure with strong loadings of extraversion, outgoingness, and leadership items on component 1, charisma, enthusiasm, and the tendency to make people comfortable items on component 2, assertiveness, organization and the tendency of being influential items on component 3 and neuroticism, non-assertiveness items on component 4. There was not enough data to implement a similar PCA approach for the audio data. However, we omitted highly correlated variables from the data.

#### 3.1 Classifications of Tutor Candidates from Various Data Inputs

Multinomial logistic regression to classify the candidates into three groups (those scored 1 or 2, those scored 3, and those scored 4 or 5) was found to be the best classification tool for all data modalities and variables investigated here. We tested all models' fitting information with pearson chi-square tests. Table 1 below shows the results of the model built with just the two experience variables, ( $df = 14$ ) = 10.73,  $p = 0.707$ . However, the model's fitting information shows that the full model does not significantly predict the scores.

Table 2 shows the results of the classification using only the audio variables. The model's goodness of fit shows that the model fits the data well, ( $df = 56$ ) = 61.63,  $p = 0.282$ . Moreover, the model's fitting information shows that the full model significantly predicts the score ( $df = 24$ ) = 41.72,  $p = 0.014$ . In the

**Table 1.** Classification based on experience variables.

Observed score	Estimated score			Percent correct
	1/2	3	4/5	
1/2	0	9	0	0.0%
3	0	22	0	100.0%
4/5	0	10	0	0.0%
Overall	0.0%	100.0%	0.0%	53.7%

**Table 2.** Classification based on audio variables.

Observed score	Estimated score			Percent correct
	1/2	3	4/5	
1/2	8	1	0	88.9%
3	2	15	5	68.2%
4/5	0	3	7	70.0%
Overall	24.4%	46.3%	29.3%	73.2%

**Table 3.** Classification based on experience and the survey variables.

Observed score	Estimated score			Percent correct
	1/2	3	4/5	
1/2	4	5	0	44.4%
3	3	17	2	77.3%
4/5	1	8	1	10.0%
Overall	19.5%	73.2%	7.3%	53.7%

**Table 4.** Multimodal classification, the experience, survey, and two of the stronger audio variables.

Observed score	Estimated score			Percent correct
	1/2	3	4/5	
1/2	8	1	0	88.9%
3	1	19	2	86.4%
4/5	0	4	6	60.0%
Overall	22.0%	58.5%	19.5%	80.5%

likelihood ratio tests, Interest - passive ( $df = 2$ ) = 15.33,  $p = 0.000$ , Emotion - anger ( $df = 2$ ) = 9.06,  $p = 0.011$ , Affect - nervous ( $df = 2$ ) = 20.19,  $p = 0.000$ , and Affect - aggressive ( $df = 2$ ) = 8.40,  $p = 0.015$ ) were found to be significant in the classification model. Table 3 shows the classification results using the survey and the experience ( $df = 56$ ) = 61.13,  $p = 0.297$ ). However, the model's fitting information shows that the full model does not significantly predict the score. In Table 4, when we built the multimodal classification model, using all modalities (the survey variables, the experience variables and two of the most significant audio variables (nervous and anger audio indicators), the model's goodness of fit shows that the model fits the data well ( $df = 52$ ) = 48.56,  $p = 0.610$ . Moreover, the model's fitting information shows that the full model significantly predicts the score, better than the intercept-only model alone ( $df = 28$ ) = 47.05,  $p = 0.014$ ). In the likelihood ratio tests, the Extrovert outgoing lead factor ( $df = 2$ ) = 11.08,  $p = 0.004$ ), the Assertive organized influential factor ( $df = 2$ ) = 13.03,  $p = 0.001$ ), the neurotic not assertive factor ( $df = 2$ ) = 8.2,  $p = 0.017$ ), the social closeness survey item ( $df = 2$ ) = 11.50,  $p = 0.003$ ), the social anger rare survey item, ( $df = 2$ ) = 7.35,  $p = 0.025$ ), tutoring experience, ( $df = 6$ ) = 14.38,  $p = 0.026$ ), and debating experience ( $df = 6$ ) = 21.92,  $p = 0.001$ ) variables were found to be significant in the classification model. The two audio variables, also found to be significant predictors: Affect nervous, ( $df = 2$ ) = 19.92,  $p = 0.000$ ), and Emotion Anger, ( $df = 2$ ) = 19.92,  $p = 0.000$ ).

## 4 Conclusions

In this paper, we showed that the extrovert leader, assertive organizer, emotionally neurotic, and charismatic personality traits as well as previous tutoring and debating experience, are all significant features of tutors who are likely to be successful at social and emotional aspects of tutoring. Support and improvement opportunities in these features should be considered as part of tutor training and CPD. Furthermore, when multimodal data was added to the classification models, the significance of the outputs has increased. Although coming from a relatively small sample size of 47 tutor candidates evaluated by three expert tutors; and, only two modalities of audio and tabular data, the results show the potential of multimodal classification models to evaluate SEL aspects of tutoring.

## References

1. Aspen Institute: From a nation at risk to a nation at hope. National Commission Final Report (2019). <https://www.aspeninstitute.org/events/national-commission-nal-report-release/>
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 423–443 (2019)
3. Blikstein, P.: Multimodal learning analytics. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 102–106. ACM (2013)
4. Cukurova, M., Luckin, R., Millan, E., Mavrikis, M.: The NISPI framework: analysing collaborative problem-solving from students' physical interactions. *Comput. Educ.* **116**, 93–109 (2018)
5. D'mello, S.K., Kory, J.: A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv. (CSUR)* **47**(3), 43 (2015)
6. Drovnikov, A.S., et al.: Teachers professional competence assessment technology in qualification improvement process. *Int. Rev. Manag. Mark.* **6**(1), 111–115 (2016)
7. Evagorou, M., Dillon, J.: Argumentation in the teaching of science. In: Corrigan, D., Dillon, J., Gunstone, R. (eds.) *The Professional Knowledge Base of Science Teaching*, pp. 189–203. Springer, Dordrecht (2011). [https://doi.org/10.1007/978-90-481-3927-9\\_11](https://doi.org/10.1007/978-90-481-3927-9_11)
8. Evans, D.E., Rothbart, M.K.: Developing a model for adult temperament. *J. Res. Pers.* **41**(4), 868–888 (2007)
9. John, O.P., Srivastava, S.: The big five trait taxonomy: history, measurement, and theoretical perspectives. *Handb. Pers. Theory Res.* **2**, 102–138 (1999)
10. Moafian, F., Ghanizadeh, A.: The relationship between Iranian EFL teachers' emotional intelligence and their self-efficacy in language institutes. *System* **37**(4), 708–718 (2009)
11. Petrides, K.V.: Psychometric properties of the trait emotional intelligence questionnaire (TEIQue). In: Parker, J., Saklofske, D., Stough, C. (eds.) *Assessing Emotional Intelligence*, pp. 85–101. Springer, Boston (2009). [https://doi.org/10.1007/978-0-387-88370-0\\_5](https://doi.org/10.1007/978-0-387-88370-0_5)
12. Siraj, I.: Nurturing 21st century skills in early childhood education and care: the way forward. In: Loble, L., Creenaline, T., Hayes, J. (eds.) *Future Frontiers Education for an AI World*, pp. 141–153. Melbourne University Press & NSW Department of Education, Australia (2017)

13. Spikol, D., Ruffaldi, E., Dabisias, G., Cukurova, M.: Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *J. Comput. Assist. Learn.* **34**(4), 366–377 (2018)
14. Tskhay, K.O., Zhu, R., Zou, C., Rule, N.O.: Charisma in everyday life: conceptualization and validation of the general charisma inventory. *J. Pers. Soc. Psychol.* **114**(1), 131 (2018)
15. Loble, L., Creenaline, T., Hayes, J.: Future frontiers education for an AI world, pp. 21–38. Melbourne University Press & NSW Department of Education, Australia (2017)



# Conscientiousness, Honesty-Humility, and Analogical/Creative Reasoning: Implications for Instructional Designs in Intelligent Tutoring Systems

Jeanine A. DeFalco<sup>1,2(✉)</sup>, Anne M. Sinatra<sup>1</sup>, Elizabeth Rodriguez<sup>3</sup>, and R. Stan Hum<sup>4</sup>

<sup>1</sup> U.S. Army CCDC Soldier Center – STTC, Natick, USA

{jeanine.a.defalco.ctr, anne.m.sinatra.civ}@mail.mil

<sup>2</sup> Oak Ridge Associated Universities, Oak Ridge, USA

<sup>3</sup> United States Military Academy, New York, USA

elizabeth.rodriguez@westpoint.edu

<sup>4</sup> Columbia University Medical Center, New York, USA

sh2117@cumc.columbia.edu

**Abstract.** This paper shares initial results of a current study to understand what factors, tools, and methods help individual military and civilian medical personnel accelerate their medical problem-solving expertise. Based on the initial data analysis, there is evidence of statistically significant positive correlations between measurements of analogical/creative reasoning with the Conscientiousness and Honesty-Humility traits as measured by the HEXACO. These results will be employed in the US Army’s Generalized Intelligent Framework for Tutoring system (GIFT) to develop a pedagogical template that adapts on relevant traits for instruction, interventions, and feedback to support accelerated learning.

**Keywords:** Intelligent · Tutoring systems · Honesty-humility · Conscientiousness · HEXACO · Analogical reasoning

## 1 Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) [1] is a domain-independent intelligent tutoring framework that the Army has been developing. The aim of the current study is to investigate how to support accelerated medical problem-solving expertise. The results of this work are intended to contribute to devising a pedagogical template for subsequent use in GIFT courses oriented towards critical care medical education. To begin this task, we have run initial correlational studies to determine what traits, as measured by the IPIP-HEXACO [4] are significantly correlated to measures of analogical/creative reasoning [2] and mental rotation tasks [5]. The results of our initial investigations are reported in this paper.

## 1.1 Investigating Correlations

In the first phase of this work, an initial correlational experiment was run at the United States Military Academy (USMA) to examine strengths of correlations between analogical/creative reasoning, mental rotation/spatial ability, and personality types. The objective was to determine what traits were statistically significantly correlated to analogical/creative reasoning so to leverage this information to help inform the experimental design of experiment two, aimed at developing a pedagogical template to support accelerated medical expertise delivered by GIFT on critical care that would adapt instruction based on the individual traits of learners.

**Analogical Finding Task Matrix (AFTM).** In the effort of supporting expertise development, Jung [6] and Hoffman [7] recommend fostering high-level reasoning skills, including creative thinking. For this study, the Analogical Finding Task Matrix (AFTM) [2] was used to measure an individual's creativity in divergent thinking, specifically analogical reasoning, a kind of reasoning that is central to creative innovation. While this is a relatively new instrument, a pilot study [2] gave evidence that confirmed that the analogies intended to be valid were identified far more frequently than analogies intended to be invalid, McNemar's within-subjects  $\chi^2 = 276.64$ ,  $P < 0.001$ . Additionally, the validity of analogies was established by domain experts, and analogy items were drawn from sets of stimuli used in previous full and pilot studies in the lab of Green [8] that obtained high rates of participant accuracy.

**Mental Rotation Tasks.** Spatial ability has been identified as relevant to high-level creative problem solving, and medical education [9]. Spatial ability and mental rotation have been linked to success in surgical skill acquisition [10, 11] and anatomical knowledge acquisition [12]. Accordingly, this first study included a mental rotation test by Ganis and Kievit, constructed and validated [5] that consists of sets of three-dimensional shapes improving on the work of Shepard and Metzler [13].

**IPIP-HEXACO.** An attribute that has been connected to analogical/creative thinking research includes personality traits [14]. In a review of the literature, Batey and Furnham [15] note that creativity in terms of the production of ideas is related to intelligence, whereas creativity as originality rests largely on personality factors. Earlier research in this area focused on Eysenck's *Gigantic Three* [16] and was followed by research on the *Big Five Factor* [17, 18]. However, a more recent instrument is finding favor in place of the Big Five: the HEXACO [4].

The value of the HEXACO rests both in the history of its development and the introduction of a six factor: *Honesty-Humility*. As such, the HEXACO model evaluates an individual's personality traits along the following criteria: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O) [4]. To reduce cognitive load and considering time constraints, the shorter validated, 60-item version of the HEXACO, the IPIP-HEXACO [19] was used. Presently, there is a gap in the literature that establishes whether there are statistically significant correlations between the AFTM instrument with traits measured in the IPIP-HEXACO. This research addresses this gap.

## 2 Methods

The first correlational study was run at USMA, in cooperation with the Department of Behavioral Sciences and Leadership and was executed online using Qualtrics. 200 participants were recruited through USMA's sign-up system, SONA. The sequence of materials for the correlational study began after participants signed an online consent form that, upon providing consent, launched a demographic survey. Following the demographic survey, participants took the AFTM [2], followed by 40 graphics from the bank of mental rotation task items [5], which was followed by the short-Grit<sup>1</sup> [20], and then the short version (60 items) of the IPIP–HEXACO [19].

## 3 Results

The descriptives of the participant pool included the following:  $N = 200$  with 3 cases missing for a total  $N = 197$ , with a mean age = 19.96, range of 18–24 years old, females  $N = 72$ , males  $N = 125$ . For the analogical reasoning tasks, the semantic distance score to measure an individual's creative reasoning strength was tallied for each matrix and is identified as AFTM Semantic matrix 1 (AFTM-S1) and AFTM Semantic matrix 2 (AFTM-S2) respectively. The sum total tally of correct responses is the measure of correct answers to the analogical reasoning tasks, and correspondingly gives us information on the strength of an individual's analogical reasoning strength. The sum total tally of correct responses is identified as AFTM Total tally matrix 1 (AFTM-TT1) and AFTM Total tally matrix 2 (AFTM-TT2). For the mental rotation tasks (MR), the total amount of correct identification of same/different images were tallied for a sum of correct responses. The IPIP–HEXACO instrument was scored according to the instructions that accompany these measures. For this paper, we are limiting our reporting to the two most positively significant correlations: Conscientious (C-IPIP) and Honesty-Humility HH-IPIP) traits.

Running a two-tailed Pearson's correlational analysis, our results provided us with the strengths of correlations between AFTM semantic distances (AFTM-S1&2) and total correct responses (AFTM-TT 1&2), total correct responses of mental rotation/spatial ability (MR), and personality types as measured by the IPIP–HEXACO —here reporting only the Honesty-Humility (HH-IPIP) and Conscientiousness (C-IPIP). We are limiting our reporting to those variables that had repeated significant correlations at the 0.01 level (2-tailed) (see Table 1).

---

<sup>1</sup> Results from GRIT will be reported in a future paper.

**Table 1.** Results Pearson correlational analysis

Correlations		AFTM-S1	AFTM-S2	AMFT-TT1	AMFT-TT2	MR	HH-IPIP	C-IPIP
AFTM-S1	Pearson correlation	1.000	.705**	.996**	.729**	.435**	.217**	.306**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.003	0.000
	N	196.000	191.000	196.000	191.000	190.000	188.000	188.000
AFTM-S2	Pearson correlation	.705**	1.000	.700**	.987**	.368**	.232**	.230**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.001	0.002
	N	191.000	191.000	191.000	191.000	190.000	188.000	188.000
AMFT-TT1	Pearson correlation	.996**	.700**	1.000	.729**	.451**	.223**	.317**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.002	0.000
	N	196.000	191.000	196.000	191.000	190.000	188.000	188.000
AMFT-TT2	Pearson correlation	.729**	.987**	.729**	1.000	.410**	.238**	.250**
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.001	0.001
	N	191.000	191.000	191.000	191.000	190.000	188.000	188.000
MR	Pearson correlation	.435**	.368**	.451**	.410**	1.000	0.118	.206**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.107	0.004
	N	190.000	190.000	190.000	190.000	190.000	188.000	188.000
HH-IPIP	Pearson correlation	.217**	.232**	.223**	.238**	0.118	1.000	.246**
	Sig. (2-tailed)	0.003	0.001	0.002	0.001	0.107		0.001
	N	188.000	188.000	188.000	188.000	188.000	188.000	188.000
C-IPIP	Pearson correlation	.306**	.230**	.317**	.250**	.206**	.246**	1.000
	Sig. (2-tailed)	0.000	0.002	0.000	0.001	0.004	0.001	
	N	188.000	188.000	188.000	188.000	188.000	188.000	188.000

\*\*Correlation is significant at the 0.01 level (2-tailed).

## 4 Conclusion

In examining the data, we find that the statistically significant positive correlation between the AMFT matrixes and the mental rotation task was not surprising, as solving analogical reasoning tasks requires the ability to call up or create mental models to visualize objects and look for common or closely related features. Similarly, comparing three-dimensional figures in the mental rotation tasks requires the ability to mentally manipulate objects to see if these objects are the same or different.

With respect to the Honesty-Humility and Conscientiousness traits, we find these to be the most compelling traits warranting further investigation, particularly as it relates to supporting creative and analogical thinking and reasoning in GIFT. Specifically,

these traits may prove to be important metrics that can be used as a point of adaption when tailoring instruction, interventions, or feedback to support creative/analogical thinking and reasoning skills to support expertise in problem solving in the medical education domain.

**Acknowledgements.** Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number **W911NF-17-2-0152**. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## References

1. Sotilare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K.: The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. Orlando, FL: U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED) (2012). [https://giftutoring.org/attachments/152/GIFTDescription\\_0.pdf](https://giftutoring.org/attachments/152/GIFTDescription_0.pdf)
2. Weinberger, A.B., Iyer, H., Green, A.E.: Conscious augmentation of creative state enhances “real” creativity in open-ended analogical reasoning. *PloS One* **11**(3), e0150773 (2016)
3. Weisberg, R.W.: *Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts*. Wiley, Hoboken (2006)
4. Ashton, M.C., Lee, K.: Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Pers. Soc. Psychol. Rev.* **11**(2), 150–166 (2007)
5. Ganis, G., Kievit, R.: A new set of three-dimensional shapes for investigating mental rotation processes: validation data and stimulus set. *Journal of Open Psychology Data Files in this Item Files Size Format View* 13-116-2-PB. pdf 567.5 Kb PDF View/Open (2015)
6. Jung, E.: Expertise development through accelerated learning: a multiple-case study on instructional principles, Doctoral dissertation, Indiana University (2016)
7. Hoffman, R.R., Ward, P., Feltovich, P.J., DiBello, L., Fiore, S.M., Andrews, D.H.: *Accelerated Learning: Training for High Proficiency in a Complex World*. Psychology Press, New York (2013)
8. Green, A.E., Spiegel, K.A., Giangrande, E.J., Weinberger, A.B., Gallagher, N.M., Turkeltaub, P.E.: Thinking cap plus thinking zap: tDCS of frontopolar cortex improves creative analogical reasoning and facilitates conscious augmentation of state creativity in verb generation. *Cereb. Cortex* **27**(4), 2628–2639 (2016)
9. Roach, V.A., Fraser, G., Kryklywy, J., Mitchell, D., Wilson, T.: Different perspectives: spatial ability influences where individuals look on a timed spatial test. *Anat. Sci. Educ.* **10**(3), 224–234 (2017)
10. Wanzel, K.R., Hamstra, S.J., Caminiti, M.F., Anastakis, D.J., Grober, E.D., Reznick, R.K.: Visual-spatial ability correlates with efficiency of hand motion and successful surgical performance. *Surgery* **134**, 750–757 (2003)
11. Brandt, M.G., Davies, E.T.: Visual-spatial ability, learning modality and surgical knot tying. *Can. J. Surg.* **49**, 412–416 (2006)
12. Lufler, R.S., Zumwalt, A.C., Romney, C.A., Hoagland, T.M.: Effect of visual– spatial ability on medical students’ performance in a gross anatomy course. *Anat. Sci. Educ.* **5**, 3–9 (2012)

13. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. *Science* **171**(3972), 701–703 (1971)
14. Lee, K., Ashton, M.C.: Prediction of self-and observer report scores on HEXACO-60 and NEO-FFI scales. *J. Res. Pers.* **47**(5), 668–675 (2013)
15. Batey, M., Furnham, A.: Creativity, intelligence, and personality: a critical review of the scattered literature. *Genet. Soc. Gen. Psychol. Monogr.* **132**(4), 355–429 (2006)
16. Eysenck, S.B., Eysenck, H.J., Barrett, P.: A revised version of the psychotism scale. *Pers. Individ. Differ.* **6**(1), 21–29 (1985)
17. Martindale, C., Dailey, A.: Creativity, primary process cognition and personality. *Pers. Individ. Differ.* **20**(4), 409–414 (1996)
18. Chamorro-Premuzic, T., Furnham, A.: Personality, intelligence and approaches to learning as predictors of academic performance. *Pers. Individ. Differ.* **44**(7), 1596–1603 (2008)
19. Ashton, M.C., Lee, K.: The HEXACO-60: a short measure of the major dimensions of personality. *J. Pers. Assess.* **91**(4), 340–345 (2009)
20. Duckworth, A.L., Quinn, P.D.: Development and validation of the Short Grit Scale (GRIT-S). *J. Pers. Assess.* **91**(2), 166–174 (2009)



# Learners' Gaze Behaviors and Metacognitive Judgments with an Agent-Based Multimedia Environment

Daryn A. Dever<sup>(✉)</sup>, Megan Wiedbusch, and Roger Azevedo

University of Central Florida, Orlando, FL 32816, USA  
{ddever, meganwiedbusch}@knights.ucf.edu,  
roger.azevedo@ucf.edu

**Abstract.** 65 undergraduate students from a North American university interacted with MetaTutorIVH, an agent-based multimedia learning environment that fosters self-regulated learning (SRL) strategy use (e.g., metacognition) while presenting information on several human body systems. Participants completed a self-paced task to study the influence of relevant content and an agents' expressed emotions on metacognitive judgments, gaze behaviors, and science learning. The goal of this study was to examine eye gaze behavior and metacognitive process use relative to their perceived relevancy of content based on discrepancies between multimedia materials and artificial agent's facial expressions. Results indicate an increase in overall fixation duration on multimedia content (e.g., text, diagram) when the text was perceived as less relevant, specifically revealing longer time spent fixating on text when the text was perceived as less relevant compared text and diagrams that were perceived as relevant. Further analyses reveal an increase in diagram fixation duration when the text was judged as being less relevant, but when compared to instances where the text and diagrams were both seen as only somewhat relevant. Across trials, there was no indication of the actual relevancy of content influencing the gaze behaviors of participants.

**Keywords:** STEM · Multimedia learning environment · Eye-tracking

## 1 Current Study

The goal of this study was to examine content evaluations (CEs) and content fixation of learners within a multimedia learning environment. This study asks the following research questions: (1) Are there significant differences in total content fixation durations based on content participants perceive as relevant versus what content is actually relevant?; and (2) Are there significant differences in total text content and diagram fixation durations based on how relevant participants perceive as relevant versus what content is actually relevant? We hypothesize there is a difference between participants' perception of relevancy and actual relevancy reflected in fixation durations of multimedia content, text content, and diagram content.

## 2 MetaTutorIVH Environment

MetaTutorIVH is an 18-trial linearly structured self-paced multimedia learning environment, about human body systems consisting of multimedia content (i.e., diagrams and text) appearing in conjunction with a pedagogical agent, emoting facial expressions based on the relevancy within each trial. Students are asked a question at the top of the screen and a CE. After submitting their own CE, the agent conveys a CE by expressing emotions (i.e., joy, confusion, neutral) designed to either facilitate or impede accurate metacognitive monitoring.

## 3 Methods

65 undergraduate students, ages ranged from 18 to 30 ( $M = 20.56$ ,  $SD = 2.60$ ), recruited from a public North American university participated in this study and compensated \$10/h. Due to incomplete data, six participants were removed from our dataset.

### 3.1 Coding and Scoring

Fixations were an 80 ms focus on an area of interest that did not exceed 100 pixels [1]. Participant CE accuracy was calculated based on CE responses compared to the actual relevancy [2]. Each trial consisted of two CEs, where participants responded on a 3-point rating scale (text or diagram is relevant, somewhat relevant, or not relevant). An accurate CE was scored as 1, a partially correct CE as 0.5, and an incorrect CE as 0. Participants scored on average 70.93% accurately overall ( $M = 1.42$ ,  $SD = 0.54$ ). When coding CEs, we grouped trials based on participant responses: (1) text as being less relevant than the diagram, (2) text as more relevant than the diagram, (3) both the text and diagram to be fully relevant, (4) all content as not relevant, and (5) all content to be somewhat relevant.

## 4 Results

### 4.1. RQ 1: Are there significant differences in total content fixation durations based on content participants perceive as relevant versus what content is actually relevant?

We conducted a non-parametric Kruskal-Wallis H test for differences in total content fixation durations between the five trial groups and found a statistically significant difference in content fixation duration,  $\chi^2(4) = 13.892$ ,  $p = 0.008$  (see Table 1). A Mann-Whitney Post Hoc analysis with a Bonferroni correction ( $p = 0.005$ ) revealed differences in fixations for trials where participants rated the text as less relevant than the diagram (Trial Group 1) and when participants rated both text and diagrams to be fully relevant (Trial Group 3;  $p = 0.002$ ). Content fixation durations were longer for Trial Group 1 compared to Trial Group 3. We also conducted a nonparametric Kruskal-

Wallis H test to examine differences in total content fixation durations between group based on actual relevancy. There were no significant differences between the groups. Results suggest that participants' total content fixation durations were different based on participant CEs but not on the actual relevancies of the content.

**Table 1.** Kruskal-Wallis of total content fixation durations between groups.

Trial group	Perceived relevancy		Actual relevancy	
	N	Mean rank	N	Mean rank
(1) Text Less Relevant	70	559.09	348	533.38
(2) Text More Relevant	312	557.62	348	518.53
(3) Both Relevant	475	490.43	348	525.59
(4) Both Somewhat Relevant	61	476.07	—	—
(5) Neither Relevant	126	558.61	—	—

- 4.2.** RQ 2: Are there significant differences in total text content and diagram fixation durations based on how relevant participants perceive as relevant versus what content is actually relevant?

A Kruskal-Wallis H test was conducted for examining the differences in total text fixation durations based on CEs. There was a statistically significant difference in text fixation durations among the five trial groups,  $\chi^2(4) = 13.169$ ,  $p = 0.010$  (See Table 2). A Mann-Whitney Post Hoc analysis with a Bonferroni correction ( $p = 0.005$ ) revealed a significant difference among the fixations on trials participants rated the text as less relevant than the diagram (Trial Group 1) and when participants rated both text and diagrams to be fully relevant (Trial Group 3;  $p = 0.001$ ). This suggests that participants fixated on text more during Trial Group 1 than Trial Group 3. We conducted a non-parametric Kruskal-Wallis H test to test for differences in total text fixation durations based on actual content relevancy. There was no significant difference among the groups. Results suggest that participants' total text fixation durations were different for perceived content relevancy but not for the actual content relevancy.

A Kruskal-Wallis H test was run for differences in total diagram fixation durations based on CEs and found a statistically significant difference of diagram fixation durations among the five trial groups,  $\chi^2(4) = 10.529$ ,  $p = 0.032$  (See Table 2). A Mann-Whitney Post Hoc analysis with a Bonferroni correction ( $p = 0.005$ ) was not significant different, potentially due to the conservative correction for multiple comparisons. A Kruskal-Wallis H test was run to test for differences in total text fixation durations based on the actual content relevancy. There was no significant difference among the groups. Results suggest that participants' total diagram fixation durations differed based on CEs but not on actual content relevancy.

**Table 2.** Kruskal-Wallis of total text fixation durations between groups.

	Text content				Diagram content			
	Perceived relevancy		Actual relevancy		Perceived relevancy		Actual relevancy	
Trial group	N	Mean rank	N	Mean rank	N	Mean rank	N	Mean rank
(1) Text Less Relevant	70	551.84	348	528.45	70	580.14	348	544.41
(2) Text More Relevant	312	558.71	348	521.99	312	528.50	348	506.91
(3) Both Relevant	475	488.06	348	517.06	475	515.17	348	516.18
(4) Both Somewhat Relevant	61	502.72	—	—	61	424.38	—	—
(5) Neither Relevant	126	555.94	—	—	126	550.75	—	—

## 5 Future Directions

Results indicate a need to monitor learners' perceptions and metacognitive judgments of content relevancy within multimedia learning environments. Results do not indicate a change in fixation duration for the actual relevancy of multimedia content type independently or in conjunction, highlighting the importance of metacognitive judgments and perceptions within these environments. From this, intelligent multimedia environments should be able to identify the patterns of learners' CEs in comparison with the correct relevancy of information to the content knowledge question. The environment would be able to intelligently identify the accuracy of learners' application of metacognitive judgments and the ability to employ correct SRL strategies. This could lead to a more intelligent scaffolding of learners' metacognitive skills through the use of artificial agents and individual and adaptive feedback within the environment to support complex learning in STEM.

**Acknowledgements.** This research was supported by funding from the National Science Foundation (DRL#1431552; DRL#1660878, DRL#1661202, CMMI#1854175, and DRL#1916417) and the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). The authors would also like to thank members of the SMART Lab for their assistance and contributions.

## References

1. Salvucci, D., Goldber, J.: Identifying fixation and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research and Applications, pp. 71–78. ACM (2000)
2. Mudrick, N., Taub, M., Azevedo, R.: Do accurate metacognitive judgments predict successful multimedia learning? In: Proceedings of the 39th Annual Conference of the Cognitive Science Society, Austin, Texas, pp. 2766–2771 (2017)



# Online Assessment of Belief Biases and Their Impact on the Acceptance of Fallacious Reasoning

Nicholas Diana<sup>(✉)</sup>, John Stamper, and Kenneth Koedinger

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA  
[ndiana@cmu.edu](mailto:ndiana@cmu.edu), [john@stamper.org](mailto:john@stamper.org), [koedinger@cmu.edu](mailto:koedinger@cmu.edu)

**Abstract.** Determining the impact of belief bias on everyday reasoning is critical for understanding how our beliefs can influence how we judge arguments. We examined the impact of belief bias on the user’s ability to identify logical fallacies in political arguments. We found that participants had more difficulty identifying logical fallacies in arguments that aligned with their own political beliefs. Interestingly, this effect diminishes with practice. These results suggest that while belief bias is a potential barrier to correctly evaluating everyday arguments, interventions focused on activating rational engagement may mitigate its impact.

**Keywords:** Belief bias · Informal reasoning · Informal logic · Logical fallacies

## 1 Introduction

For decades, research in formal logic has demonstrated that our prior-knowledge or beliefs can interfere with our ability to reason logically about an argument [2, 4–8]. This phenomenon, known as *Belief Bias*, is precisely defined as a tendency to “base [our] judgments on the believability of the conclusions” rather than the “logical form of the arguments” [8].

We used LIFTS (Logic and Informal Fallacy Tutoring System) [1] to test the impact of belief bias on one’s ability to identify informal logical fallacies. Inside the tutor, we asked participants to identify informal fallacies in short arguments, given some context. Some of these arguments were political in nature, with conclusions designed to support either a typical conservative or typical liberal position on an issue. We hypothesized that participants would have more difficulty identifying fallacies in problems with conclusions that align with their own political beliefs (i.e., conclusions they may agree with), despite the fact that all of the arguments presented were fallacious.

## 2 Methods

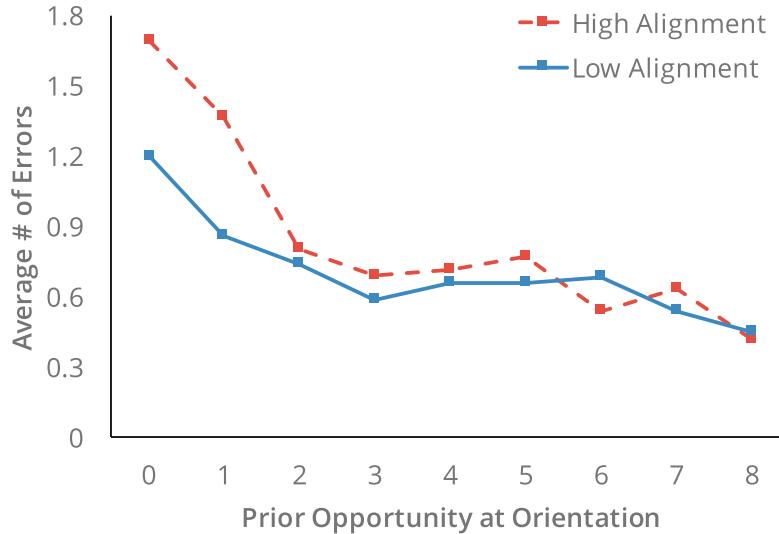
Sixty-three participants were recruited for the experiment. In order to select a politically diverse sample, subjects were recruited using Amazon Mechanical

Turk with the restriction that they must reside in the United States. To mitigate concerns about data quality, we analyzed the log data to find participants who appeared to be “gaming” the tutor. Participants who provided an answer less than a second after seeing a problem were classified as “gamers” and excluded. Six participants were excluded for gaming, two for possessing clear outlier values (values above the 99.9th percentile) on the outcome variable (number of errors), and one for having a large time gap (more than an hour) between actions. Of the remaining 54 participants, 23 identified as female, 30 as male, and 1 as agender. The average age of participants was 31.31 years old ( $SD = 6.67$ ).

In this experiment, participants were asked to identify the fallacy present in an argument from a list of 3 different fallacies. The tutor consisted of 18 problems total, with each of the 3 fallacies being the correct answer 6 times. Problems were presented in a random order. Of the 18 problems, 9 were designed to contain a conservative conclusion and 9 were designed to contain a liberal conclusion. All of the arguments presented contained an informal logical fallacy, but we expected that participants would have more difficulty identifying the fallacy when their personal political orientation matched (or aligned with) the political orientation of the problem. Instruction was provided in expandable drop-down boxes that contained definitions and examples of the fallacies. After completing the tutor, participants were asked to complete a post-test questionnaire that included three questions that directly assessed beliefs about the specific political issues used in the study, and general demographics questions.

### 3 Results

We built a mixed linear regression model with *number of errors* as the outcome variable and *participant* as a random effect. Input into the model was a participant-by-problem table, such that each row represented one participant’s performance on one problem. The mixed linear regression included the following as features: **Prior Opportunities at Fallacy** (**oppFallacy**) represents the number of times the participant has seen the fallacy in a question before. If learning occurs over the course of the experiment, we expect this feature’s coefficient to be negative (i.e., inversely related to *number of errors*). **Prior Opportunities at Orientation** (**oppOrientation**) represents the number of times the participant has seen a problem with this political orientation (i.e., with a conservative or liberal conclusion) before. We do not expect this feature to be a significant predictor outside of an interaction. In other words, by itself, orientation should not add any difficulty to the problem. **Alignment** represents the degree to which the participant’s political beliefs (as measured using the direct questions discussed above) align with the political orientation of the problem. We expect this feature to be a significant, positive predictor of *number of errors* outside of an interaction. **oppOrientation \* Alignment** represents the interaction between the number of prior opportunities at an orientation and the degree to which the participant’s political beliefs align with that orientation. It may be the case that belief bias has a strong effect on performance at the beginning



**Fig. 1.** In general, LIFTS succeeded at teaching fallacy identification (i.e., users made fewer errors with practice). We also see that users were biased by their beliefs, as evidenced by the differences between early performance on high alignment and low alignment problems. Alignment group was determined by score (1–5) on the self-reported alignment questions (High = 4 or 5, Low = 1 or 2). These results suggest that belief bias impacted performance early in the experiment, but that the effect diminished with practice. At least in this case, LIFTS was successful in reducing belief bias.

of the experiment, but as participants continue to practice identifying fallacies in arguments that align with their beliefs, the impact of belief bias diminishes. This term was designed to capture that interaction.

As predicted, the number of *Prior Opportunities at Fallacy* was a significant predictor ( $\beta = -0.102, p < .05$ ) and inversely related to *number of errors*. As participants had more opportunities practicing a fallacy, their performance improved. In other words, learning occurred inside LIFTS.

We also found that the interaction between the number of *Prior Opportunities at Orientation* and the participant's *Alignment* to that orientation was a significant, negative predictor ( $\beta = -0.021, p < .05$ ). If we plot this interaction (see Fig. 1), it appears that belief bias impacts performance early in the experiment, but that the effect diminishes with practice. This interpretation is supported by the coefficient for *Alignment* by itself ( $\beta = 0.124, p < .05$ ). We see that higher *Alignment* is associated with worse performance (i.e., higher *number of errors*) when *oppOrientation* is 0. Outside of the interaction, *oppOrientation* was not a significant predictor.

## 4 Discussion

There are two possible interpretations for the diminishing impact of belief bias with practice. First, it is possible that an improved understanding of the logical fallacies makes the fallacious features of an argument more salient. If this is the

case, then reducing belief bias is a matter of better training in argument evaluation. However, it is also possible that it's not learning that is reducing belief bias, but rather that some typically dormant critical thinking faculties are coming online (as the task requires them) and overpowering the influence of belief bias. This interpretation seems to support the main assertion of Haidt's Social Intuitionist Model of moral reasoning [3], which argues that everyday moral reasoning happens quickly and is primarily based on intuitions (as opposed to a rational assessment of the argument). Rationalization enters into the model *after* a moral decision has been reached, to justify the decision (or conversely, to undermine an opposing position). With respect to the current experiment, it is possible that the belief bias effect seen early in the experiment is evidence of an intuitions-based moral reasoning, and performance improves as participants discover that the task requires rational reasoning. If this interpretation is correct, then performance on the earlier problems is representative of how we typically evaluate everyday arguments (i.e., in the absence of heightened critical thinking). Moreover, the difference observed between the *High Alignment* and *Low Alignment* groups on these early problems suggests that being susceptible to belief bias may be the typical case.

If this second hypothesis is true, then mitigating belief bias in everyday reasoning may not simply be a matter of better training in argument evaluation. Instead, systems designed to combat our susceptibility to weak arguments or misleading news stories should place a greater emphasis on understanding the user's beliefs and how those beliefs (1) relate to the beliefs present in the content they are consuming, and (2) impact their judgment of that content's validity.

## 5 Conclusion

We demonstrated that a participant's political beliefs impacted their ability to identify logical fallacies in arguments that aligned with those political beliefs. The larger impact of belief bias on earlier problems may be evidence of an intuitionist model of moral reasoning. As such, while our results suggest that the key to overcoming belief bias may be to simply think more critically about the argument in question, they also imply that we naturally forgo this critical evaluation when we agree with the argument. Combating the negative effects of belief bias in real-world contexts such as advertising and politics may benefit from some external agent that can relate the user's values to the values latent in the text, and prime us to think critically about invalid arguments that may intuitively seem true.

## References

1. Diana, N., Stamper, J., Koedinger, K.: An instructional factors analysis of an online logical fallacy tutoring system. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 86–97. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_7](https://doi.org/10.1007/978-3-319-93843-1_7)

2. Evans, J.S., Barston, J.L., Pollard, P.: On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* **11**(3), 295–306 (1983). <https://doi.org/10.3758/BF03196976>
3. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* **108**(4), 814 (2001)
4. Henle, M., Michael, M.: The influence of attitudes on syllogistic reasoning. *J. Soc. Psychol.* **44**(1), 115–127 (1956). <https://doi.org/10.1080/00224545.1956.9921907>
5. Lefford, A.: The influence of emotonal subject matter on logical reasoning. *J. Gen. Psychol.* **34**, 127–151 (1946). <https://doi.org/10.1080/00221309.1946.10544530>
6. Markovits, H., Nantel, G.: The belief-bias effect in the production and evaluation of logical conclusions. *Mem. Cogn.* **17**(1), 11–17 (1989). <https://doi.org/10.3758/BF03199552>
7. Morgan, J.J.B., Morton, J.T.: The distortion of syllogistic reasoning produced by personal convictions. *J. Soc. Psychol.* **20**(1), 39–59 (1944). <https://doi.org/10.1080/00224545.1944.9918830>
8. Revlin, R., Leirer, V., Yopp, H., Yopp, R.: The belief-bias effect in formal reasoning: the influence of knowledge on logic. *Mem. Cogn.* **8**(6), 584–592 (1980). <https://doi.org/10.3758/BF03213778>



# Early Dropout Prediction for Programming Courses Supported by Online Judges

Filipe D. Pereira<sup>1</sup>(✉), Elaine Oliveira<sup>2</sup>, Alexandra Cristea<sup>3</sup>, David Fernandes<sup>2</sup>, Luciano Silva<sup>1</sup>, Gene Aguiar<sup>1</sup>, Ahmed Alamri<sup>3</sup>, and Mohammad Alshehri<sup>3</sup>

<sup>1</sup> Department of Computer Science, Federal University of Roraima,  
Boa Vista, Roraima, Brazil

{filipe.dwan,luciano.silva,gene.charles}@ufrr.br

<sup>2</sup> Computer Institute, Federal University of Amazon, Manaus, Amazon, Brazil  
{elaine,david}@icomp.ufam.edu.br

<sup>3</sup> Durham University, Durham, UK  
{alexandra.i.cristea,ahmed.s.alamri,mohammad.a.alshehri}@durham.ac.uk

**Abstract.** Many educational institutions have been using online judges in programming classes, amongst others, to provide faster feedback for students and to reduce the teacher’s workload. There is some evidence that online judges also help in reducing dropout. Nevertheless, there is still a high level of dropout noticeable in introductory programming classes. In this sense, the objective of this work is to develop and validate a method for predicting student dropout using data from the first two weeks of study, to allow for early intervention. Instead of the classical questionnaire-based method, we opted for a non-subjective, data-driven approach. However, such approaches are known to suffer from a potential overload of factors, which may not all be relevant to the prediction task. As a result, we reached a very promising 80% of accuracy, and performed explicit extraction of the main factors leading to student dropout.

**Keywords:** Learning analytics · Programming online judges · Dropout

## 1 Introduction

As dropout prediction based on data-driven solutions has been studied recently intensively in MOOCs [2, 5, 14], at first glance, similar approaches seem applicable using data from Programming Online Judges (POJ) [6, 13]. Especially early prediction has been advocated [4, 5, 10], as it is the only type of prediction that allows for interventions, for students as well as in supporting teachers. However, POJ are more challenging than ‘simple’ e-learning systems, including MOOCs, which mainly deliver content, or even online tests and evaluations, usually only based on multiple choice tests or questionnaires. The main complexity lies in the ‘free’ nature of the student input, in the form of a program. Hence, the data created is both richer and more complex. The complexity increases when

online judges are complemented by IDEs, which allow students to input multiple programs and receive iterative feedback. Moreover, MOOCs usually have very high numbers of students (of the orders of thousands or tens of thousands) [5, 12], whereas online judges with embedded IDE, as here, have lower numbers [8]. This leads to the data being potentially less reliable, and the prediction more difficult. Thus, here, we tackle, to the best of our knowledge, for the first time, the challenging problem of early prediction of dropout using data collected from Introductory Programming courses supported by IDEs embedded in POJs. To do so, we defined two research questions: **RQ1**. How can early dropout prediction for students on Introductory Programming courses be achieved, for medium-sized cohorts using IDE embedded in online judges? **RQ2**. Which early student behaviours (here, features) are leading indicators of dropout (for the case above)? (answering to questions such as: *why?* and *how?*)

## 2 Methodology

In our work dropout is interpreted as having an attendance level less than 75% in the course of Introduction to Programming, since we collected data from the Federal University of Amazonas. In Brazil there is a law which establishes that for every University course, students can not pass if their absence is higher than 25%. We collected data from the online judge CodeBench system, which was developed by one of the authors, used as support for instructors and students in programming courses. The data were collected from 9 introductory programming classes. In this paper, only the data from the first two weeks were used as training data for the prediction task, as we aim at as early prediction as possible.

To construct the predictive model we first collected and defined 20 initial ML features, starting from the state of the art, from related domains, which could be applied to Programming Classes with online judges. We also added our own self-devised features, which were introduced based on knowledge extracted from discussions with teachers that were using IDEs with online judges. For instance, we used number of comments; number of logical lines, time spent programming (in minutes), and etc. However, after performing Recursive Feature Selection [7], only 5 of the 20 features (which will be discussed more in depth at the end of this section) were relevant for the task of predicting dropout, which are: **lloc** - number of logical lines for each submitted code [9]; **correctness** - number of test cases passed for each problem [3]; **correctness\_with\_effort**: represents the same as *correctness*, but in this case we considered correct only student solutions with more than 50 log lines<sup>1</sup>; **access\_num** - number of student logins between the beginning and end of a session; **keystroke\_latency** - keystroke latency of the students (in seconds) when typing in the embedded IDE;

Furthermore, because of the unbalanced nature of the dataset, where approximately 79% of the students did not dropout, we applied random undersampling.

---

<sup>1</sup> Number of log lines on attempt to solve problems. To illustrate, each time the student presses a button in the embedded IDE of the ‘online judge’, this event is stored as a line in a log file (adapted from [1, 3]).

For prediction, we employed the ML algorithm C4.5 [11] because besides being efficient, it provides an easy interpretation of the existing relationships in the data. The model was optimized using grid search and validated with 10-folds cross-validation method.

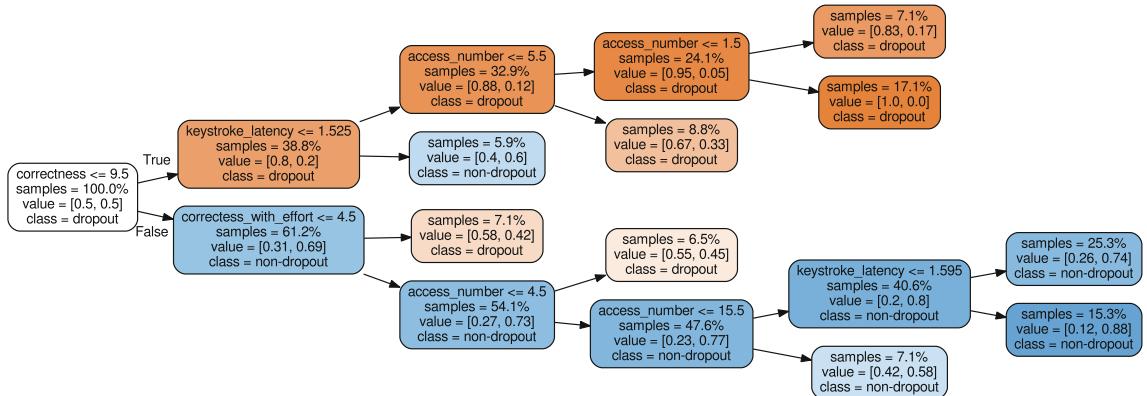
### 3 Results and Discussion

Using the method explained in the previous section, the predictive model achieved 80% of accuracy. The model was able to identify students who dropped out and those who did not, with a similar hit rate, as shown in Table 1.

**Table 1.** Results using a balanced database.

Precision	Recall	Class	Situation
0.82	0.76	0	Dropout
0.78	0.84	1	Complete

As our goal is to analyse which early student behaviours (here, features) are leading indicators of dropout or completion, we retrained the same model that achieved 80% of accuracy with the entire dataset. The resulting tree can be seen in Fig. 1, where the nodes in orange represent the dropout estimation of the tree and the blue nodes represent completion. The difference in color tones is due to the division of the parent node. In other words, the darker the color, the higher the information gain in the prediction (less entropy).



**Fig. 1.** Decision tree created using the entire balanced database. (Color figure online)

The decision tree nodes of Fig. 1 might contain four fields. At the top of each internal node, there is a *condition* that is used for the estimator to make a decision, which may be true or false, where the upward-tilted arrow indicates

true and the down-tilt arrow indicates false. In all nodes there is a field *samples* that shows the percentage of the sub-sample that was received by a child node after the split, based on the condition of the parent node. Just below the *samples* field, we have the field *value* =  $[x, y]$ , represented by an ordered pair, where  $x$  contains the percentage of the *samples* dropout and  $y$  brings the value of the non-dropout sub-sample. Finally, there is a field called *class* that represents the *decision* made based on the previous conditions, that is, whether or not the student has dropped out.

It is noticed that the **correctness** feature was the most relevant for the model and therefore it was placed in the root of the tree. Thus, if **correctness** is less than or equal to 9.5 and the feature *keystroke\_latency* is greater than 1.525, then the student is classified as non-dropout. An analysis of this rule allows us to understand that students who do well on the problem lists and code quickly are more likely not to drop out. Another aspect of this is the fact that students coding fast may indicate that they already had previous programming experience.

However, when the student has a **correctness** grade below 9.5, and a **keystroke\_latency** less than or equal to 1.525, as well as a very low number of accesses to the online judge (*access\_num*), the probability of him to dropout is very high. This can be seen in the leaves at the top of the tree (orange), where the confidence level of the decision is 83% and 100% (almost without entropy).

On the other hand, if the **correcteness** is greater than 9.5, the student has often accessed the online judge and **correctness\_with\_effort** is greater than 4.5, then the student is classified as non-dropout. Noteworthy is that this rule shows that dedicated students, who solve the problems list, frequently access the online judge and have many lines of log in solving the problems (**correctness\_with\_effort**) in the first two weeks of the course, they usually complete the course. However, even when a student has solved many problems, if they generated only a few **log\_lines**; if such a student has additionally accessed the online judge only a few times, then this student is classified as dropout. Observe that if the students have few **log\_lines** on a particular submitted solution, this may mean that they did not solve the problem from scratch in the IDE.

## 4 Conclusion

In our view, these rules are very interesting as they could be presented as warnings to the students, perhaps as pop-up messages when they are programming in the IDE of the online judge. It might be helpful for students to know in advance that some programming behaviours might lead to dropping out. For example, knowing in advance that it is important to solve all the programming problems from the lists of problems, but is also important to undertake effort doing it by themselves, without many “copy and paste” actions, could lead the student to a more conscientious attitude, as well as being empowered and in charge of their learning. Another point is that students who have lower keystroke latency

(code slowly) could be hesitating or procrastinating and some recommendation about this issue could be important to make the students reflect about their behaviour. However, in general, the interventions could help students to improve upon identified weaknesses in their programming skills, by recommending them, for example, to revisit specific parts of the material, post their doubts on the forums, and talk to the teacher/tutor. From the perspective of the instructors, some information could be displayed to them, such as a list (group) of students who have a high probability to dropout or not. With this information in hand, the instructors could do some interventions.

## References

1. Ahadi, A., Lister, R., Haapala, H., Vihavainen, A.: Exploring machine learning methods to automatically identify students in need of assistance. In: ICER 2015, pp. 121–130 (2015)
2. Alamri, A., et al.: Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In: Coy, A., Hayashi, Y., Chang, M. (eds.) ITS 2019. LNCS, vol. 11528, pp. 163–173. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-22244-4\\_20](https://doi.org/10.1007/978-3-030-22244-4_20)
3. Castro-Wunsch, K., Ahadi, A., Petersen, A.: Evaluating neural networks as a method for identifying students in need of assistance. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, pp. 111–116. ACM (2017)
4. Chen, W., Brinton, C.G., Cao, D., Mason-Singh, A., Lu, C., Chiang, M.: Early-detection prediction of learning outcomes in online short-courses via learning behaviors. IEEE Trans. Learn. Technol. (2018)
5. Cristea, A.I., Alamri, A., Kayama, M., Stewart, C., Alsheri, M., Shi, L.: Earliest predictor of dropout in moocs: a longitudinal study of futurelearn courses. In: 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden. Association for Information Systems (2018)
6. Dwan, F., Oliveira, E., Fernandes, D.: Predição de zona de aprendizagem de alunos de introdução à programação em ambientes de correção automática de código. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol. 28, p. 1507 (2017)
7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. J. **46**(2), 389–422 (2002)
8. Ihantola, P., et al.: Educational data mining and learning analytics in programming: literature review and case studies. In: Proceedings of the 2015 ITiCSE on Working Group Reports, pp. 41–63. ACM (2015)
9. Otero, J., Junco, L., Suarez, R., Palacios, A., Couso, I., Sanchez, L.: Finding informative code metrics under uncertainty for predicting the pass rate of online courses. Inf. Sci. **373**, 42–56 (2016)
10. Pereira, F.D., Oliveira, E., Fernandes, D., Cristea, A.: Early performance prediction for CS1 course students using a combination of machine learning and an evolutionary algorithm. In: The 19th IEEE International Conference on Advanced Learning Technologies (ICALT 2019) (2019)
11. Quinlan, J.R.: C4. 5: Programming for Machine Learning, vol. 38, p. 48. Morgan Kauffmann (1993)

12. Vivian, R., Falkner, K., Falkner, N.: Addressing the challenges of a new digital technologies curriculum: MOOCs as a scalable solution for teacher professional development (2014)
13. Wasik, S., Antczak, M., Laskowski, A., Sternal, T., et al.: A survey on online judge systems and their applications. ACM Comput. Surv. (CSUR) **51**(1), 3 (2018)
14. Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., Tingley, D.: Delving deeper into MOOC student dropout prediction. arXiv preprint [arXiv:1702.06404](https://arxiv.org/abs/1702.06404) (2017)



# Developing a Deep Learning-Based Affect Recognition System for Young Children

Amir Hossein Farzaneh<sup>1</sup>(✉), Yanghee Kim<sup>2</sup>, Mengxi Zhou<sup>2</sup>, and Xiaojun Qi<sup>1</sup>

<sup>1</sup> Department of Computer Science, Utah State University, Logan, UT 84322, USA  
farzaneh@aggiemail.usu.edu, xiaojun.qi@usu.edu

<sup>2</sup> Department of Educational Technology, Research, and Assessment,  
Northern Illinois University, DeKalb, IL 60115, USA  
ykim9@niu.edu, z1841378@students.niu.edu

**Abstract.** Affective interaction in tutoring environments has been of great interest among several researchers in this community, which has spurred the development of various systems to capture learners' emotional states. Young children are one of the biggest learner groups in digital learning environments, but these studies have rarely targeted them. Our current study leverages computer vision and deep learning to analyze young children's learning-related affective states. We developed an effective recognition system to compute the probability for a child to present neutral or positive affective state. Our results showed that the prototype was able to achieve an average affective state prediction accuracy of 93.05%.

**Keywords:** Emotion recognition · Deep learning · Computer vision · Young children · Learner affect

## 1 Introduction

Advances in Artificial Intelligence (AI) over recent decades have led to a growing interest in the development of AI-based approaches to education, as well as broadening the use of AI applications in education. The AIED community acknowledges the important role of affect for learning and has examined the relationship between affective states and learning gains in various domains and with various groups of learners [2, 5, 13]. Representative on-going efforts of the community include the quantitative method (called BROMP) to observe student behaviors and affective states [12], the virtual character to collect students' reported emotional states [15], and human annotators to detect learner emotions [14]. Likewise, AI-enhanced emotion recognition has also been proliferating, helping to interpret emotional states of users for both educational and therapeutic purposes. In particular, children's interactions with digital devices are rich in emotions and involve a lot of non-verbal responses [3, 18]. Designing and implementing a system that recognizes children's emotions using a non-verbal channel

such as facial expression, is needed to substantiate and expedite the analysis of children’s behaviors in a digital learning environment.

A majority of conventional recognition systems are built on smaller datasets and rely on compact hand-crafted features [7, 10, 16]. As a result, they fail to incorporate the variability in facial expressions among different demographics. With the emergence of larger Facial Expression Recognition (FER) databases, modern deep learning techniques [4, 8, 9, 20] have increasingly been implemented to operate directly on image pixels to automatically extract complex features from facial images to represent emotions at different layers and handle challenging factors for emotion recognition in the wild. However, these emotion recognition engines have been built from adult face databases that represent the fine-tuned dynamics of mature faces. The performance of such predictive models for children is therefore sub-optimal.

In this study, we have leveraged deep learning techniques to predict the emotions of young children.

## 2 Deep Learning-Based Facial Emotion Recognition

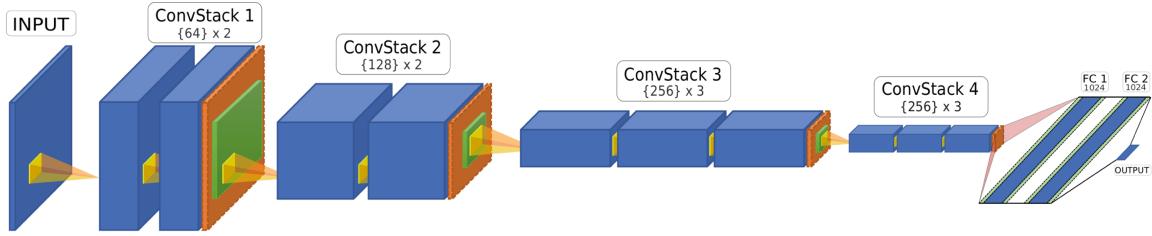
In the current on-going study, we aim to develop a Deep Convolutional Neural Network (DCNN) based emotion recognition prototype that automates effective extraction of sophisticated facial features and thereby more accurate classification of affective states. We continuously test this prototype with kindergarten-aged children as they interact in a natural classroom environment.

To train DCNN, we use an enhanced FER (FER+) dataset [1] that contains 35,887 face images annotated with eight emotions: *neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt*. FER+ images have been captured under diverse illuminations, head poses, and occlusions and have a broad spectrum of demographics including people of different ages and races. On the other hand, other FER datasets containing children’s faces such as National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS) [6] offer few labeled images under lab-controlled conditions with limited head poses. Since there are various and unexpected bodily movements of children in a natural classroom setting, we decided to use FER+ to train DCNN.

For the testing dataset, we use three video sequences which include children interacting with a teaching assistant. In these trial tests, we consider two affective categories: positive and neutral. For annotation, two researchers and two graduate students discussed the annotation criteria first, individually annotated ten clips of face images, and discussed the individual results until they reached consensus. They repeated this process four times for each child.

### 2.1 Training

We use a VGG-like standard deep architecture [1] to train an inference model on the FER+ set. This VGG model achieves close to state-of-the-art performance while offering a simple architecture [17]. Figure 1 presents the architecture of



**Fig. 1.** VGG-12 architecture: blue, yellow, orange, and green are activation, convolution, max-pooling, and drop-out layers, respectively. Each number in the bracket reflects the depth of the corresponding convolutional layer. (Color figure online)

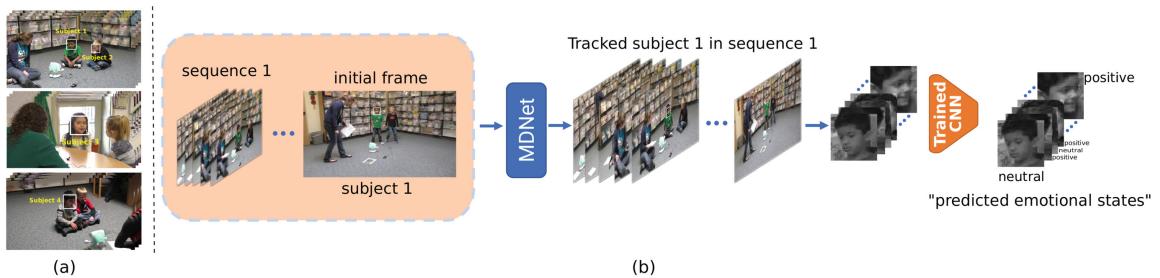
the VGG-like CNN, where 10 layers are convolutional layers in 4 stacks and 2 layers are linear classification layers. The final layer produces the probability of *positive* emotion and the probability of being *neutral* for each candidate.

Before the training process, we re-label *happiness* and *neutral* face images of the FER+ dataset as *positive* and *neutral* to match with the expected labels of children in a classroom setting. We then scale the original face images to the size  $48 \times 48$  and feed the scaled and labeled 8,733 *positive* and 7,284 *neutral* face images to the VGG-12 network.

Training is carried out by optimizing the cross-entropy loss using the back-propagation algorithm. To provide better generalization on the test data, we perform on-the-fly data augmentation during training by applying random affine transformations [19] and random horizontal flipping on input images to generate significantly more perturbed training images.

## 2.2 Testing

To test the performance of our emotion recognition system, we first track each child's face in three video sequences using a CNN-based Multi-Domain Convolutional Neural Network (MDNet) tracker [11]. We then crop the face region, apply histogram equalization to increase its contrast, and pass the processed face through the emotion recognition system. The predicted emotion for each child in all frames is saved for evaluation. Figure 2a presents the MDNet-tracked



**Fig. 2.** Illustration of the proposed system. (a) Tracking results on three sample video frames of the test dataset. (b) Proposed emotion recognition system for subject 1 in sequence 1.

face image bounding boxes in yellow in three sample frames. Figure 2b shows the proposed end-to-end emotion recognition system for subject 1 in test video sequence 1.

### 3 Evaluation

Table 1 summarizes the prediction accuracy results for four children in each test video sequence. We calculated the accuracy of the proposed emotion recognition system as the ratio of *the number of correct predictions* and *the total number of predictions* and evaluated prediction at the rate of one frame per second.

Due to the random parameter configuration for DCNNs during training, we evaluated our system in five trials and reported an average prediction accuracy for each child. Our proposed model achieved an average accuracy of 93.05% with above 90% of accuracy for three children and 83.46% for the fourth child. The lower rate for the fourth kid was mainly due to poor lighting in the room and more non-frontal faces.

**Table 1.** Emotional state testing accuracy for 4 children in 3 test video sequences

Target subject	Trials					Average accuracy
	1	2	3	4	5	
<i>Subject 1 Sequence 1</i>	89.01%	96.70%	87.91%	96.70%	94.51%	92.97%
<i>Subject 2 Sequence 1</i>	98.04%	96.08%	96.08%	100.00%	98.04%	97.65%
<i>Subject 3 Sequence 2</i>	96.96%	98.48%	98.18%	98.78%	98.18%	<b>98.12%</b>
<i>Subject 4 Sequence 3</i>	82.69%	84.62%	82.69%	82.69%	84.62%	83.46%
Average accuracy	91.67%	93.97%	91.21%	<b>94.54%</b>	93.84%	

### 4 Discussion and Future Work

In this study, we developed a system that recognizes young children's affective states (e.g., positive and neutral). The system achieves an average prediction accuracy of 93.05% in the five running trials with four children.

Some challenges at this stage are in line with previous research in the AIED community. These include the need for psychological and theoretical frameworks to more clearly define the categories of children's learning-related emotions. Some emotions in the highly recognized emotion database like FER+ are not specifically related to learning behaviors [2].

Lastly, the team acknowledges the need for the detection and analysis of dynamic affect (i.e., transition and reciprocity between affective states) beyond static affect to be able to fully understand learning behaviors in natural settings. This will be achieved effectively when complemented by other behavioral data

that include speech, voice, and bodily movements, which leads us to continuous computational exploration to coordinate multi-modal datasets and interpret multiple sources of information meaningfully.

## References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016)
2. Bosch, N., D'Mello, S.: The affective experience of novice computer programmers. *Int. J. Artif. Intell. Educ.* **27**(1), 181–206 (2017)
3. Breazeal, C.L.: Designing Sociable Robots. MIT Press, Cambridge (2004)
4. Caramihale, T., Popescu, D., Ichim, L.: Emotion classification using a tensorflow generative adversarial network implementation. *Symmetry* **10**(9), 414 (2018)
5. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. *Learn. Instr.* **22**(2), 145–157 (2012)
6. Egger, H.L., et al.: The NIMH child emotional faces picture set (NIMH-ChEFS): a new set of children's facial emotion stimuli. *Int. J. Methods Psychiatr. Res.* **20**(3), 145–156 (2011)
7. Kalsum, T., Anwar, S.M., Majid, M., Khan, B., Ali, S.M.: Emotion recognition from facial expressions using hybrid feature descriptors. *IET Image Process.* **12**(6), 1004–1012 (2018)
8. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
9. Matsugu, M., Mori, K., Mitari, Y., Kaneda, Y.: Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw.* **16**(5), 555–559 (2003)
10. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* **4**(2), 151–160 (2013)
11. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4293–4302 (2016)
12. Ocumpaugh, J., Baker, R., Rodrigo, M.: Monitoring protocol (BROMP) 2.0 technical & training manual. Teachers College, New York, NY (2015)
13. Ocumpaugh, J., Baker, R.S.J., Gaudino, S., Labrum, M.J., Dezendorf, T.: Field observations of engagement in reasoning mind. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 624–627. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_74](https://doi.org/10.1007/978-3-642-39112-5_74)
14. Okur, E., Aslan, S., Alyuz, N., Arslan Esme, A., Baker, R.S.: Role of socio-cultural differences in labeling students' affective states. In: Artificial Intelligence in Education, pp. 367–380 (2018)
15. Ranjbartabar, H., Richards, D., Makhija, A., Jacobson, M.J.: Students' responses to a humanlike approach to elicit emotion in an educational virtual world. In: Artificial Intelligence in Education, pp. 291–295 (2018)
16. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* **27**(6), 803–816 (2009)

17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
18. Turkle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Hachette, New York (2017)
19. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442 (2015)
20. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)



# Using Exploratory Data Analysis to Support Implementation and Improvement of Education Technology Product

Mingyu Feng<sup>1</sup>(✉) , Daniel Brenner<sup>1</sup>, and Andrew Coulson<sup>2</sup> 

<sup>1</sup> 400 Seaport Ct, Redwood City, CA 94063, USA

{mfeng, dbrenne}@wested.org

<sup>2</sup> 111 Academy, Suite 100, Irvine, CA 92617, USA

acoulson@mindresearch.org

**Abstract.** ST Math is a visual instructional game-based program that builds a deep conceptual understanding of mathematics through rigorous learning and creative problem solving. It is widely adopted in many elementary schools in the US. In this paper, we describe the exploratory data analysis we conducted on system log data of kindergarten students to discover patterns in students' interaction with the system, and to examine productivity and engagement of students with different profiles. The findings informed the implementation of the program in schools as well as improvement of individual games in the program to render more effective student learning.

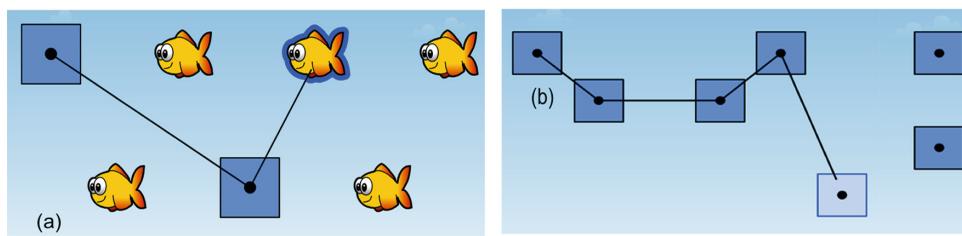
**Keywords:** ST Math · Game-based learning · Education technology · Exploratory data analysis

## 1 Introduction

Today, advances in technology have created a much richer learning environment than before. There are abundant visions for using technologies in schools to help improve mathematics achievement and close the achievement gaps (Bohrnstedt et al. 2015). Some studies have begun to find meaningful effects of newer technological interventions on student outcomes (Pape et al. 2010; Roschelle et al. 2010; Roschelle et al. 2016; VanLehn 2011). But many other studies of educational technologies found small or no effects (Bielefeldt 2005; Campuzano et al. 2009; Pane et al. 2014; Rutherford et al. 2014; What Works Clearinghouse 2016). Noticing the mixed results, different kind of efforts are being made to improve education technologies to increase the learning outcome in such environments. The work described here represents such kind of effort to improve an educational technology product for mathematics learning using data-driven approach.

Created by the nonprofit MIND Research Institute (MIND), Spatial-Temporal (ST) Math provides a distinctive a game context for individualized instruction. The program is designed to teach mathematical reasoning through intuitive spatial temporal representations where key concepts are illustrated with dynamic imagery and minimal mathematical symbols and terminologies (Rutherford et al. 2014). The program has

been adopted by over 1,400 schools in the United States. ST Math contains a large suites of interactive games (puzzles) that are formulated to engage and motivate students to solve mathematics problems. Learning in ST Math is self-paced and competency-based. Students have two “lives” in each level, and they must repeat the level after they lost both lives. Figure 1 showed two puzzles in the Object “Numbers and Objects to 10” of Kindergarten curriculum where students count and recognize quantities from 0 to 10. In puzzle (a), counting is supported by black lines that trace the counting sequence of objects that are aligned on the screen. Puzzle (b) is slightly harder as the items are not laid out as well. In addition to the scaffolding shown in Fig. 1, ST Math provides a corrective feedback and allows for the gradual extrapolation of mathematics principles within lessons to help build students self-confidence and motivation.



**Fig. 1.** Puzzles of progressing difficulty levels in ST Math

## 2 Identify Patterns of Usage and Areas for Improvement

In 2018, we conducted an exploratory data analysis study on data collected by MIND from previous use of ST Math, to examine how students of different backgrounds interact with the program, identify which levels are particularly challenging or easy for students, and explore any relationships between a student’s overall standardized math NWEA test score and his/her performance in ST Math. The analysis reported will also help identify areas of improvement in the product design and implementation in classrooms. The research questions that guided the study were: (a) Do students of different backgrounds show different patterns of learning, such as failing specific puzzles in ST Math, engagement level, time on task, or intensity of use, etc.? (b) What patterns of behavior within ST Math are associated with more productive learning (fewer failures)? (c) Which levels/puzzle in ST Math are particularly challenging or easy for students based on performance from the student perspective?

We received data sets for 254 students from 9 Kindergarten classrooms in Chicago Public Schools. The data sets include the aggregated system use data that characterized student use of ST Math in school during the 2017–18 school year. Student’s basic demographic information (gender and ethnicity), and their scale scores from three NWEA tests (fall, winter, spring) during the school year were also included. The system use data was provided in multiple forms aggregated at different grain sizes: year, week, session, objective and level. Unfortunately, the puzzle level data is not logged, which prevented us from analyzing student’s behaviors and responses during their interactions with individual puzzles.

Based on these data, statistical analyses and visualizations were conducted to examine the patterns of ST Math usage from the four aspects: intensity of use, engagement of students with ST Math, productivity, and performance in ST Math. Generally, the analysis was exploratory. Depending on the characteristics of the dataset being examined, descriptive statistics were calculated to quantitatively describe the sample and their usage of ST Math. Analysis of Variances (ANOVA) and linear regression models were used to analyze the differences among group means in the sample data. When applicable, different kinds of plots, such as dot plot, line plot, box plot, bar graph, time series plot, and heatmap, were generated to help illustrate patterns over time, or to depict differences among groups.

## 2.1 Findings

Students in the sampled schools spent on average 2,044 min over the school year with **large variance among teachers and within the same students over time**. The analysis showed that students of different ethnicities vary in their time on task, even within the same teacher's classroom (ANOVA  $p = 0.001$ ), which is surprising as ST Math was used mostly in schools during class in these schools. It suggests that (a) students' progress during the first 10 weeks or so was fairly consistent across ethnicity, (b) White and Asian students started demonstrating faster progress than other ethnicity groups after approximately the 15<sup>th</sup> week and continued to exhibit this trend throughout the use of ST Math

Based on student's level attempt time data, we categorized teachers' classrooms into 2 groups based on whether students accessed ST Math objectives in the same sequence as designated in ST Math program (Group A), or not (Group B). We then analyzed the use data to see whether students' productivity in the program differed between the two groups. The results showed that: students in Group B spent significantly more time in the program ( $p = 0.056$ ) in shorter sessions ( $p = 0.06$ ). Students in Group B made significantly more extra tries, replayed more levels, and passed the same levels repeatedly more often than students in Group A, even though the two groups did not differ on the total number of unique levels passed. The results suggested that when following the designated order in ST Math, **students tended to be more "productive", namely completing the same amount of work in a shorter time**, with fewer replays or failure.

We examined the relationship between the total number of minutes students have spent in ST Math and student progress in the program. The results suggested that for students of mid- or low-level performances, there was **a positive association between spending time practicing problem solving in ST Math and increased learning gain as measured by NWEA tests**, especially if students were able to progress through the program further. However, given this is not a controlled study, the finding does not permit making any causal inferences between of use of the ST Math program and a student's learning outcome.

We developed two **metrics to examine difficulty of objectives**: (a) objective completion rate (i.e., the percentage of students who completed an objective out of all students who have worked on an objective), and (b) average number of minutes students needed to complete an objective. We then looked at the average amount of time

students needed to complete one level in each objective, sorted by the time needed. We noticed that several objectives under Operations and Algebraic Thinking tended to require more time to complete, followed by those under the goal of Measurement and Data, Number and Operations, and Geometry. Combining this with data on completion rate, we identified the 4 objectives that are relatively challenging for students and shared with MIND. To visually examine the difficulty of Levels, we visualized each individual level with the X-axis being the number of attempts students needed to make in order to pass a level, and Y-axis the cumulative percentage of students who passed the level. Generally speaking, in this form of representation, Levels with longer smooth lines could be more challenging, and the “passing probability added” per attempt was higher for the ones with steeper regression slopes. Overall, the low passing rate on the first attempt, along with multiple attempts being needed to pass the level, suggested that Levels 1, 2, and 4 in Objective 4 were more challenging, comparing to other levels in the same objective.

### 3 Recommendations for Developers

The results from the study suggest several focus areas for MIND to pursue including examination of ST Math program content and further in-depth analysis via observations, cognitive labs, and more quantitative analysis of usage data of a bigger sample of students. The recommendations based on the findings are

- Considering changes for classroom implementations. Students of lower incoming knowledge may need more support to make progress in the program; it may be more important to improve learning efficiency than spending longer time in the program. The results suggested that simply allocating excessive amounts of time (esp. after 2000 min) will not help all students make progress.
- Considering examining the content regarding difficulty of Levels and Objectives. Several levels in the game were particularly challenging for students. Levels within an objective sometimes do not progress from easiest to most challenging. The findings may be used to inform discussions around reordering or inserting levels. The data also showed that for certain levels in the program, a portion of students had to repeatedly attempt the levels over 10 times to pass, and for selected levels, repeated attempts did not significantly help increase passing rate.
- Considering order of Objectives. Including results from this study in discussions during professional development with teachers may help teachers understand the reasoning behind the recommended order of objectives.

### 4 Conclusion

Analyses were conducted to address research questions that focused on the use pattern, variance among students of different characteristics, difficulty of content in the program, and the relationship between usage and students’ external test scores. We highlighted a few key findings and recommendations that may help improve the design

and implementation of ST Math. While the findings are considered valuable, these analyses and corresponding finding were exploratory and had limitations. There was limited data to draw more definitive conclusions. The students, teachers, and schools were not randomly sampled and may not be representative of all ST Math users in general. Limited information was available about the small sample of users. The analysis conducted during this study was exploratory by nature and doesn't support making any causal inferences between of use of ST Math and a student's learning outcome.

## References

- Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., Chan, D.: School composition and the black–white achievement gap (NCES 2015-018). U.S. Department of Education. National Center for Education Statistics, Washington, DC (2015). <http://nces.ed.gov/pubsearch>. Accessed 24 Sept 2015
- Campuzano, L., Dynarski, M., Agodini, R., Rall, K.: Effectiveness of reading and mathematics software products: findings from two student cohorts (NCEE 2009-4041). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC (2009)
- Pane, J., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of cognitive tutor Algebra I at scale. *Educ. Eval. Policy Anal.* **36**(2), 127–144 (2014)
- Pape, S.J., et al.: Classroom connectivity in Algebra I classrooms: results of a randomized control trial (2010)
- Roschelle, J., et al.: Integration of technology, curriculum, and professional development for advancing middle school mathematics: three large-scale studies. *Am. Educ. Res. J.* **47**(4), 833–878 (2010)
- Roschelle, J., Feng, M., Murphy, R., Mason, C.: Online mathematics homework increases student achievement. *AERA Open* **2**(4) (2016). <https://doi.org/10.1177/2332858416673968>
- Rutherford, T., et al.: A randomized trial of an elementary school mathematics software intervention: spatial-temporal (ST) math. *J. Res. Educ. Effectiveness* **4**, 358–383 (2014). <https://doi.org/10.1080/19345747.2013.856978>
- VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**(4), 197–221 (2011)
- What Works Clearinghouse: WWC intervention report: a summary of findings from a systematic review of the evidence—cognitive tutor. Institute of Education Sciences, US Department of Education, Washington, DC (2016). [http://ies.ed.gov/ncee/wwc/pdf/intervention\\_reports/wwc\\_cognitivetutor\\_062116.pdf](http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_cognitivetutor_062116.pdf). Accessed 24 June 2016



# Bayesian Diagnosis Tracing: Application of Procedural Misconceptions in Knowledge Tracing

Junchen Feng<sup>(✉)</sup>, Bo Zhang, Yuchen Li, and Qiushi Xu

17zuoye, Chaoyang District, Beijing 100020, China  
junchen.feng@17zuoye.com

**Abstract.** Bayesian diagnosis tracing model (BDT) replaces the generic “wrong” response in the classical Bayesian knowledge tracing model (BKT) with a vector of procedure misconceptions. Using a novel dataset with actual student responses, this paper shows the BDT model has better interpretability of the latent factor and minor improvement in out-sample predictability in some specification than the BKT model.

**Keywords:** Procedural misconception · Bayesian knowledge tracing · Bayesian diagnosis tracing · Hidden Markov model

## 1 Introduction

### 1.1 Motivation

In our frequent exchanges with front-line teachers, a question often arises: “What does the 84% mastery mean in reality? Could you show us what students actually submitted?” Teachers are not only interested in predicting whether a student gets a question wrong, but also how they get it wrong. For example, the most frequent wrong answer to  $54 - 26$  is 38: students forget to trade a ten from the digit in tens. A less frequent wrong response is 32, which is caused by misunderstanding the rule of decomposition and treat the larger number in each digit as minuend ( $5 - 2 = 3$ ,  $6 - 4 = 2$ ). The latter error exposes a more critical procedure misconception of subtraction. However, The Bayesian Knowledge Tracing (BKT) model (Corbett and Anderson [1]) cannot answer the question of “how” because of an implicit assumption that the response is a binary variable, thus all wrong responses are qualitatively the same.

### 1.2 Literature Review

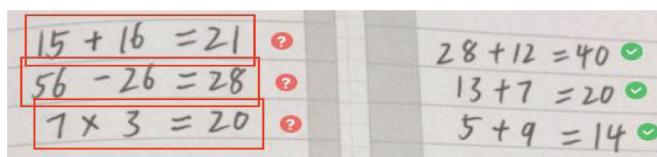
Pelánek and Desmarais both provid the latest literature review on this extending the BKT model [2, 3]. Among them, the most influential innovations are contextual slip and guess parameter (Baker et al. [4]), individualized model (Yudelson et al. [5], Pardos and Neil [6]), and Deep knowledge tracing (Piech et al. [7]). Instead of elaborating the latent factor structure, this paper proposes to enlarge the observations. Such idea draws inspirations from VanLehn [8]’s work on procedure misconceptions. Liu et al. [9]

encodes the misconception in the structures of knowledge components. In contrast, this paper treats the misconceptions as observable responses.

## 2 Diagnosis of Procedure Misconceptions

### 2.1 Dataset

The dataset comes from the Optical Character Recognition (OCR) of mental arithmetic practice booklet. Mental arithmetic means no vertical procedure. A student writes the answer on the booklet and takes a photo. An app auto-mark the photographed booklet so that a teacher does not need to. Figure 1 is a screenshot of a marked booklet.



**Fig. 1.** The OCR of an mental arithmetic practice booklet

The paper extracts two-digit subtraction items from the OCR data submitted during December 2018. It excludes students who practiced less than 5 times or more than 200 times. The remaining dataset includes 627,330 practices from 22,395 students, with a correct percentage of 92%.

### 2.2 Misconception Diagnosis

This paper identifies the following procedure misconceptions: forget borrowing a ten ( $54 - 26 = 38$ ), miss one ( $54 - 36 = 27/39$ ), miss the digit of tens ( $54 - 36 = 8$ ) and general misconception of subtraction. The last category includes unnecessary trading a ten from the next digit ( $56 - 24 = 22$ ) and treating larger number as the minuend in each digit ( $54 - 26 = 32$ ). “skip” is not procedure misconceptions but frequent enough to merit its own category: leave a line empty (“ $54 - 26 = \underline{\hspace{2cm}}$ ”) or fill it with a number from the expression (“ $54 - 26 = 26$ ”). Table 1 lists the distribution of wrong responses.

**Table 1.** Distribution of wrong responses

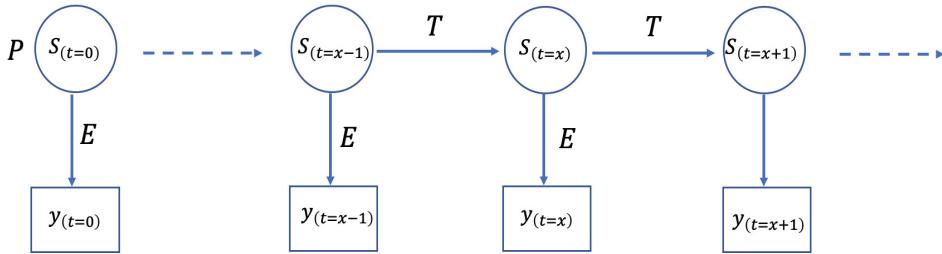
Pattern	Example	Percentage
Skip	$54 - 26 = \underline{\hspace{2cm}}; 54 - 26 = 54$	12.4%
Forget Borrowing a Ten	$54 - 26 = 38$	13.5%
Miss One	$54 - 26 = 27; 54 - 26 = 29$	8.9%
General Misconception	$54 - 26 = 32; 56 - 24 = 22$	7.1%
Miss the Digit of Tens	$54 - 26 = 8$	3.2%
No-diagnosis	$54 - 26 = 1;$	54.8%

It should be noticed that more than half of the wrong responses are not diagnosed: Even for such a quite simple arithmetic operation, the distribution of misconceptions has a very long tail.

### 3 Bayesian Diagnosis Tracing Model

The misconception-as-observation model is called Bayesian Diagnosis Tracing Model (BDT), to distinguish it from the classical BKT model [5, 10, 11]. The BDT model consists of three parameters: the priors ( $P$ ), transition matrix ( $T$ ) and emission matrix ( $E$ ). The likelihood function of BDT model is given in Eq. (1) [12]: (Fig. 2)

$$P(Y_t|S_t) = \left( P(S_0) | P(Y_0|S_0) \prod_{t=1}^x P(S_t) | P(Y_t|S_t) \right) / P(Y_{0:t}|S_{0:t}) \quad (1)$$



**Fig. 2.** HMM representation of the Bayesian diagnosis tracing model.

#### 3.1 Two-State Latent Factor Model

The BKT model does not allow for forgetting. However, such specification performs poorly in this dataset. Therefore, the BKT model reported in this paper has a full transition matrix. For the sake of comparison, the BDT parameters are reformatted in the form of BKT by ignoring the intermediate state. Table 2 shows the two models have very similar parameters. The out-sample AUC of two models are both around 0.943. In the simplest latent structure, the two models are essentially equivalent.

**Table 2.** Parameter comparison in the forms of BKT model

	Guess	Slip	Learn rate	Forget rate	Prior density
BDT	0.45	0.05	0.1	0.01	0.9
BKT	0.51	0.04	0.09	0.01	0.87

Table 3 reports the BDT emission probabilities. The mastery students do not skip or incur the two misconception (general misconception and miss the digit of tens).

**Table 3.** Emission probabilities of two-state BDT model

	Right	No diagnosis	Skip	Miss one	Forget borrowing	General misconception	Miss the digit of tens
No mastery	0.45	0.29	0.14	0.03	0.04	0.03	0.01
Mastery	0.95	0.03	<0.01	0.01	0.01	<0.01	<0.01

### 3.2 Three-State Latent Factor Model

This section employs a three-state model (No Mastery, Intermediate, Mastery) to better illustrate the benefit of misconception as observation. For better parameter convergence, the latent factor can only transit to the adjacent state. For the theoretical motivation of such specification, see Chap. 1 of Feng [4].

Table 4 reports the emission probabilities. The factors of the BDT model are more interpretable compared with the BKT: The no mastery state skips a lot; the intermediate state is prone to various misconceptions; the mastery state performs almost perfectly except for the most common misconceptions. The interpretable states are not only easy to communicate but also are helpful in constructing remedial instruction. In this case, students who skip and students who slip shall be treated differently: The no mastery students may need heavy intervention, such as interactive course or video tutoring; while the intermediate students can receive light-weight help, such as hint or more practices.

**Table 4.** Emission probabilities of the three-state model

Parameter	No mastery		Intermediate		Mastery	
	BDT	BKT	BDT	BKT	BDT	BKT
Right	0.32	0.11	0.63	0.79	0.96	0.97
Wrong	–	0.89	–	0.21	–	0.03
No diagnosis	0.01	–	0.24	–	0.02	–
Skip	0.67	–	0.01	–	<0.01	–
Miss one	<0.01	–	0.03	–	<0.01	–
Forget borrowing a ten	<0.01	–	0.04	–	0.01	–
General misconception	<0.01	–	0.03	–	<0.01	–
Miss the digit of tens	<0.01	–	0.01	–	<0.01	–

Besides the gain of interpretability, the BDT model also performs better in out-sample predictability. The out-sample AUC of the BDT model is 0.9243 while that of the BKT model is 0.9038.

## 4 Discussion

This paper explores the benefit of using procedure misconceptions as observation in the HMM model. The BDT model is more accurate in prediction and more interpretable in diagnosis for high dimension latent state model, when compared with the BKT model.

However, there is more work to be done. For one thing, little is known about the tail of the distribution, whose diagnosis can improve BDT performance. For another thing, the BDT model has great potential in analyzing problems that has multiple knowledge components because identified misconceptions can accurately find the component(s) to blame.

## References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User Adapt. Interact.* **4**(4), 253–278 (1994)
2. Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Model. User Adapt. Interact.* **27**(3–5), 313–350 (2017)
3. Desmarais, M.C., Baker, R.S.: A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User Adapt. Interact.* **22**(1–2), 9–38 (2012)
4. Feng, J.: Essays on learning through practice. Doctoral dissertation, The University of Chicago (2017)
5. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized Bayesian knowledge tracing models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 171–180. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_18](https://doi.org/10.1007/978-3-642-39112-5_18)
6. Pardos, Z.A., Heffernan, N.T.: Modeling individualization in a Bayesian networks implementation of knowledge tracing. In: De Bra, P., Kobsa, A., Chin, D. (eds.) UMAP 2010. LNCS, vol. 6075, pp. 255–266. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13470-8\\_24](https://doi.org/10.1007/978-3-642-13470-8_24)
7. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
8. VanLehn, K.: Mind Bugs: The Origins of Procedural Misconceptions. MIT Press, Cambridge (1990)
9. Liu, R., Patel, R., Koedinger, K.R.: Modeling common misconceptions in learning process data. In: Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, pp. 369–377. ACM (2016)
10. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
11. Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Beyond knowledge tracing: modeling skill topologies with Bayesian networks. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 188–198. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07221-0\\_23](https://doi.org/10.1007/978-3-319-07221-0_23)
12. Ghahramani, Z.: An introduction to hidden Markov models and Bayesian networks. In: Hidden Markov Models: Applications in Computer Vision, pp. 9–41 (2001)



# Analysis of Gamification Elements. A Case Study in a Computer Science Course

Miguel García Iruela<sup>1</sup>(✉), Manuel J. Fonseca<sup>2</sup>, Raquel Hijón Neira<sup>1</sup>, and Teresa Chambel<sup>2</sup>

<sup>1</sup> Universidad Rey Juan Carlos, Móstoles, Spain  
{miguel.garciai,raquel.hijon}@urjc.es

<sup>2</sup> LASIGE, Facultade de Ciências, Universidade de Lisboa, Lisbon, Portugal  
{mjfonseca,mtchambel}@ciencias.ulisboa.pt

**Abstract.** Nowadays, researchers are increasingly interested in the study of gamification. Gamification is the application of typical game elements in other areas. This technique can be used in different sectors like health, marketing, politics or education. In this paper we have focused on education. There are many papers related to the impact that this methodology has on the student's learning, but only a few of them delve into the gamified elements. They pay attention to other aspects such us engagement, flow, or motivation. A gamified course can be made up of a great variety of components which provide students with a game experience. The aim of our research is to compare and evaluate a list of components. Our results can serve as guidance for choosing components in educational environments and, furthermore, they can be a great support for teachers to design gamified courses.

**Keywords:** Gamification · Game-based learning · Elements of gamification · Satisfaction

## 1 Introduction

Gamification is the application of typical game elements in other areas. Although these elements were already used previously, the term “gamification” does not appear until 2002 by Pellin [1]. Gamification is considered a branch of “Game-Based Learning” [2]. Game mechanics are used in search of an improvement of motivation, to obtain an objective, to reinforce behaviour in the resolution of problems, to improve productivity or to increase engagement.

There is obvious interest in the use of gamification in various areas to achieve better results. We can find examples in marketing campaigns [3], applications that encourage good habits [4], companies dedicated to the sports sector [5], education [6], etc. At the same time, scientific research has started to provide evidence of the benefits of these techniques [7, 8]. This paper is not going to study the benefits of using gamification. We focus on studying the satisfaction of the students doing a gamified course and the evaluation of the different gamified elements used.

## 2 The Gamified System

To design the gamified system, we first consult studies that provide us with a list of the most common elements. In [9] the authors mentioned a component's list. We designed a Moodle course using this list as basis. The Moodle platform already includes some elements: badges, time limits, unlock content, profile image (Avatar) and comments. To include more components, we used a plugin that allows us to incorporate points, levels, progress bar and classification. We decided not to incorporate all the elements because that could be excessive and detrimental to the satisfaction of the users [10].

## 3 Method

Two groups of students were assigned to do the experience. All the students had exactly the same tasks. The number of points assigned, the number of levels, the number of badges and other elements coincided. The time in which the tasks were performed was the same.

### 3.1 Participants

Initially 200 students were registered, all of them were enrolled in the database course of the Faculty of Sciences of the Universidade de Lisboa. The course belongs to the first semester of the second year of the degree in Information Technology. The course designed for the experience served as support for the delivery of the entity relationship model and the relational model.

As is typical in the second year of bachelor's degrees, the age of the majority of the students was around 20, with some students of higher ages. We divided the students into two groups. The first group was composed of 96 students, of which, 57 completed the experience, while the second group started with 94 participants, of which 58 finished.

### 3.2 Experience

The students were randomly distributed between the two groups. The course implemented consisted of four missions, each of them lasting one week. Group one had the first two gamified weeks, the last two weeks stopped having gamified components. Group two started as a control group in the first two weeks, and then the gamified components were enabled.

### 3.3 Data Collected

This study seeks to evaluate the users' satisfaction and the elements of gamification used. In both groups, users completed a survey midway and another at the end. We had one that only measured satisfaction applied after the two non-gamified weeks, and another that measured the satisfaction and the evaluation

of the components, after the two gamified weeks. In group one, we went from gamification to non-gamification, while in the second group it was the opposite. To measure satisfaction, we created our questions by adapting the satisfaction part of the USE [11] in which seven questions are elaborated valued from 1 (Strongly disagree) to 7 (Strongly agree) and N/A. The questions regarding the components are based on the research [12] and they are valued from 1 to 5. In our case, instead of interviewing students, they were left with an open question to leave their comments.

## 4 Results

During the experience two groups of students were analyzed, the first of them started with gamification the first two weeks and ended without it. The second group started without gamification and ended gamified. Through the survey conducted halfway and at the end, we seek to analyze the satisfaction of the students and their evolution. In addition, the students were able to evaluate the different components used in the gamification.

### 4.1 Satisfaction Analysis

In the following table we can find the answers of the students to the different questions about satisfaction in the first questionnaire as well as the second one (Table 1).

**Table 1.** Answers about satisfaction.

	Group 1		Group 2	
	Quest. 1	Quest. 2	Quest. 1	Quest. 2
1. I am satisfied with it	4,45	4,06	4,16	4,10
2. I would recommend it to a friend	4,12	3,77	3,81	3,84
3. It is wonderful	3,71	3,35	3,18	3,51
4. It is pleasant to participate	4,34	3,98	3,84	3,88
5. Learning experience was worthwhile	4,91	4,54	4,44	4,71
6. I learned about the course topic with the tool	4,95	4,63	4,84	4,98
7. I was involved	4,88	4,71	4,54	4,69

### 4.2 Elements Valuation

Points, levels, feedback and missions have a value higher than a 3. Highlight the 4 obtained when considering the necessary feedback, the students also choose a

3 for the need to obtain a report immediately, on the other hand, with 2.5 if the use of feedback in each task was excessive. Regarding the points, they recognized that they liked the points and that they were motivated, they awarded with a 3 that the points should be obtained only in important tasks.

Badges are also well valued with more than a 3, although they considered them less motivating than the points. Leaderboard obtained a 2.67, compared with a 2.55 which motivated them to work harder and did not consider a big problem with a 2.34 possible negative effect that may have. Regarding the blocked content students felt good to have access to new content and they valued with a 2.91 the importance of getting the content gradually, although, on the contrary, they like to have access to all the content from the beginning. The time limit makes students feel stressed, but at the same time it makes them work harder.

## 5 Conclusions

This paper presents the analysis of data about the satisfaction of students in a gamified course and the assessment of the different elements. The data collected showed a slightly higher satisfaction of students belonging to a gamified course versus students in a non-gamified course. In group one the satisfaction fell when they stopped being gamified, while satisfaction in group two went up in the second part when they started to be gamified. The data obtained reflects that gamification achieves a slightly higher satisfaction among students. This small difference may be due to the short period of time used.

All the components were evaluated positively above 2.5 out of 5. The feedback is an element highly valued by the students and which they consider to be of great help. The points levels, missions and badges were also well valued, their inclusion can have positive effects to achieve greater student satisfaction. The time limit achieved an improvement in the work, although it caused them some stress, so this must be well controlled for proper functioning. Although students prefer to have full access to content, this can motivate students to do regular work. The leader's table only had a 2.6 valuation, as an advantage, it emphasizes that the students did not consider that it could have too many negative effects.

To achieve high student satisfaction, it is necessary to take into account a wide variety of factors. The gamification, if it is well thought out, can provide an increase in satisfaction, but a good design of the tasks, a good programming and a good adaptation of the contents are of great importance. This study gathers data from a wide group of Portuguese students of Information Technology, but it could change for another group of students with different characteristics or for another course. However, the study provides valuable information that may facilitate additional studies in this area.

As future work, it is proposed to expand the student spectrum of the test in order to check the effectiveness of the components in groups of students with different profiles and even of different nationalities.

**Acknowledgment.** This research received financial support from Ministerio de Economía y Competitividad (TIN2015-66731-C2-1-R), Comunidad Autónoma de

Madrid (P2018/TCS-4307), and Rey Juan Carlos University (30VCPIGI15) and the LASIGE Research Unit (UID/CEC/00408/2019).

## References

1. Marczewski, A.: Gamification: A Simple Introduction and a Bit More, 1st edn. [E-book], United Kingdom (2012)
2. Deterding, S., Dixon, D.: From game design elements to gameness: defining “Gamification”. In: Proceedings of the 15th International Academic MindTrek Conference, pp. 9–15. ACM, Tampere (Finland) (2011)
3. Chou, Y.-k.: Gamification and behavioral design, top 10 marketing gamification cases you won’t forget. <https://yukaichou.com/gamification-examples/top-10-marketing-gamification-cases-remember/>. Accessed Jan 2019
4. Chou, Y.-k.: Gamification and behavioral design, the 10 best productivity apps that use gamification in 2018. <https://yukaichou.com/lifestyle-gamification/the-top-ten-gamified-productivity-apps/>. Accessed Jan 2019
5. Chou, Y.-k.: Gamification and behavioral design, top 10 fitness gamification examples to get fit in 2017. <https://yukaichou.com/gamification-examples/top-10-fitness-gamification-examples-get-fit-2017/>. Accessed Jan 2019
6. Duolingo Homepage. <https://es.duolingo.com/>. Accessed Jan 2019
7. García Iruela, M., Hijón Neira, R.: How gamification impacts on vocational training students. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 99–103. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_19](https://doi.org/10.1007/978-3-319-93846-2_19)
8. Jang, J., Park, J.J.Y., Yi, M.Y.: Gamification of online learning. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 646–649. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_82](https://doi.org/10.1007/978-3-319-19773-9_82)
9. Biel, L.A., García Jiménez, A.M.: Gamificar: el uso de los elementos del juego en la enseñanza del español. In: L Congreso internacional de la AEPE (Asociación Europea de Profesores de Español), pp. 73–83 (2015)
10. Werbach, K., Hunter, D.: For the Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press, Philadelphia (2012)
11. Lund, A.: Measuring usability with the USE questionnaire. Usability Interface **8**(2), 3–6 (2001)
12. Aldemir, T., Celik, B., Kaplan, G.: A qualitative investigation of student perceptions of game elements in a gamified course. Comput. Hum. Behav. **78**, 235–254 (2018)



# Towards Adaptive Worked-Out Examples in an Intelligent Tutoring System

Nicholas Green<sup>1</sup>, Barbara Di Eugenio<sup>1()</sup>, and Davide Fossati<sup>2</sup>

<sup>1</sup> University of Illinois at Chicago, Chicago, IL 60607, USA

[bdieugen@uic.edu](mailto:bdieugen@uic.edu), [contact@nickgreen.co](mailto:contact@nickgreen.co)

<sup>2</sup> Emory University, Atlanta, GA 30322, USA

[davide@fossati.us](mailto:davide@fossati.us)

**Abstract.** Worked-out examples (WOEs) have been shown to be effective for learning, but they need to be adapted to student characteristics. We experimented with three versions of our Intelligent Tutoring System for Computer Science, one that does not include WOEs, and two that differ as concerns WOE length and content. We found that shorter WOEs are more effective for advanced students, whereas novice students learned the same, no matter the presence or length of the WOE.

**Keywords:** Worked out examples · Intelligent tutoring · Computer Science education

## 1 Introduction

Worked-out examples (WOEs) provide a step-by-step example in solving a problem. WOEs have been studied since the mid '80s, and often found to be effective, especially for novices. Their effectiveness is often linked to reduced cognitive load in problem solving [12]. WOEs have been used and experimented with in a variety of Intelligent Tutoring Systems (ITSs) in scientific domains, e.g. logic and Computer Science (CS) [5, 10]. Results are mixed as concerns what types of students benefit the most, and from what kind of WOEs: e.g. a WOE that thoroughly illustrates a correct solution, or a faded WOE where the student has to actively engage in problem solving [11]. Here, we pose the question of whether length and content affect the effectiveness of WOEs for novice and advanced students. We conducted experiments with our ITS, ChiQat-Tutor (ChiQat): we found that novices learned the same amount no matter the presence or length of WOEs; however, short WOEs were more effective for advanced students.

## 2 Related Work

WOEs have been used in several ITSs in scientific domains, with positive, but often mixed results: features of the learner and/or of the examples themselves can affect learning and/or efficiency thereof. McLaren et al. [9] integrated WOEs into

a chemistry ITS that also includes tutored problem solving. No significant gains were observed over problem solving only, however students were able to learn faster. Further research [8] showed there were no significant learning gain differences over four conditions, including WOEs and erroneous examples (EWOE), but students who studied via WOEs spent 46%–69% less time to complete the activity. Barnes et al. [7, 10] used WOEs in an open-ended data-driven logic tutor. The authors showed that WOEs were effective for novices, however had the same effect as hint-based systems; advanced students did not benefit. In a CS domain, we compared providing WOEs to providing analogies, with mixed results [6]. Mitrovic et al. [1] explored adaptivity of several strategies in their ITS for teaching SQL queries. First, they showed that alternating EWOEs to WOEs engenders more learning than WOEs alone; and that an adaptive strategy that provides students with either a WOE or an EWOE based on their performance, results in more efficient learning.

### 3 Worked-Out Examples in ChiQat-Tutor

ChiQat is an ITS that helps students learn core CS data structures, and mostly focuses on linked lists: students are provided with a linked list curriculum consisting of seven problems. A problem consists of an initial state, and a goal state that the student has to reach by manipulating the list via code (Java or C++); ChiQat provides various types of verbal feedback in order for the student to reach the correct answer [3]; additionally, a graphical representation is automatically updated after executing the student’s code. For each problem, students can call up a WOE on demand by clicking on the *Example* button in the interface. Each problem includes its own specific WOE (in two guises, see below): e.g., the WOE for Problem 1 that asks for a number to be inserted into the middle of the list, will demonstrate the operation on a different list. A student needs not finish a WOE before going back to solving the problem, however, returning to the WOE will restart from the beginning.

**Table 1.** Short worked-out example for problem 1 (Step types not shown to students)

1	Definition	Here is how you insert ‘3’ in between ‘2’ and ‘4’
2	Operation	First create a node ‘3’, pointed to by ‘Z’
3	Operation	Next we need to find the node ‘2’ and assign it to a new variable, S
4	Operation	Connect ‘3’ to the ‘4’
5	Operation	and then ‘2’ to ‘3’
6	Operation	Remove variables that are not needed, and we are done

Each problem in the linked list module contains two WOEs demonstrating how to solve the same problem, in two guises: *Short* (Table 1) and *Long* (Table 2).

**Table 2.** Long worked-out example for problem 1 (Step types not shown to students)

1	Introduction	OK, (USER), we're going to take a look at how we can insert items into a linked list
2	Introduction	Take a look at this list
3	Definition	What we want to do is to insert a '3' in between '2' and '4'
4	Operation	The very first step is simple, lets create a new node with a value of '3', and why don't we call it 'Z'
5	Explanation	Now we need to insert a node AFTER '2', the one that's flashing
6	Operation	Firstly, we should get access to the second node. This can be done by going through the root, T, and getting its next node. This can be assigned to a variable, S
7	Operation	From here, we could assign 2's next pointer to the node containing 3, like so
8	Reflection	However, there is a problem here, think about it...
9	Explanation	How do you reattach the node containing 4? The connection now has been lost!
10	Explanation	Lets take a step back and see how we can do this without losing this vital connection
11	Operation	Lets connect the node containing 3 to the node containing 4
12	Operation	Now lets do what we done before and connect 2 to 3
13	Operation	and then tidy up some of the references, this being S and Z
14	Conclusion	And there we go, 3 has been inserted into the list between 2 and 4!

Long WOEs faithfully reproduce WOEs from our human-human tutoring dialogues on introductory CS data structures [2]; short WOEs are systematically derived from Long WOEs by removing non-essential steps and more verbose language - namely, keeping only *definition* and correct *operation* steps. In Problem 1, the Long WOE also includes the discussion of an erroneous solution (steps 7 through 9), thereby connecting to the earlier studies we discussed that compare WOEs to EWOEs. This specific Long WOE has 14 steps and 212 words; the corresponding Short WOE for Problem 1 has 6 steps and 56 words.

## 4 Initial Student Knowledge, WOEs, and Learning

We ran experiments with students in two introductory CS courses at our institution for four consecutive semesters. Students used ChiQat in one regular laboratory section, held about a week after their first exposure to linked lists in class. Each student worked on the same problem set in the system, but was randomly assigned to one of three conditions: no WOE available, long WOE, or short WOE. Students took a pre-test before interacting with the system, and then

took the same post-test after interaction; as learning gain, we use the difference in score between pre and post-test.

In general, we did not find differences in learning gains across conditions when considering all students together. However, given the earlier results on WOEs being more effective and/or efficient for novices [7, 10] we investigated our data from this angle. We define a novice student as a student who scored at or below the median score of all students in the pre-test; the others are advanced students, for a 132/141 split (Table 3). Whereas novice students could be defined in other ways, this distinction is borne out by comparing the two groups' pre-tests, ( $\mu = 0.30, \sigma = 0.09$ ) and ( $\mu = 0.65, \sigma = 0.15$ ) respectively, which are significantly different ( $t = -23.141, df = 271, p < 2.2e^{-16}$ ).

Table 3 shows the distribution and learning gains of novice and advanced students, per condition. The number of subjects in the Long WOE condition is larger because it is an aggregate of three subconditions, in which the same Long WOEs were presented to students, and different timing features were experimented with. In several experiments, students could peruse Long WOEs without any constraints (115 subjects); in one experiment, students had to bring the WOE to conclusion (28 subjects); in another one, a time limit was imposed on perusing the example (32 subjects).

**Table 3.** Novice/advanced student learning gains and distribution, per condition

	No WOE			Short WOE			Long WOE			Total		
	$\mu$	$\sigma$	$N$	$\mu$	$\sigma$	$N$	$\mu$	$\sigma$	$N$	$\mu$	$\sigma$	$N$
Initial knowledge												
Novice	0.19	0.21	25	0.14	0.16	25	0.15	0.17	82	0.16	0.17	132
Advanced	0.04	0.17	28	0.10	0.14	20	0.01	0.16	93	0.03	0.16	141
Total	0.11	0.20	53	0.12	0.15	45	0.07	0.18	175	0.09	0.18	273

The difference in learning gains between novice ( $\mu = 0.16, \sigma = 0.17$ ) and advanced ( $\mu = 0.03, \sigma = 0.16$ ) is highly significant ( $t = 6.4453, df = 271, p = 5.257e^{-10}$ ). This result is also borne out by a two factor ANOVA: there is an effect of prior knowledge ( $F = 40.8845, p = 7.181e^{-10}$ ), but no effect of condition, and no interaction.

The result that novices learn more than advanced students when interacting with ChiQat is reassuring, but not surprising. We further analyzed novices and advanced students separately. For novices, we did not find any significant difference among conditions. For advanced students, we conducted a planned comparison ANOVA that distinguishes between learning gains as follows: we compared no WOEs ( $\mu = 0.04, \sigma = 0.17$ ), short WOEs ( $\mu = 0.10, \sigma = 0.14$ ), and long WOEs ( $\mu = 0.01, \sigma = 0.16$ ). We found a significant difference ( $F = 3.2, p = 0.0438$ ). Tukey post-hoc tests revealed that it is the difference between short and long WOEs which is significant ( $p = 0.03897$ ). This result suggests that, when provided with WOEs, advanced students learn more from more concise examples.

## 5 Conclusions

In this work, we explored whether learning is affected by longer or shorter WOEs in ChiQat. We found that advanced students learned more from short WOEs than from long WOEs; however, novice students learned the same amount, no matter the presence or length of a WOE. There are several limitations to our experiments. First, the three conditions (no WOE, Long WOE, and Short WOE) are not balanced as concerns number of subjects. Second, the definition of novice /advanced was based on a median split, rather than on an objective evaluation by an expert on what should be considered as beginner or advanced knowledge (notice that the pre-/post-test was developed by experts). These findings could be used for an adaptive WOE subsystem to tailor WOEs for estimated student ability levels. We explored mining the logs of ChiQat to infer the knowledge level of students from their behavior on the first problem; some preliminary results can be found in [4].

## References

1. Chen, X., Mitrovic, A., Matthews, M.: Learning from worked examples, erroneous examples and problem solving: towards adaptive selection of learning activities. *IEEE Transact. Learn. Technol.* (2019). <https://doi.org/10.1109/TLT.2019.2896080>
2. Di Eugenio, B., Chen, L., Green, N., Fossati, D., AlZoubi, O.: Worked out examples in computer science tutoring. In: 16th International Conference on Artificial Intelligence in Education. Memphis, TN, short paper, July 2013
3. Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., Chen, L.: Data driven automatic feedback generation in the iList intelligent tutoring system. *Technol. Instr. Cogn. Learn. (TICL) Spec. Issue Role Data Instr. Processes* **10**(1), 5–26 (2015)
4. Green, N.: Example Based Pedagogical Strategies in a Computer Science Intelligent Tutoring System. Ph.D. thesis, University of Illinois at Chicago (2017)
5. Green, N., Di Eugenio, B., Harsley, R., Fossati, D., AlZoubi, O.: Behavior and learning of students using worked-out examples in a tutoring system. In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) *ITS 2016. LNCS*, vol. 9684, pp. 389–395. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-39583-8\\_46](https://doi.org/10.1007/978-3-319-39583-8_46)
6. Green, N., Di Eugenio, B., Harsley, R., Fossati, D., AlZoubi, O., Alizadeh, M.: Student behavior with worked-out examples in a computer science intelligent tutoring system. In: International Conference on Educational Technologies. Florianopolis, Santa Catarina, Brazil, November 2015
7. Liu, Z., Mostafavi, B., Barnes, T.: Combining worked examples and problem solving in a data-driven logic tutor. In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) *ITS 2016. LNCS*, vol. 9684, pp. 347–353. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-39583-8\\_40](https://doi.org/10.1007/978-3-319-39583-8_40)
8. McLaren, B.M., van Gog, T., Ganoe, C., Karabinos, M., Yaron, D.: The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Comput. Hum. Behav.* **55**, 87–99 (2016)

9. McLaren, B.M., Lim, S.-J., Koedinger, K.R.: When is assistance helpful to learning? Results in combining worked examples and intelligent tutoring. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 677–680. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69132-7\\_75](https://doi.org/10.1007/978-3-540-69132-7_75)
10. Mostafavi, B., Zhou, G., Lynch, C., Chi, M., Barnes, T.: Data-driven worked examples improve retention and completion in a logic tutor. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 726–729. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_102](https://doi.org/10.1007/978-3-319-19773-9_102)
11. Renkl, A.: The worked-out-example principle in multimedia learning. The Cambridge Handbook of Multimedia Learning, pp. 229–245 (2005)
12. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. *Cogn. Instr.* **2**(1), 59–89 (1985)



# Orchestrating Class Discussion with Collaborative Kit-Build Concept Mapping

Yusuke Hayashi<sup>(✉)</sup>, Toshihiro Nomura, and Tsukasa Hirashima

Graduate School of Engineering, Hiroshima University, Higashihiroshima, Japan  
hayashi@lel.hiroshima-u.ac.jp

**Abstract.** In the collaborative learning context, typical classroom practices tend to create three distinguishable levels of activity: individual-activities, small-group work, and whole-class discussion. It is important to connect the analysis of the levels for teachers to understand and improve the dynamics of students' understanding of collaborative learning. This study, with the Kit-Build Concept Map method, proposes a method to analyze the three levels of activities in the classroom. Kit-Build Concept Map is a type of close-ended concept map and provides decomposed concepts and links from the concept map made by a teacher. This mechanism enables teachers to check students' understanding and to facilitate his or her coordination of learning in a classroom. This paper illustrates the method with the result of a case study in a junior high school in Japan.

**Keywords:** Concept map · Kit-build · Classroom orchestration · Learning analytics

## 1 Introduction

Teachers have an important role in class discussion. They look up students' understandings or opinions and facilitate to activate their discussion. This is a part of classroom orchestration that is the conceptualization of management of class by teachers [1]. This study proposes a method to help teachers to be aware of students' understandings or opinions as the results of individual and group activities before class discussion. If teachers know this information, they can coordinate classroom discussion highlighting the difference between students' understandings or opinions. This study realizes the teachers' awareness of students' understandings or opinions with a kind of concept mapping method called Kit-Build (KB) [5].

Martinez-Maldonado proposes and develops a multi-tabletop classroom and dashboard to support collaborative learning [2]. Their study provides a special environment for learners to collaboratively work with concept maps and for teachers to capture the verbal and physical interactions of learners. In contrast to their focus on activity, this study focuses on the content learners communicate and produce in class. It does not mean that this study focuses on shallow level content, for example, keyword level network analysis [4]. This study proposes to capture deeper level content with KBmap [3].

## 2 Class Discussion with Kit-Build Concept Map

The goal of the class discussion in this study is that learners organize what they have learned through the previous classes as their shared knowledge in order to make discussion based on it. In order to realize it, it is necessary for the teacher to be aware of learners' understanding and to provide them with feedback to complement and correct their understanding. For this requirement, in this study, the teacher uses the KBmap system to capture learners' collaborative learning in classes.

The main task of students in this lesson is to consider the things and their relation in what they have learnt in the lecture and compose it as a concept map. In this lesson, students build a concept map with the decomposed component of the teacher's concept map as the representation of their understanding and then exchange each other. KBmap editor becomes a learning material for learners to represent their understanding and KBmap analyzer becomes a tool for teachers to capture learners' understanding during and after class.

In the lessons what students do is the following two things:

- to organize their knowledge on KBmap inductively from materials and what they have learned in the previous lessons and
- to compare and correct their knowledge represented on KBmap through discussion.

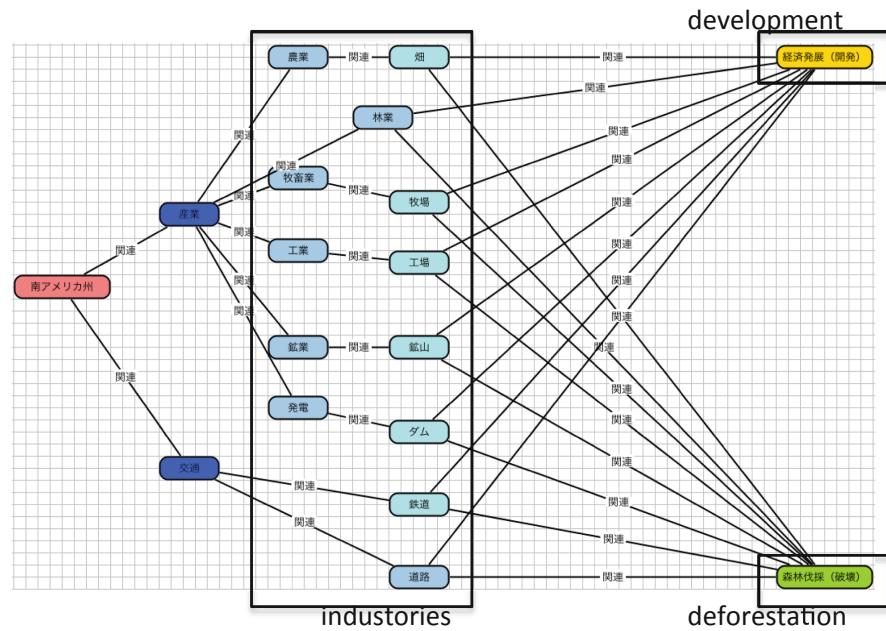
After that, the teacher explains difference between the teacher's and the students' maps. Through this process, this lesson expects that students are aware of missing relations in their understanding and that the teacher identifies students' understandings remained after discussion and taught them carefully.

The lessons are composed of the following steps:

1. Reviewing the previous lessons through instruction by the teacher.
2. Building a concept map by an individual (pre-map).
3. Building a concept map in a group (collaborative-map).
4. Modifying the pre-map by an individual (post-map).
5. Whole-class instruction by the teacher (group-map).

## 3 Data from a Case of In-Class Collaborative Learning

We conducted three lessons in a junior high school. Participants are 76 students from three classes in the first grade of the junior high school in Japan. These classes are conducted in regular classes and replace usual tasks with paper worksheets by tasks with the KBmap system on tablet computers. The topic of this class is the characteristics of Latin America, especially the relationship between economic development and deforestation. Figure 1 shows the map that a teacher builds as the correct answer and students should build, called "goal-map." This may represent that the teacher wants learners to have an image of the relationship between industries in Latin America and development or deforestation.

**Fig. 1.** A goal-map

During the classes, the teacher uses only the group-map of the collaborative-maps. With the group-map teacher was aware of learners' lack or misunderstanding of knowledge and gave feedback on them in the whole-class instruction. In our initial plan, the teacher was also planning to use group-maps of individual pre-maps during the group work. However, it was difficult to use them in the limited time of the group work.

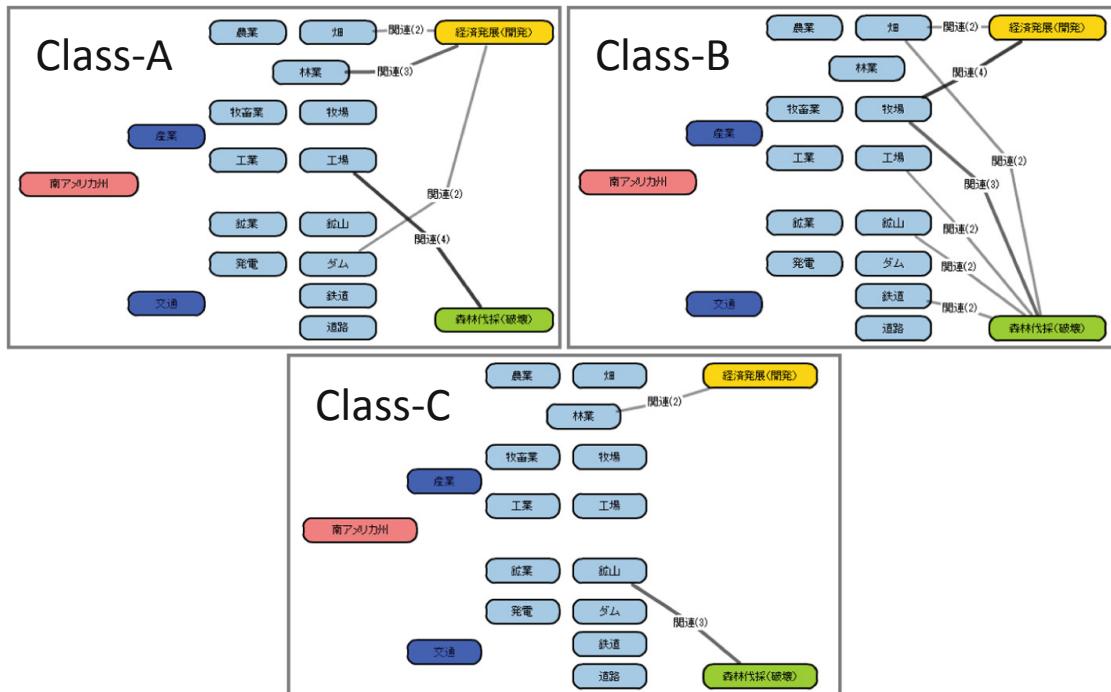
**Fig. 2.** Group maps

Figure 2 shows the group-maps emphasizing lacking links that over half of the group have not set in the collaborative-maps. In the goal-map, all the industries are linked to both development and deforestation. Therefore, the lacking links show the relation of which many learners are not aware of. As shown in Fig. 2 emphasized lacking links are different from classes. Before the classes, the teacher supposed that the difficulty of understandings is the same in different classes. However, according to group-maps, such a forecast is not always correct. After the lessons, the teacher said that this is the first time to get information about the understanding of the students on time in the classroom.

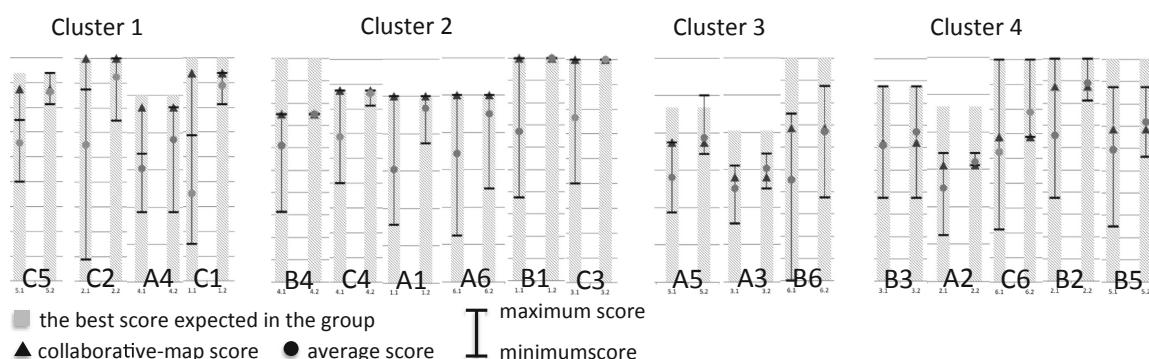
Figure 3 shows the classification of the group by hierarchical cluster analysis. As a result, there are four types of group.

**Cluster 1:** the scores of the collaborative-maps are higher than the maximum score of the pre-map in the group.

**Cluster 2:** the scores of the collaborative-maps are the same as the maximum score of the pre-map in the group.

**Cluster 3:** the scores of the collaborative-maps are lower than the maximum score of the pre-map in the group.

**Cluster 4:** the scores of the collaborative-maps are not the same as the maximum score of the pre-map in the group.



**Fig. 3.** Clusters of groups

## 4 Conclusion

This paper presents the result of a case study of implementation and data analysis of in-class collaborative learning with the KBmap approach. The teacher was aware of the understanding the status of students not quantitatively but qualitatively and give feedback to them. It is difficult for teachers to be aware of the process and the product in in-class collaborative learning. Usual methods to be aware of it are to let students give a presentation about the product or to check their conversation carefully. Although KBmap provides a kit for constructing a concept map and beats the bounds of the map, their discussion in groups and whole-class is not limited to the boundary. In other words, KBmap itself is a closed-end learning environment.

## References

1. Dillenbourg, P., Jermann, P.: Technology for classroom orchestration. In: Khine, M., Saleh, I. (eds.) *New Science of Learning*, pp. 525–552. Springer, New York (2010). [https://doi.org/10.1007/978-1-4419-5716-0\\_26](https://doi.org/10.1007/978-1-4419-5716-0_26)
2. Martinez-Maldonado, R., Clayphan, A., Yacef, K., Kay, J.: Towards providing notifications to enhance teacher's awareness in the classroom. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) *ITS 2014. LNCS*, vol. 8474, pp. 510–515. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07221-0\\_64](https://doi.org/10.1007/978-3-319-07221-0_64)
3. Nomura, T., Hayashi, Y., Suzuki, T., Hirashima, T.: Knowledge propagation in practical use of kit-build concept map system in classroom group work for knowledge sharing. In: *ICCE 2014 Workshop Proceedings*, pp. 463–472 (2014)
4. Oshima, J., Oshima, R., Matsuzawa, Y.: Knowledge building discourse explorer: a social network analysis application for knowledge building discourse. *Educ. Technol. Res. Dev.* **60** (5), 903–921 (2012)
5. Tergan, S.-O.: Digital concept maps for managing knowledge and information. In: Tergan, S.-O., Keller, T. (eds.) *Knowledge and Information Visualization. LNCS*, vol. 3426, pp. 185–204. Springer, Heidelberg (2005). [https://doi.org/10.1007/11510154\\_10](https://doi.org/10.1007/11510154_10)



# Automating the Categorization of Learning Activities, to Help Improve Learning Design

Wayne Holmes<sup>(✉)</sup> and Juliette Culver

Institute of Educational Technology, The Open University, Milton Keynes, UK  
wayne.holmes@open.ac.uk

**Abstract.** As part of the large-scale implementation of Learning Design at The Open University, the UK’s largest higher education institution, a taxonomy of learning activities informs the development of course modules. The taxonomy is also used to map a module’s Learning Design, to categorize its learning activities, after it has been developed. This enables course teams to compare a module’s Learning Design with student outcomes, in order to determine which Learning Designs are most effective and in which circumstances. However, the mapping process is labor-intensive and open to inconsistencies, making the outcomes less trustworthy and less useful for learning analytics. In this paper, we present an exploratory study that investigates the automatization of the mapping process by means of both unsupervised and supervised machine learning approaches. For the supervised machine learning (Logistic Regression), we use a labelled set of 35,000 activity descriptions classified as either reflective or non-reflective (i.e., whether or not the activity involves student reflection) drawn from 267 modules. Our outcomes, with ~ 79% accuracy, are sufficiently promising for this approach to merit further work, extending it in particular to a larger set of Learning Design activities.

**Keywords:** Learning design · Learning activities · Machine learning

## 1 Introduction

In this paper, we present a proof-of-concept study that investigated automatizing the process of mapping the learning activities in a university module to a Learning Design (LD) taxonomy, in order to facilitate better LD and robust learning analytics.

### 1.1 Learning Design

In recent years, there has been a growing interest in how teaching might be informed by an approach known as Learning Design: “a methodology for enabling teachers/designers to make more informed decisions in how they go about designing learning activities and interventions, which is pedagogically informed and makes effective use of appropriate resources and technologies” [1]. At a base level, LD provides a ‘notational framework’ to facilitate the sharing of learning designs, so that they may be iteratively improved upon [2]. Building upon teachers’ best practices [3], LD also recognises that “*different teaching approaches may be used for different subjects, and*

*at different stages in learning”* [4]. There have been multiple LD projects and initiatives, including the *SoURCE* project [5], the *Larnaca Declaration* [4], and the *Open University Learning Design Initiative* [6].

## 1.2 Learning Design at the Open University

The Open University’s Learning Design Initiative (OULDI) has now been in operation for more than ten years (currently, there is at least partial mapping data for more than 500 modules). It is “*a collaborative design approach in which OU module teams, curriculum managers and other stakeholders make informed design decisions with a pedagogical focus, by using representation in order to build a shared vision*” [7]. OULDI centers on a taxonomy of seven types of learning activities, as shown in Table 1.

**Table 1.** OULDI taxonomy of learning activities, adapted from [8].

Type of activity	Description	Examples
<i>Assimilative</i>	Attending to information	Reading, watching, listening, thinking about, accessing
<i>Finding and handling information</i>	Searching for and processing information	Listing, analyzing, collating, finding, discovering, gathering
<i>Communication</i>	Discussing module related content with at least one other person	Communicating, debating, discussing, sharing, collaborating
<i>Productive</i>	Actively constructing an artefact	Creating, building, making, designing, constructing
<i>Experiential</i>	Applying learning in a real-world setting	Practicing, applying, experiencing, exploring, investigating
<i>Interactive/adaptive</i>	Applying learning in a simulated setting	Exploring, experimenting, trialing, modeling, simulating
<i>Assessment</i>	Summative, formative and self-assessment	Writing, presenting, reporting, demonstrating, critiquing

The OULDI taxonomy of learning activities informs the development of new course modules (i.e., a module’s intended LD). In addition, after a module has been developed, the taxonomy is used to notate the final module’s activity texts (i.e., a module’s actual LD), which is combined with workload estimates (i.e., the anticipated time that an activity will take for the student to complete).

This process, known as module ‘mapping’, serves four key purposes: (i) prioritizing the student perspective of a module, (ii) aligning module planning with the final module, (iii) providing a snapshot of a module, and (iv) researching effective patterns of learning design. Although module-mapping is a neutral descriptive process (i.e., it describes but does not evaluate a module’s LD), the mapping outcomes can feed into research on effective LDs, to determine which LDs are most effective in which

circumstances (for a summary of this research see [7]). A typical finding is that the primary predictor for student retention is the relative amount of time mapped as being spent on ‘communication’ activities [7].

## 2 Automating the LD Mapping Process

### 2.1 Rationale

Although the OULDI mapping process has clear benefits, it is not without challenges. In particular, it is labor-intensive (it typically takes an LD specialist around four days to complete the mapping of one module); and it is open to inconsistencies, making the outcomes less trustworthy and less useful for learning analytics. There is therefore a large incentive to automate, or at least partially automate, the module mapping process. Such automation would enable a live view of the module’s LD at any point during a module’s development process, which is currently not possible (and would be prohibitive in terms of the labor costs). Further, automating module mapping will require a focus on, and thus might facilitate, improvements in mapping consistency.

Possible approaches to automating the module mapping process are rules-based, machine learning, or a combination. Some unpublished work using a rules-based approach has been conducted at The Open University (OU), for example using a bag-of-words model (mainly the verbs used in activity texts). However, hand-designing suitable rules has severe challenges and limitations. Accordingly, in the study presented here, we investigate the potential of machine learning approaches.

### 2.2 Dataset

The complete content of 267 modules, spanning all disciplines and levels of study taught at the OU, was obtained in XML format. Scripts were written to extract from the content all the activity texts, to remove XML tags, to exclude all duplicates and all activities with instructions in languages other than English, and to remove all parts of the activity text which were not part of the instructions (e.g., an activity might ask students to comment on a poem and then include the text of the poem, in which case the poem was removed). This process resulted in an Activity Text Dataset of approximately 35,000 activity texts.

### 2.3 Unsupervised Learning

We tested the following clustering algorithms on the Activity Text Dataset using a bag-of-words approach with the stop words removed, firstly on all the activity texts words, then restricted to the activity text’s verbs: (i) One-hot Encoding and K-Means Clustering, (ii) Tf-idf encoding and K-Means Clustering, (iii) Word2Vec and K-Means Clustering, and (iv) Latent Dirichlet Allocation. The aim was to investigate whether this approach might identify alternative clusters to the OULDI learning activity types. The results from all four algorithms, both on all words and when restricted to the verbs, produced clusters that were unhelpful: they were not focused on any specific learning

activities. Instead, the clusters centered on subject domains or subdisciplines, or on similar words (e.g., the words ‘view’ and ‘video’ might be clustered together as passive activities, even if the word ‘produce’ preceded the word ‘video’ thus making it an active activity).

## 2.4 Supervised Learning

Initially, the intention was to label the Activity Text Dataset using the seven types of learning activities from the OULDI taxonomy. However, it soon became clear that both the learning activity types and the activity text were ambiguous, such that labelling the dataset using the taxonomy would be extremely challenging. Accordingly, in order to reduce the effort required and to provide us with a robust starting point, we restricted the labelling to a single binary category: whether or not an activity was ‘reflective’. An activity was labelled as ‘reflective’ if it encouraged students to reflect on their learning or on their study process. A simple web platform was built to enable this categorization, by means of which all 35,000 activities were manually labelled by the researchers as either being ‘reflective’ or ‘not reflective’.

We investigated a variety of supervised learning algorithms (Logistic Regression, Decision Trees, and Support Vector Machines) with varying parameters: (i) custom stop word lists, (ii) stemming and tokenization, (iii) n-grams for various values of n, (iv) limiting to certain parts of speech, (v) changing the number of words used by the count vectorizer and (vi) examining false positives and false negatives. The most promising results were obtained using Logistic Regression and 20-fold cross-validation (i.e., the data was divided into 20 groups, and each group was used in turn as a test subset while the others functioned as the training subset, and then the results were averaged) with 2-grams and a custom stop word list without any stemming or tokenization. This resulted in an F1-score of 0.79. By contrast, stemming decreased the F1 value to 0.64, while just using verbs to 0.50.

## 3 Discussion and Future Directions

Although modest, the results obtained in this proof-of-concept work are sufficiently promising for the approach to merit further investigation, in particular extending it to other learning activity types from the OULDI taxonomy.

However, the fact that our attempt to use machine learning reaffirmed the OULDI learning activity types’ ambiguities, suggested that they might benefit from being revisited. Ideally, an OULDI 2.0 would comprise a comprehensive but unambiguous set of learning activity types. This would enable human mappers to be more consistent, with the differences between the activity types indicating pedagogical distinctiveness, and would provide a more robust starting point for machine learning. Once such a set of OULDI 2.0 activity types has been identified, a supervised machine learning approach would again depend on the Activity Text Dataset being manually labelled – which will require a significant amount of initial manual effort. It might also be worthwhile including workload data (the amount of time a student is expected to

expend on any particular learning activity), as well as investigating whether a combination of rules-based and machine learning approaches yields better results.

Being able to automatize the process of mapping learning activities has clear benefits (for learning designers, learning analytics and potentially learners). The question remains whether a set of more robust OULDI activity types, workload data, and a combination of rules-based and machine learning approaches will make this ambition realizable.

## References

1. Conole, G.: Designing for Learning in an Open World. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-1-4419-8517-0>
2. Koedinger, K.R., Booth, J.L., Klahr, D.: Instructional complexity and the science to constrain it. *Science* **342**, 935–937 (2013)
3. Laurillard, D., et al.: A constructionist learning environment for teachers to model learning designs. *J. Comput. Assist. Learn.* **29**, 15–30 (2013)
4. Dalziel, J., et al.: The Larnaca declaration on learning design. *J. Interact. Media Educ.* **2016**, 7 (2016)
5. Laurillard, D., McAndrew, P.: Virtual teaching tools: bringing academics closer to the design of e-learning (2001)
6. Cross, S., Galley, R., Brasher, A., Weller, M.: OULDI-JISC project evaluation report: the impact of new curriculum design tools and approaches on institutional process and design cultures (2012)
7. Rienties, B., Nguyen, Q., Holmes, W., Reedy, K.: A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK. *Interact. Des. Archit. J. IxDA* **33**, 134–154 (2017)
8. Rienties, B., Toetenel, L.: The impact of learning design on student behaviour, satisfaction and performance: a cross-institutional comparison across 151 modules. *Comput. Hum. Behav.* **60**, 333–341 (2016)



# Identifying the Structure of Students' Explanatory Essays

Simon Hughes<sup>1</sup>(✉) , Peter Hastings<sup>1</sup> , and M. Anne Britt<sup>2</sup>

<sup>1</sup> School of Computing, DePaul University, Chicago, USA

[simonhughes22@hotmail.com](mailto:simonhughes22@hotmail.com)

<sup>2</sup> Psychology Department, Northern Illinois University, DeKalb, USA

**Abstract.** Recent educational standards stress that students should learn how to read and understand scientific explanations and create explanations of their own. But these skills are difficult for teachers to evaluate, so they often assess them at a shallow level or avoid giving such assignments. Previous approaches for automatically evaluating explanatory and other types of structured essays have relied on the use of shallow features or bag-of-words methods. These methods might allow for a reasonable holistic assessment of an essay, but they fail to identify which concepts students included and which causal connections they made. In this paper, we investigate which natural language processing methods are most successful at locating conceptual information in student explanations and the causal connections between them. We found that a combination of a recurrent neural network for identifying concepts along with a novel causal relation parser produced very good accuracy in two different scientific domains, significantly improving on the prior state-of-the-art.

## 1 Introduction

The US Common Core standards and the Next-Generation Science Standards reflect an increasing emphasis in education on how important it is for students to learn how to read and comprehend science theories, models, and explanations, integrate information from multiple sources, and to create their own explanations [1, 6]. Teachers often find it challenging to evaluate such texts in more than a cursory manner [13, 22]. Automated Essay Scoring mechanisms could be used to reduce the load on teachers, but they tend to rely on surface-level features of text aggregated across the essay [14] or bag-of-words approaches like LSA [8], correlated with expert scores or pre-scored essays. These approaches are not sophisticated enough to identify the structure of the students' explanations. In other words, they cannot determine which components of an ideal explanation

---

S. Hughes—The assessment project described in this article was funded, in part, by the Institute for Education Sciences, U.S. Department of Education (Grant R305G050091 and Grant R305F100007). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

the students have included, and how they have connected them together. In this paper, we attempt to determine the optimal natural language processing (NLP) techniques for identifying conceptual information and causal relations in explanations, including a novel relation-parsing method.

## 2 Materials

As part of a larger project focused on understanding students' reading processes, approximately 1,300 14–15 year old students in a large U.S. city were asked to read a small set of documents on a particular scientific phenomenon and create their own explanation of that phenomenon. Two different topics were used: skin cancer and coral bleaching. Each student worked with both topics. There were 5 documents of less than 1 page for each topic. One gave a general overview and the others gave related information including images, maps, and charts. With input from topic experts, a causal model was created for each topic, indicating the important concepts described in the documents and the causal connections made between them. The coral bleaching causal model included 13 concepts, and the skin cancer model had 9.

Over 1100 student essays were collected for each topic. The **brat** tool [23, 24] was used to annotate word spans as concepts and explicit connections between them as causal relations. Inter-rater reliability was high, with  $\kappa$  values of 93%. In the next sections, we present the evaluation of several successful NLP techniques for identifying the concepts and causal relations in the essays.

## 3 Concept Identification

The five techniques we compared have each been previously shown to produce state-of-the-art results on various NLP tasks. Each had different representational approaches to handling the challenges of ambiguity in text, interrelationships between words, and relative probabilities of classification. We compared the approaches using micro-averaged  $F_1$  scores, because they capture performance with the relative frequencies of the codes in actual texts. All were tested with 5-fold cross validation

- **Window-based taggers** [21, for example] classify an item using that item and features about its neighboring items as inputs. Previously, we evaluated a window-based method with an SVM classifier, yielding an  $F_1$  score of 0.73 [10]. Here, we extended that approach, finding the best performance by using logistic regression on positional stemmed unigrams, non-positional unigrams, Brown cluster labels [3], and dependency parser relations.
- A **Conditional Random Field** (CRF) [12] learns a graphical model which constitutes a linear chain of probabilities, expressing relationships between random variables [12]. We used the CRFSuite [16] implementation and trained the model with the L-BFGS gradient descent method.

- A **Hidden Markov Model** (HMM) is also a linear chain probabilistic model [18, 25], but it is a *generative* model. It learns to predict the probability of observing a particular *word* based on the label from the training set and the label of the previous word.
- A **Structured Perceptron** was used to perform multi-class classification [5, for example]. Being an online model allows this approach to more easily incorporate its own previous predictions as features to predict the next label in the sentence.
- A **Recurrent Neural Network** learns to build its own representation as it iterates through the words in a sentence [7]. We used the bi-directional Gated Recurrent Unit (GRU) variant of RNN, with 100-dimensional GloVe embeddings [17] as inputs. The best-performing network followed the inputs with two bi-directional GRU layers of 256 units, then a softmax output layer, and it was trained with the Adam optimizer [11].

The performance metrics for the five different concept identification methods on the testset are shown in the top of Table 1. Averaging across topics, the RNN performed best. In comparison with previous results, the average  $F_1$  of 0.84 found here was significantly higher than the 0.73 previously reported.

**Table 1.** Testset accuracy for concept and causal relation identification

	Coral bleaching			Skin cancer		
	Recall	Precision	$F_1$	Recall	Precision	$F_1$
Window-based tagger	0.802	<b>0.885</b>	<b>0.842</b>	0.779	0.853	0.814
CRF	0.797	0.787	0.835	0.759	0.855	0.804
HMM	0.799	0.702	0.747	0.731	0.628	0.675
Structured perceptron	0.794	0.884	0.837	0.773	0.860	0.814
Bi-directional RNN	<b>0.830</b>	0.855	<b>0.842</b>	<b>0.807</b>	<b>0.869</b>	<b>0.837</b>
RNN word tagger	0.656	0.698	0.676	<b>0.798</b>	0.786	<b>0.792</b>
Stacked model	0.674	<b>0.736</b>	0.704	0.719	0.816	0.765
Dependency parser	<b>0.766</b>	0.693	<b>0.728</b>	0.760	<b>0.823</b>	0.790

## 4 Causal Relation Identification

Causal relation identification is a much more challenging task than concept identification because a concept tends to be described by a relatively small set of contiguous words, whereas causal relations are inherently spread across a wider range of words and variety of patterns. Previous work on detecting causal relations in text reflects the difficulty of the problem, either restricting the forms of relations that were considered [2, 9] or achieving rather low performance (e.g.,  $F_1 = 0.41$  [19],  $F_1 = 0.39$  [20]). Our previous work with an SVM classifier

achieved  $F_1 = 0.63$  for the two topics [10], but it was limited to detecting only the presence or absence of *any* causal relation within a sentence. Here, we evaluated three techniques:

- **RNN Word Tagger:** We trained a bi-directional RNN to predict, for each word, the label of the causal connection that it was involved in (if any). The same RNN architecture described above performed best.
- As a **Stacked Model:** [15, for example], we used predictions for all codes in a sentence, and their combinations from the best concept identifier, the RNN, as inputs to a logistic regression classifier, because it is robust to overfitting and can learn from arbitrary input features.
- **Transition-based Dependency Parser:** We developed a novel parsing mechanism which learns to detect causal relations between concept codes predicted by the RNN model. The parsing mechanism was adapted from dependency parsers, such as [4].

The performance of the different causal relation identification techniques on the test sets for both topics is shown in the bottom part of Table 1. The dependency parser produced the top combined performance with an average  $F_1$  score of 0.759, compared to 0.734 for the RNN Word Tagger and 0.735 for the stacked model. The parser's advantages are reflected in the pattern of results. In the coral bleaching topic, students mentioned 85 different relations, compared to 49 relations between the smaller set of concepts in the skin cancer topic. Accordingly, the average number of examples of each causal relation was much higher in the skin cancer topic (20.3 compared to 7.0). The parser learns when it can combine two concept codes into a causal relation instead of treating each relation as a separate label. This allows it to generalize better over all of the relations, as reflected in the higher recall scores for the parser over the other models on the coral bleaching topic. The higher precision for the parser on the skin cancer topic than on coral bleaching can be attributed to the higher number of training examples.

## 5 Conclusions

In this paper, we compared the performance of several highly competitive techniques for identifying explanation structure, including a novel adaptation of a parsing mechanism to the task of causal relation identification. The bi-directional RNN showed the best performance on the concept identification task, achieving an average  $F_1$  score of 0.84, significantly higher than that found in previous research. Although the Word-Tagging RNN achieved slightly higher performance than the Dependency Parser for causal relation identification on the skin cancer topic, overall the parser provided better performance, with an average  $F_1$  of 0.76. Here too, we have achieved a significant increase in accuracy over previous research. This level of performance indicates that these techniques can be confidently used by an intelligent system to give feedback on the concepts and causal structure in students' scientific explanations.

## References

1. Achieve Inc.: Next Generation Science Standards (2013)
2. Blanco, E., Castell, N., Moldovan, D.: Causal relation extraction. In: LREC (2008)
3. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. *Comput. Linguist.* **18**(4), 467–479 (1992)
4. Collins, M.: Head-driven statistical methods for natural language parsing. Unpublished PhD thesis, University of Pennsylvania (1999)
5. Collins, M.: Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 1–8. Association for Computational Linguistics (2002)
6. Council of Chief State School Officers (CCSSO): The Common Core Standards for English Language Arts and Literacy in History/Social Studies and Science and Technical Subjects (2010). <http://www.corestandards.org>
7. Dietterich, T.G.: Machine learning for sequential data: a review. In: Caelli, T., Amin, A., Duin, R.P.W., de Ridder, D., Kamel, M. (eds.) SSPR /SPR 2002. LNCS, vol. 2396, pp. 15–30. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-70659-3\\_2](https://doi.org/10.1007/3-540-70659-3_2)
8. Foltz, P.: Latent semantic analysis for text-based research. *Behav. Res. Methods Instrum. Comput.* **28**, 197–202 (1996)
9. Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., Yuret, D.: SemEval-2007 task 04: classification of semantic relations between nominals. In: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 13–18 (2007). <http://acl.ldc.upenn.edu/W/W07/W07-2003.pdf>
10. Hughes, S., Hastings, P., Britt, M.A., Wallace, P., Blaum, D.: Machine learning for holistic evaluation of scientific essays. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 165–175. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_17](https://doi.org/10.1007/978-3-319-19773-9_17)
11. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning 2001, ICML 2001, pp. 282–289 (2001)
13. Magliano, J.P., Graesser, A.C.: Computer-based assessment of student-constructed responses. *Behav. Res. Methods* **44**(3), 608–621 (2012)
14. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
15. Menahem, E., Rokach, L., Elovici, Y.: Troika—an improved stacking schema for classification tasks. *Inf. Sci.* **179**(24), 4097–4122 (2009)
16. Okazaki, N.: CRFsuite: a fast implementation of conditional random fields (CRFs) (2007). <http://www.chokkan.org/software/crfsuite/>
17. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. *EMNLP.* **14**, 1532–1543 (2014)
18. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1989)
19. Riaz, M., Girju, R.: Recognizing causality in verb-noun pairs via noun and verb semantics. *EACL* **2014**, 48 (2014)

20. Rink, B., Bejan, C.A., Harabagiu, S.M.: Learning textual graph patterns to detect causal event relations. In: Guesgen, H.W., Murray, R.C. (eds.) FLAIRS Conference. AAAI Press (2010)
21. Sánchez-Villamil, E., Forcada, M.L., Carrasco, R.C.: Unsupervised training of a finite-state sliding-window part-of-speech tagger. In: Vicedo, J.L., Martínez-Barco, P., Muñoz, R., Saiz Noeda, M. (eds.) EsTAL 2004. LNCS (LNAI), vol. 3230, pp. 454–463. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30228-5\\_40](https://doi.org/10.1007/978-3-540-30228-5_40)
22. Shermis, M.D., Burstein, J.: Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge, Abingdon (2013)
23. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Association for Computational Linguistics, Avignon, France, April 2012. <http://brat.nlplab.org>
24. Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.D., Tsujii, J.: BioNLP shared task 2011: supporting resources. In: Proceedings of BioNLP Shared Task 2011 Workshop, pp. 112–120. Association for Computational Linguistics, Portland, Oregon, USA, June 2011. <http://www.aclweb.org/anthology/W11-1816>
25. Welch, L.R.: Hidden markov models and the Baum-Welch algorithm. IEEE Inf. Theor. Soc. Newslett. **53**(4), 10–13 (2003)



# A Systematic Approach for Analyzing Students' Computational Modeling Processes in C2STEM

Nicole Hutchins<sup>1</sup>✉, Gautam Biswas<sup>1</sup>, Shuchi Grover<sup>2</sup>,  
Satabdi Basu<sup>3</sup>, and Caitlin Snyder<sup>1</sup>

<sup>1</sup> Vanderbilt University, Nashville, TN 37212, USA

nicole.m.hutchins@vanderbilt.edu

<sup>2</sup> Looking Glass Ventures, Palo Alto, CA 94306, USA

<sup>3</sup> SRI International, Menlo Park, CA 94025, USA

**Abstract.** Introducing computational modeling into STEM classrooms can provide opportunities for the simultaneous learning of computational thinking (CT) and STEM. This paper describes the C2STEM modeling environment for learning physics, and the processes students can apply to their learning and modeling tasks. We use an unsupervised learning method to characterize student learning behaviors and how these behaviors relate to learning gains in STEM and CT.

**Keywords:** Learning by modeling · Computational model-building · STEM+CT

## 1 Introduction

Modeling is fundamental to science. The Next Generation Science Standards (NGSS) [12] have reinforced the importance of model-based STEM learning to engage students in authentic STEM practices. Our Collaborative Computational STEM (C2STEM) [8] learning environment provides opportunities for students to construct computational models in STEM domains (e.g., [5, 16, 17]) and use these models for problem solving [1, 18]. Such “constructionist” approaches have helped students learn STEM and computational thinking (CT) concepts and practices [3, 16], but some students face difficulties in translating STEM knowledge into computational models [5]. Therefore, students’ learning and model building processes merit further investigation.

This paper adopts an exploratory approach to characterize students’ learning and model building processes in C2STEM. We apply hierarchical clustering on students’ activity data to address the research questions: (1) What patterns of behavior do learners exhibit during computational modeling tasks in a science domain? and (2) What can we glean from these patterns about student learning of science and CT?

## 2 Background

Our learning-by-modeling paradigm helps students learn by developing, testing, and refining computational models. Such modeling environments provide mechanisms for students to work with multiple representations, receive rapid feedback through the visualization of model behaviors [5, 11], and engage in CT practices [18]. Classroom studies conducted with systems such as CTSiM [3], ViMap [14] and CT-STEM [10] have produced successful summative learning results [5, 15, 18]. We aim to extend this work by analyzing students' model building processes, including impact on learning gains.

Early efforts in the analysis of log data from students' programming process focused on methods to quantify students' modeling progress at each model revision by calculating the distance between the student and expert model [1]; identify program states and assess the likelihood of reaching a "sink" state in which a student was likely to get stuck [5]; and apply exploratory data-driven approaches to design partial solution feedback [13]. In this work, we used unsupervised learning to closely examine the processes students used towards mutually supportive learning of physics and CT, and made attempts to relate their learning performance to groups of student behaviors (e.g., [4, 19]).

## 3 The C2STEM Environment

C2STEM scaffolds students' model-building by creating a block-based DSML [8] that provides domain-relevant variables (e.g., acceleration and velocity), and explicit constructs (blocks) for initializing and updating the values of these variables (see Fig. 1). This supports exploratory learning by allowing students to execute their developing models and observe the behaviors generated using animations and data tools [8]. While an initialization block (e.g., *green\_flag*) is common across block-based environments, we provide additional scaffolding by explicitly providing a *simulation\_step* block to help students separate initialization steps from the dynamic update step. In contrast to equation-based modeling, this sets up a temporal *step-by-step* approach to modeling to gain a better understanding of how the behavior of a system evolves over time.

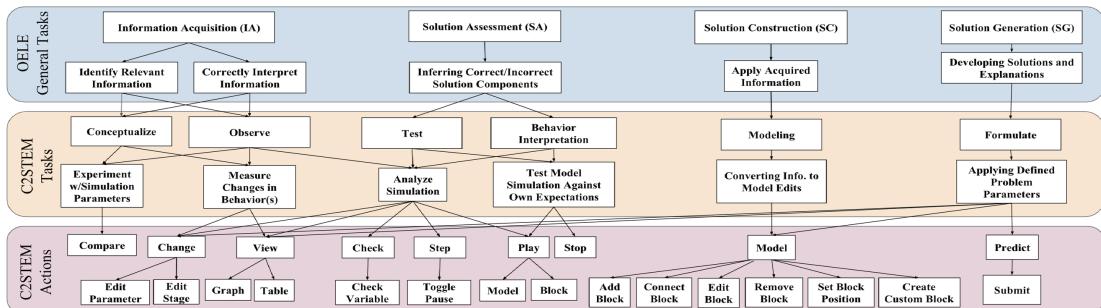


**Fig. 1.** A completed C2STEM model incorporating DSML blocks. (Color figure online)

## 4 Methods

Thirty-five middle school students worked on a 1D motion module in C2STEM that consisted of a training unit and 4 modeling tasks. We used a summative assessment adapted from other studies to measure disciplinary knowledge in physics [2, 7] and CT [1, 6]. Normalized learning gains calculated using  $\frac{\text{Posttest} - \text{Pretest}}{\text{Max Possible Score} - \text{Pretest}}$ .

We performed cluster analysis to characterize students' model building behaviors based on actions employed on a constant velocity task (Fig. 1). We analyzed data for 29 students, excluding data from students who did not complete either the pre-test or post-test or performed less than five actions. Student actions were recorded in log files with timestamps. We extend a task model developed in our previous work [1] (Fig. 2) to interpret students' model building actions. The lowest level captures the discrete model building actions possible, the middle associates a specific purpose for the actions as C2STEM subgoals and the top provides more generic labels to the actions, typically useful for understanding student behaviors across multiple learning environments.



**Fig. 2.** C2STEM task model.

The following features helped cluster students by their model-building activities:

1. *Ratio of total simulations runs to total actions performed (RTP)*: Frequency of SA operations performed.
2. *Ratio of data tool access to total number of simulation runs (RDT)*: Frequency of IA and/or SA operations performed.
3. *Average time per access of data tools (TDT)*: IA/SA related actions.
4. *Average number of actions between simulation runs (ABP)*: Average size of SC tasks; actions between plays imply a construction process influenced by debugging.
5. *Number of blocks Under Green Flag (NBF)*: a SC task related to variable initialization demonstrating conceptual understanding of problem domain.
6. *Number of blocks in simulation step construct (NST)*: Updating functions (in SC).

We derived a dendrogram structure using the UPGMA hierarchical clustering scheme [9]. The maximum distance between levels heuristic was used to determine the cut-off level and the number of clusters formed. The groups were characterized by distinguishing features, which were then used to explain groups' pre-post learning gains (Table 1).

**Table 1.** Characterizing clusters based on frequency (*mean, sd*) of features.

GR	RTP (SA)	ABP (SA/SC)	RDT (IA/SA)	TDT (IA/SA)	NBG (SC)	NST (SC)
1	0.29 (0.09)	1.96 (0.5)	0.05 (0.05)	4,946.1 (6070.7)	4.25 (0.96)	0.25 (0.5)
2	0.24 (0.03)	2.55 (0.52)	0.24 (0.06)	17,007.4 (20,436.7)	3 (0)	7 (2)
3	0.31 (0.12)	2.14 (0.0)	0.05 (0.05)	21,822.7 (30,106.5)	3.75 (0.34)	8 (0)

## 5 Results

Summative assessment results showed that normalized learning gains were *statistically significant*, with t-tests in Physics ( $p = 0.009$ ) and CT ( $p = 0.0001$ ). Cluster analysis produced three distinct groups. Group 1 achieved the highest learning gains in CT [0.50 (0.19)] and moderate Physics learning gains [0.21 (0.23)]. Group 1 is defined by their minimal use of the data tools, highest number of initialization blocks, and very little in terms of update actions to generate dynamic behavior (SC actions). This group also had the least amount of actions between plays (ABP) and the second highest ratio of total plays to total actions (RTP). This may indicate their reliance on trial and error. Given the significant CT learning gains and trial and error approach, we conjecture that these results suggest a focus on programming.

Group 2 achieved the highest learning gains in Physics [0.31 (0.17)] and lowest CT gains [0.22 (0.25)]. Group 2 used data tools (RDT) the most, and had the second largest time usage (TDT). The group had few initialization blocks, forgetting to initialize the simulation step size block [*set delta t to [n] seconds*]. This may have impacted their ability to interpret results from the data tools (for instance, setting delta-t to 1 s would have resulted in variable values updating as integers). Finally, this group had the highest ABP and lower RTP indicating the least amount of testing, implying possible weakness in CT practices such as debugging (as indicated by their low CT gains).

A review of the clustering dendrogram indicates that at the next largest distance, Group 3 breaks into 1 outlier and two subgroups. Subgroup 1 showed higher physics gains [0.23 (2.56)], but lower CT gains of [0.37 (0.36)] (markedly higher than Group 2). All students in this group utilized the data tools, with the highest average TDT and implemented the highest, indicating a more systematic debugging process. Subgroup 2 demonstrated moderate Physics gains, 0.21 (0.16) and higher CT gains, 0.50 (0.16). Their feature values indicate a similar trial and error approaches to Group 1, with low ABP and high RTP, but differences in SC actions may provide useful information into how this approach may impact Physics learning.

## 6 Discussion and Conclusions

This paper presents initial analyses in linking students' model building behaviors to their pre-post assessment scores. High performers showed better ability to model the update functions. Although exploratory, this work provides unique insights and approaches to the evaluation of block-based computational model building processes in STEM classrooms. As next steps, we are continuing our pattern analysis with larger student populations across different science topics. In addition, we are building more sophisticated logging mechanisms to better understand synergistic learning processes and design adaptive feedback to help students overcome their conceptual difficulties.

**Acknowledgments.** We thank Marian Rushdy, Naveed Mohammed, and our other collaborators at Vanderbilt University, Stanford University, Salem State University, SRI International, and ETS. This research is supported by NSF grant #1640199.

## References

1. Basu, S., Biswas, G., Kinnebrew, J.S.: Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Model. User-Adap. Inter.* **27**(1), 5–53 (2017)
2. Basu, S., McElhaney, K., Grover, S., Harris, C., Biswas, G.: A principled approach to designing assessments that integrate science and computational thinking. In: Proceedings of ICLS 2018 (2018)
3. Basu, S., Dickes, A., Kinnebrew, J.S., Sengupta, P., Biswas, G.: CTSiM: A computational thinking environment for learning science through simulation and modeling. In: Conference on Computer Supported Education, pp. 369–378, Germany (2013)
4. Berland, M., Martin, T., Benton, T., Smith, C.P., Davis, D.: Using learning analytics to understand the learning pathways of novice programmers. *J. Learn. Sci.* **22**(4), 564–599 (2013)
5. Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., Koller, D.: Programming pluralism: using learning analytics to detect patterns in the learning of computer programming. *J. Learn. Sci.* **23**(4), 561–599 (2014)
6. Grover, S., Jackiw, N., Lundh, P.: Concepts before coding: non-programming interactives to advance learning of introductory programming concepts in middle school. *Comput. Sci. Educ.* (2019). <https://doi.org/10.1080/08993408.2019.1568955>
7. Hestenes, D., Wells, M., Swackhamer, G.: Force concept inventory. *Phys. Teach.* **30**(3), 141–158 (1992)
8. Hutchins, N., Biswas, G., Maroti, M., Broll, B., Ledezci, A.: C2STEM: a design-based approach to a classroom-centered OELE. In: Proceedings of AIED 2018 (2018)
9. Johnson, S.C.: Hierarchical clustering schemes. *Psychometrika* **32**(3), 241–254 (1967)
10. Jona, K., Wilensky, U., Trouille, L., Horn, M. S., Orton, K., Weintrop, D., Beheshti, E.: Embedding computational thinking in science, technology, engineering, and math (CT-STEM). In: Future Directions in Computer Science Education Summit Meeting, Orlando, FL (2014)
11. Jonassen, D., Strobel, J., Gottdenker, J.: Model building for conceptual change. *Interact. Learn. Environ.* **13**(1–2), 15–37 (2005)

12. NGSS Lead States: Next Generation Science Standards: For states, by states. National Academies Press, Washington, DC (2013)
13. Piech, C., Huang, J., Nguyen, A., Phulsuksombati, M., Sahami, M., Guibas, L.: Learning program embeddings to propagate feedback on student code. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France pp. 1093–1102 (2015)
14. Sengupta, P., Dickes, A., Farris, A.V., Karan, A., Martin, D., Wright, M.: Programming in K-12 science classrooms. *Commun. ACM* **58**(11), 33–35 (2015)
15. Sengupta, P., Farris, A.V., Wright, M.: From agents to continuous change via aesthetics: learning mechanics with visual agent-based computational modeling. *Technol. Knowl. Learn.* **17**(1–2), 23–42 (2012)
16. Sengupta, P., Kinnebrew, J.S., Basu, S., Biswas, G., Clark, D.: Integrating computational thinking with K-12 science education using agent-based computation: a theoretical framework. *Educ. Inf. Technol.* **18**(2), 351–380 (2013)
17. Shen, J., Lei, J., Chang, H.-Y., Namdar, B.: Technology-enhanced, modeling-based instruction (TMBI) in science education. In: Spector, J.M., Merrill, M.D., Elen, J., Bishop, M.J. (eds.) *Handbook of Research on Educational Communications and Technology*, pp. 529–540. Springer, New York (2014). [https://doi.org/10.1007/978-1-4614-3185-5\\_41](https://doi.org/10.1007/978-1-4614-3185-5_41)
18. Weintrop, D., et al.: Defining computational thinking for mathematics and science classrooms. *J. Sci. Educ. Technol.* **25**(1), 127–147 (2016)
19. Werner, L., McDowell, C., Denner, J.: A first step in learning analytics: pre-processing low-level Alice logging data of middle school students. *J. Educ. Data Min.* **5**(2), 11–37 (2013)



# Intelligent Tutoring System for Negotiation Skills Training

Emmanuel Johnson<sup>(✉)</sup>, Gale Lucas, Peter Kim, and Jonathan Gratch

University of Southern California, Los Angeles, CA 90007, USA  
e.johnson@ict.usc.edu

**Abstract.** Intelligent tutoring systems have proven very effective at teaching hard skills such as math and science, but less research has examined how to teach “soft” skills such as negotiation. In this paper, we introduce an effective approach to teaching negotiation tactics. Prior work showed that students can improve through practice with intelligent negotiation agents. We extend this work by proposing general methods of assessment and feedback that could be applied to a variety of such agents. We evaluate these techniques through a human subject study. Our study demonstrates that personalized feedback improves students’ use of several foundational tactics.

**Keywords:** Negotiation training · Individualized feedback · Soft skills training

## 1 Introduction

Research in the intelligent tutoring systems community has shown that these systems are effective at teaching hard skills including math [1–3], reading [4, 5], even computer literacy [6, 7]. Research has tried to extend these techniques to softer skills such as public speaking [8], collaborative problem solving [9] and more specifically negotiation [10–12]. Compared to the application of intelligent tutoring systems to hard skills training, the systems designed to teach softer skills are limited.

In this article, we tackle the domain of negotiation. Like most social skills, negotiation falls within what Aleven and colleagues [13] define as an ill-defined domain, and presents a challenge for intelligent tutors. Social skills lack clear assessment metrics and prescribed formulas to guarantee success. Though hard to teach, social skills are becoming increasingly crucial for students entering the modern workforce. The US Academy of Sciences and the World Economic Forum identify negotiation as a foundational social skill essential for the future of work through its impact on organizational creativity and productivity [14, 15]. Deficits in negotiation ability contribute to the underrepresentation and lack of advancement of women and minorities in STEM fields [16, 17]. Unfortunately, negotiation training is inaccessible to most workers who need it (e.g., even a short 5-day seminar can cost more than \$10,000 per student).

Yet there is reason for optimism. Most recently, researchers have shown that students who practice negotiating with intelligent agents can improve their skills [11, 18]. Although these systems have been shown to improve negotiation skills, with some exceptions [12], they mainly allow users to practice and do not provide feedback.

Feedback is one of the most crucial aspects of the learning process [19]. In this paper, we illustrate how to build upon general intelligent agent technology to provide both experiential practice and personalized feedback. We introduce a general (domain- and algorithm-independent) approach to incorporate automatic assessment and personalized feedback into intelligent negotiation agent technology and describe a study to assess the effectiveness of our approach. In Sect. 2, we discuss our method for automatically assessing students’ ability to create and claim value. In Sect. 3, we incorporate these metrics into a publicly-available online negotiation platform called IAGO [20], and present experiments that assess the benefits of “mere practice” with this system compared with practice coupled with either generic or personalized feedback. In Sect. 4, we discuss our results and several lessons on how to improve these techniques in future research.

## 2 Automated Assessment and Feedback

The ultimate goal of negotiation is to obtain good outcomes (i.e., maximize the value of the negotiated agreement, establish a fair and positive reputation, etc.). However, most negotiation training addresses tactics to achieve these ends. Thus, an intelligent negotiation tutor must assess outcomes, and the means by which students achieve them. Here, we review the assessments used by negotiation instructors and show how to automatically make these assessments and provide personalized feedback.

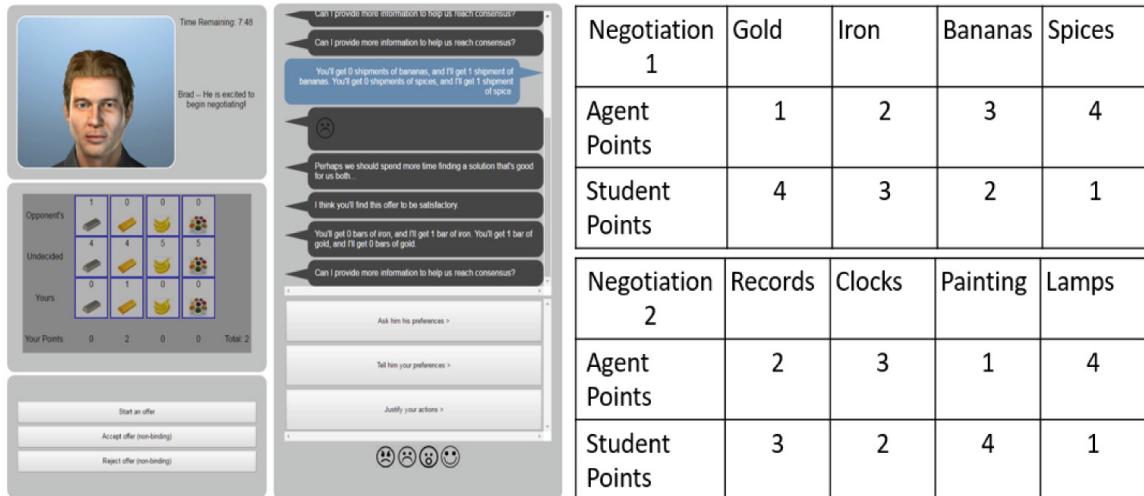
We adopt a set of general assessment metrics that address specific tactics for creating and claiming value. In our previous work we highlight these metrics and show how they are automatically calculated [21]. We assess a student’s ability to create value by measuring the joint points achieved in the negotiated agreement (i.e., the points obtained by both the student and the agent). We evaluate several process measures to gain insight into why a student may have failed to create value. For example, we assess if a student employed the tactic of logrolling by the extent to which they made tradeoffs in their initial offer to the agent (specifically, the number of highest-value items they claimed minus the number of lowest-value items they offered). We also evaluate a student’s ability to claim value by measuring the individual points they obtained in the final deal. We assess one process measure to gain insight into why they may have failed to claim value. Specifically, we look at the point value of the student’s initial offer.

After completing a simulated negotiation, students are assessed using the above-mentioned metrics and then receive automatically-generated feedback. The generated feedback describes the extent of good outcomes achieved, and how they followed specific strategies to achieve these outcomes (e.g., did they exchange information with their opponent? Did they make ambitious offers?). They are then provided specific actionable strategies for improving in the future. When students achieve good outcomes or follow recommended tactics, this is positively reinforced (e.g., “The first offer you made would have gotten you about 76% of the points. Pretty good.”) and the principle emphasized (“By claiming most of what you want early in the negotiation, you can manage your negotiation partner’s expectations of what they will receive.”). When students fail, it is highlighted (e.g., You failed to fully understand your opponent’s preferences. This prevented you from making good tradeoffs”) and specific suggestion is

provided (e.g., “For example, if you realized your opponent wanted bananas the most, a win-win solution would be giving them all bananas and taking all the gold for yourself.”).

### 3 Evaluating the Effectiveness of Personalized Feedback

*Negotiation Task:* Participants were asked to engage in two negotiations using the IAGO online negotiation platform [22]. IAGO is designed to support tactics that expert negotiators used to create and claim value. Negotiators can exchange offers but also information (do you like A more than B?) and send other messages such as threats. The platform also provides tools to customize agent behavior including the ability to incorporate common biases shown by negotiators (such as the fixed-pie bias). It has been used by a number of researchers to build human like negotiating agents [23]. Each negotiation has the same mathematical structure (a 4-issue, 6-level multi-issue bargaining task) but used a different cover story and a different ordering of the issues to obscure this similarity. The tasks were framed as a negotiation between antique dealers on dividing the contents of an abandoned storage locker. Both the agents and participant had distinct preferences across the items, and neither knew the other’s preference. Figure 1 shows the number of points each party could get for each item.



**Fig. 1.** The left image illustrates the IAGO agent interface. The tables on the right illustrate the issues and payoffs for the two negotiations

*Measures:* We gathered basic demographic information and self-reported negotiation skill level prior to the negotiation. During the negotiation, we automatically derived the metrics discussed in Sect. 2.

*Participants:* 120 English speaking America participants who were recruited via Mechanical Turk. To motivate their performance, participants were paid \$3/hour for their participation in the study and entered into a lottery to win a prize of \$10. Of these participants, 19 were excluded from analysis (9 failed the attention check and 10 failed to reach an agreement or experience software failure).

*Experimental Manipulation:* Participants were randomly assigned to one of three experimental conditions, Personalized Feedback, Generic Feedback or No Feedback. Participants in the Personalized Feedback condition were provided personalized feedback on their initial claim, understanding of their opponent's preferences and the overall value of their final claim using the methods described in Sect. 2. Those in the Generic Feedback condition received feedback on the same metrics as the personalized feedback condition except that it was based on a hypothetical negotiation. For example, they are provided suggestions on how good that person did and how their results could have been improved. Those in the No Feedback condition were told the points they received but provided no other information.

## 4 Results and Discussion

We evaluated the effects of practice and feedback with a 3 (feedback: none v. generic v. personalized)  $\times$  2 (time: negotiation 1 v. negotiation 2) mixed ANOVA. For value claiming, students benefited from practice alone and this benefit was enhanced by feedback (both in tactics and final outcome). Students made stronger initial claims on the second negotiation ( $F(1, 98) = 33.47, p < .001$ ) than the first, and the interaction with the type of feedback nearly reached significance ( $F(2, 98) = 3.01, p = .054$ ). Participants who received feedback (either personalized or generic) claimed more value. In terms of final outcome, we see a significant main effect of time ( $F(1, 98) = 30.40, p < .001$ ) and a significant interaction with the type of feedback ( $F(2, 98) = 3.808, p = .026$ ). Participants obtained more points in the second negotiation and those who received personalized feedback gained the most points. For creating value, we found a clear benefit of practice and a strong effect of feedback for logrolling and joint points but not for questions asked. Concerning the final outcome, we find a significant benefit of practice on joint points as they created more value in the second negotiation than the first ( $F(1, 98) = 7.322, p = .008$ ). Personalized feedback yielded the highest joint points, the interaction was significant ( $F(2, 98) = 8.187, p = .001$ ). Students engaged in logrolling more with practice ( $F(1, 98) = 37.495, p < .001$ ) and there was a significant interaction with condition such that this improvement in logrolling from the first negotiation to the second was strengthened by personalized feedback ( $F(2, 98) = 4.930, p = .009$ ). Students asked more questions with practice ( $F(1, 98) = 24.461, p < .001$ ) and asked the most with personalized feedback, though the interaction with condition was not significant ( $F(2, 98) = 1.711, p = .186$ ).

We show students improve in their use of both value-claiming tactics through a combination of practice and personalized feedback. Personalized feedback further increased learning by helping students to make more ambitious offers and use log-rolling. Although this work is promising, our ultimate goal is to show that the benefits accrued through such automated practice, assessment and feedback will generalize outside these simulations. Future planned studies will examine if students improve in both computer-mediated and face-to-face negotiations with other students.

## References

1. Koedinger, K.R.: Cognitive tutors. In: The Cambridge Handbook of the Learning Sciences, pp. 61–78 (2005)
2. Koedinger, K.R., Corbett, A., et al.: Cognitive tutors: technology bringing learning sciences to the classroom, na (2006)
3. Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A.: Intelligent tutoring goes to school in the big city. *Int. J. Artif. Intell. Educ. (IJAIED)* **8**, 30–43 (1997)
4. Mills-Tettey, G.A., et al.: Improving child literacy in Africa: experiments with an automated reading tutor. In: 2009 International Conference on Information and Communication Technologies and Development (ICTD) (2009)
5. Wijekumar, K., Meyer, B., Spielvogel, J.: Web-based intelligent tutoring to improve reading comprehension in elementary and middle schools: design, research, and preliminary findings. In: E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (2005)
6. Olney, A., Bakhtiari, D., Greenberg, D., Graesser, A.C.: Assessing computer literacy of adults with low literacy skills. In: EDM (2017)
7. Guo, P.J.: Codeopticon: real-time, one-to-many human tutoring for computer programming. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (2015)
8. Chollet, M., Sratou, G., Shapiro, A., Morency, L.-P., Scherer, S.: An interactive virtual audience platform for public speaking training. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (2014)
9. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the science of collaborative problem solving. *Psychol. Sci. Public Interest* **19**, 59–92 (2018)
10. Monahan, S., Johnson, E., Lucas, G., Finch, J., Gratch, J.: Autonomous agent that provides automated feedback improves negotiation skills. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 225–229. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_41](https://doi.org/10.1007/978-3-319-93846-2_41)
11. Gratch, J., DeVault, D., Lucas, G.: The benefits of virtual humans for teaching negotiation. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) IVA 2016. LNCS (LNAI), vol. 10011, pp. 283–294. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-47665-0\\_25](https://doi.org/10.1007/978-3-319-47665-0_25)
12. Kim, J.M., et al.: BiLAT: a game-based environment for practicing negotiation in a cultural context. *Int. J. Artif. Intell. Educ.* **19**, 289–308 (2009)
13. Aleven, V., Ashley, K., Lynch, C., Pinkwart, N.: Intelligent tutoring systems for ill-defined domains: assessment and feedback in ill-defined domains. In: The 9th International Conference on Intelligent Tutoring Systems (2008)
14. National Academies of Sciences Engineering and Medicine, Promising Practices for Strengthening the Regional STEM Workforce Development Ecosystem. National Academies Press (2016)
15. Forum, W.E.: The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution. World Economic Forum, Geneva (2016)
16. Hernandez, M., Avery, D.R.: Getting the short end of the stick: racial bias in salary negotiations. *MIT Sloan Manag. Rev.* (2016)
17. Goldman, E.G.: Lipstick and labcoats: undergraduate women's gender negotiation in STEM fields. *NASPA J. Women High. Educ.* **5**, 115–140 (2012)

18. Lin, R., Oshrat, Y., Kraus, S.: Investigating the benefits of automated negotiations in enhancing people's negotiation skills. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, vol. 1 (2009)
19. Kolb, A.Y., Kolb, D.A.: Experiential learning theory. In: Seel, N.M. (ed.) Encyclopedia of the Sciences of Learning, pp. 1215–1219. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-1-4419-1428-6\\_227](https://doi.org/10.1007/978-1-4419-1428-6_227)
20. Mell, J., Gratch, J.: IAGO: interactive arbitration guide online. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (2016)
21. Johnson, E., DeVault, D., Gratch, J.: Towards an autonomous agent that provides automated feedback on students' negotiation skills. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems (2017)
22. Mell, J., Gratch, J.: Grumpy and Pinocchio: answering human-agent negotiation questions through realistic agent design. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, Richland (2017)
23. Mell, J., Gratch, J., Baarslag, T., Aydogan, R., Jonker, C.M.: Results of the first annual human-agent league of the automated negotiating agents competition. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, New York, NY, USA (2018)



# Robot Lecture for Enhancing Non-verbal Behavior in Lecture

Akihiro Kashihara<sup>1</sup>(✉), Tatsuya Ishino<sup>1</sup>, and Mitsuhiro Goto<sup>2</sup>

<sup>1</sup> The University of Electro-Communications, Chofu, Tokyo 182-8585, Japan  
akihiro.kashihara@inf.uec.ac.jp

<sup>2</sup> NTT Service Evolution Laboratories, Yokosuka, Kanagawa 239-0087, Japan

**Abstract.** In lecture with presentation slides such as e-Learning lecture on video, it is important for lecturers to control their non-verbal behavior involving gaze, gesture, and paralanguage to attract learners' attention to slide or oral contents they intend to emphasize. However, it is not so easy even for well-experienced lecturers to properly use and maintain non-verbal behavior in their lecture to promote learners' interest and understanding. This paper proposes robot lecture, in which a communication robot substitutes for human lecturers, and reconstructs their non-verbal behavior to enhance their lecture. Towards such reconstruction, we have designed a model of non-verbal behavior in lecture, and developed a robot lecture system, which follows the model to detect and reconstruct insufficient/inappropriate behavior, and which conducts the reconstructed lecture. This paper describes a case study with the system, whose purpose was to ascertain the benefits of robot lecture by comparing video lecture conducted by human, robot lecture simply reproducing the original one, and robot lecture reconstructing the original one. The results suggest that the robot lecture involving reconstruction promotes learners' understanding of the lecture slides more than the video lecture and the robot lecture involving simple reproduction.

**Keywords:** Robot lecture · Lecture enhancement · Non-verbal behavior

## 1 Introduction

In lecture, it is important to present the lecture contents as lecture slides with oral explanation so that learners' understanding could be promoted. This requires lecturers to consider not only what to present but also how to present. In particular, they need to control the attention of learners using gaze, gesture, paralanguage, etc., which are viewed as non-verbal behavior [1]. If lectures want to induce learners to pay more attention to an important point in a slide, for example, they should direct their face to it, and point it out by pointing gesture in concurrence with its oral explanation. On the other hand, conducting non-verbal behavior excessively and unnecessarily would prevent learners from keeping attention to understand the lecture contents. It is accordingly indispensable to properly use non-verbal behavior in lecture [2–4] (called lecture behavior in this paper). However, it is not so easy even for well-experienced lecturers.

In this paper, we propose robot lecture, in which a robot substitutes for human lecturers. The main purpose of robot lecture is to reproduce their own unique behavior

as much as possible with their lecture contents, and to reconstruct their improper and insufficient behavior for enhancing their lecture. Toward the lecture behavior reconstruction, we have designed a model of lecture behavior, which shows how lecturers should conduct lecture behavior to promote learners' interest and understanding [5].

We have been also developing a robot lecture system, which deals with gaze, face direction, gesture, and paralanguage as lecture behavior [5]. This system records the lecture given by human to detect and reconstruct inappropriate or insufficient behavior by following the model. The robot then reproduces the reconstructed behavior.

This paper describes a case study with the system whose purpose was to compare video lecture by human, robot lecture simply reproducing the original one, and robot lecture reconstructing the original one. The results suggest that the robot lecture involving reconstruction significantly promotes learners' understanding of the lecture slides.

## 2 Robot Lecture

Related work on using non-verbal behavior in interaction between human and robot suggest non-verbal behavior conducted by robot could control the attention of human [6–9]. Following the findings, we consider how robot should conduct lecture behavior according to lecturers' intention to control the attention of learners.

In considering robot lecture, we have designed a model of lecture behavior by referring to related work on nonverbal behavior [6–8]. Lecturers should use non-verbal behavior according to their intention in lecture, which could be determined with learning states of learners. We divide the states into four, which are “not listening (state 1)”, “listening (state 2)”, “noticing (state 3)”, and “understanding (state 4)” of the lecture contents. Lecturers should intend to change learning states from state 1 to 4. We accordingly define lecture intention as changing learning states, and classify it into *Intention 1* (from *state 1* to *2*), *Intention 2* (from *state 2* to *3*), and *Intention 3* (from *state 3* to *4*). The model of lecture behavior represents nonverbal behavior appropriate to each lecture intention (See [5] in detail). The model is composed of three layers, which are lecture intention, behavior category, and basic components of behavior. It derives lecture behavior from the relationships among them.

In order to follow the model to reconstruct non-verbal behavior conducted by human, we have developed the robot lecture system. In this system, we currently assume video lecture in which the lecturer has learners in the learning state 2 with the intention 2. The learning state and intention are also supposed to be unchanged during lecture. In addition, the system can currently reconstruct gaze, gesture, and pitch/volume of paralanguage as non-verbal behavior. It also uses Sota as robot [9].

## 3 Case Study

The purpose of the case study was to ascertain whether robot lecture with reconstruction could be more beneficial for understanding the lecture contents than robot lecture with simple reproduction and video lecture by human. Comparing the robot

lecture with reconstruction and the one with simple reproduction, we can ascertain the validity of reconstruction with the lecture behavior model.

Participants were 22 graduate and undergraduate students. We prepared three video lectures whose topics were learning model, social learning, and learning technology, which were recorded from lectures conducted by the same lecturer who was one of the authors. These lectures included 11 or 12 slides, and were about 5 to 6 min. We also prepared three robot lectures that reconstructed the corresponding lectures with the lecture behavior model, and three robot lectures that simply reproduced the corresponding lectures without reconstruction.

Group	Lecture topic		
	Learning model	Social learning	Learning technology
Group 1 (3 participants)	Robot-lecture with reconstruction	Robot-lecture with simple reproduction	Video lecture
Group 2 (4 participants)	Robot-lecture with reconstruction	Video lecture	Robot-lecture with simple reproduction
Group 3 (4 participants)	Video lecture	Robot-lecture with reconstruction	Robot-lecture with simple reproduction
Group 4 (4 participants)	Robot-lecture with simple reproduction	Robot-lecture with reconstruction	Video lecture
Group 5 (4 participants)	Video lecture	Robot-lecture with simple reproduction	Robot-lecture with reconstruction
Group 6 (3 participants)	Robot-lecture with simple reproduction	Video lecture	Robot-lecture with reconstruction

**Fig. 1.** Procedure for taking lectures.

We set three conditions: (a) Taking video lecture (Video condition), (b) Taking robot lecture involving reconstruction (Robot-Reconstruction condition), and (c) Taking robot lecture involving simple reproduction (Robot-Reproduction condition). As within-subject design, each participant took the three lectures under these three conditions. In order to counterbalance the order effects of the conditions, as shown in Fig. 1, we randomly assigned 22 participants to six groups.

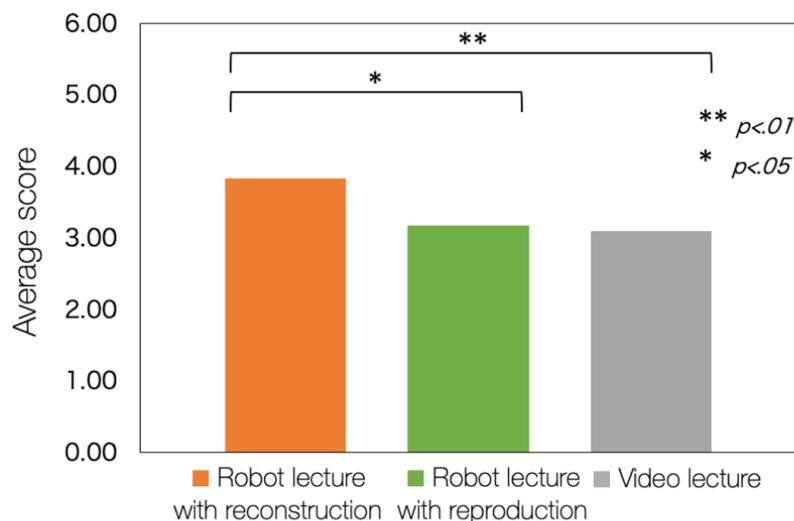
After taking each lecture, the participants were required to have an understanding test including 3 in-slide questions and 3 between-slides questions. The in-slide questions asked the contents within slide. The between-slides questions asked the relations between the contents of two slides. Each question was scored one point (The perfect score was 6 points). The hypotheses we set up in this study were as follows:

H1 (H2): Robot lecture involving reconstruction promotes understanding of the slide contents more than robot lecture involving simple reproduction (more than lecture video).

Figure 2 shows the average scores of the understanding tests in each condition. From the one-sided t-test, there were significant differences between Robot-Reconstruction and Video conditions ( $t(21) = 2.59$ ,  $p < .01$ ), and between Robot-Reconstruction and Robot-Reproduction conditions ( $t(21) = 1.93$ ,  $p < .05$ ). As for the

average scores of understanding test in each condition, there was a marginally significant difference between Robot-Reconstruction and Robot-Reproduction conditions ( $t(21) = 1.64, p < .10$ ) in in-slide questions. As for between-slides questions, there was a significant difference between Robot-Reconstruction and Video conditions ( $t(21) = 3.11, p < .01$ ).

From these results, the robot lecture involving reconstruction promotes understanding of the slide contents more than the video lecture and the robot lecture involving simple reproduction, which supports H1 and H2. The results also suggest the necessity and importance of reconstructing lecture behavior with robot. In addition, the results of the understanding test suggest that the lecture behavior reconstruction promoted understanding of the contents within slides rather than the relation between the slides. This seems reasonable since the current system mainly deals with lecture behavior for presenting the lecture contents in every slide, not with the one for presenting the relations embedded in a number of slides. The results also indicate the validity of the lecture behavior model within non-verbal behavior for keeping/controlling attention and promoting understanding of important points.



**Fig. 2.** Average scores for understanding test.

## 4 Conclusion

In this paper, we have proposed robot lecture, in which a communication robot substitutes for human lecturers by means of reconstructing their lecture behavior. We have also reported a case study with the system whose results suggest reconstructing lecture behavior contributes to promotion of understanding the lecture contents more than simple reproduction of lecture behavior by robot and video lecture by human.

In future, we attempt to detect the learning states dynamically changed during lecture to reconstruct lecture contents and behavior as interactive robot lecture.

**Acknowledgments.** This work is supported in part by JSPS KAKENHI Grant Number 17H01992.

## References

1. Collins, J.: Education techniques for lifelong learning: giving a powerpoint presentation: the art of communicating effectively. *Radiographics* **24**(4), 1185–1192 (2004)
2. Melinger, A., Levelt, W.J.M.: Gesture and the communicative intention of the speaker. *Gesture* **4**(2), 119–141 (2004)
3. Arima, M.: An examination of the teachers' gaze and self reflection during classroom instruction: comparison of a veteran teacher and a novice teacher. *Bull. Grad. Sch. Educ. Hiroshima Univ.* **63**, 9–17 (2014 in Japanese)
4. Goldin-Meadow, S., Alibali, M.W.: Gesture's role in speaking, learning, and creating language. *Ann. Rev. Psychol.* **64**, 257–283 (2013)
5. Ishino, T., Goto, M., Kashihara, A.: A robot for reconstructing presentation behavior in lecture. In: Proceedings of the 6th International Conference on Human-Agent Interaction (HAI), pp. 67–75 (2018)
6. Kamide, H., Kawabe, K., Shigemi, S., Arai, T.: Nonverbal behaviors toward an audience and a screen for a presentation by a humanoid robot. *Artif. Intell. Res.* **3**(2), 57–66 (2014)
7. Goto, M., Ishino, T., Inazawa, K., Matsumura, N., Nunobiki, T., Kashihara, A.: Authoring robot presentation for promoting reflection on presentation scenario. In: Proceedings of the 14th Annual ACM/IEEE International Conference on Human Robot Interaction (HRI), pp. 660–661 (2019)
8. McNeill, D.: Hand and mind: What Gestures Reveal about Thought. University of Chicago press, Chicago (1992)
9. Presentation Sota. [https://sota.vstone.co.jp/home/presentation\\_sota/](https://sota.vstone.co.jp/home/presentation_sota/)



# Design Prompts for Virtual Reality in Education

Lawrence Kizilkaya<sup>1</sup>, David Vince<sup>1</sup>, and Wayne Holmes<sup>2</sup> 

<sup>1</sup> Learning Innovation, The Open University, Milton Keynes, UK  
lawrence.kizilkaya@open.ac.uk

<sup>2</sup> Institute of Educational Technology, The Open University, Milton Keynes, UK  
wayne.holmes@open.ac.uk

**Abstract.** The Learning Innovation team at the Open University, the UK's largest higher education institution, is leading a project to support the design and implementation of educational Virtual Reality (VR). Using an iterative design methodology enables the team to explore VR's educational affordances, and the new opportunities these represent to educators over established media forms. The design and development of two VR applications utilizing IBM Watson's Artificial Intelligence (AI), for language learning and phlebotomy, has resulted in design principles that reinforces understanding of implementing AI in VR. The design principles are part of a broader 'suitability toolkit' enabling a practical assessment of the appropriateness of VR in educational contexts early on in the design process.

**Keywords:** Virtual reality · Mixed reality · Instructional design

## 1 Introduction

In this paper, we discuss the experimental use of IBM Watson Artificial Intelligence (AI) in two applications as part of a broader project to explore the affordances of Virtual Reality (VR) in a higher education context. From this exploratory work, a series of design prompts have emerged, which will help educators consider VR as a medium for learning more generally, and is of relevance to those incorporating AI with VR.

### 1.1 Designing Immersive Learning Experience

VR uses head-mounted displays and software to generate realistic three-dimensional images, sounds and other sensations [1]. VR's uniquely immersive and experiential properties enable a sense of immersive presence [2, 3] which can potentially support pedagogic practices [4].

VR experiences can draw on a plurality of technologies and software (e.g. gaze, gesture control, haptics, voice interfaces and artificial intelligence) in such a way that the real environment and virtual/artificial environment overlap, giving rise to a mixed reality experience [5]. However, the challenge for pedagogic and learning design is to harness the increased immersion (afforded by mixed reality experiences) as a new form

of narrative where few guidelines exist [6], otherwise the potential of technologies like VR in a teaching context will be limited [7].

## 2 Suitability Toolkit Design Prompts

A ‘suitability toolkit’ has emerged from our exploratory work. It enables practical assessment of VR that is being designed to support educational practices, even before the process of ideation is complete. Framed as a set of design prompts, the toolkit offers practitioners an insight into common, yet easily overlooked, problems with designing learning experiences in VR. It is these prompts which formed the basis for our use of AI in VR.

Each prompt has a heading (indicative of its theme), a question, a brief explanation of what the prompt means, and examples of viable and limited applications of it. These are intended to be thought provoking ideas, which allow educators to assess the value of their VR design as a learning experience. The toolkit also addresses the consistency of effective pedagogy and its relationship with the hardware. Table 1 contains a summary list of the prompts included in the toolkit, with the full version being free to access on our website at [www.learninginnovation.org.uk](http://www.learninginnovation.org.uk).

**Table 1.** Summary list of suitability toolkit design prompts

Theme	Prompt
Immersive	Does your learning experience require an immersive environment?
Interactive	Will students be interacting with the environment?
Accurate	Is student understanding contingent on the accuracy of the environment?
Experiential	Is your learning experience an experiential one?
Self-contained	Is your learning experience studied as part of a larger activity?
Evaluable	Can students evaluate their learning solely within the experience?

## 3 AI and VR Case-Studies, Using the Design Prompts

From the suitability toolkit design prompts, many more context-specific principles can be elicited. To illustrate, we now consider two case study applications: the use of AI and VR in language learning and to support clinical phlebotomy practice. While the use of AI in both these applications took a similar form, the drivers were very different. However, in both, the AI was a mechanism by which users could engage a non-playable avatar in conversation. Using a voice interface, developed using the IBM Watson service, the user can talk to an avatar and expect a response. Rather than simply pursuing a natural conversation, the AI sought to serve one underlying purpose: to progress interactions through shared understanding.

### 3.1 Case Study 1: AI and VR in Language Learning

In our first application, we attempted to use AI as a vehicle for practicing basic conversational skills in a foreign language (French). In the absence of a real second player in the VR environment, AI needed to fill the role of conversational partner. However, recognizing that the AI cannot achieve this fully realistically, we focused on something it could achieve: conversational fluency - allowing users to practice without worrying about issues of accuracy. From this, and other similar decisions, the ‘self-contained’ toolkit design prompt was updated.

When it comes to learning to speak a second language, a student’s approach changes depending on their need. For example, learning a language in order to pass a written exam requires accuracy (i.e., getting grammar, syntax and spelling correct). On the other hand, communicating with people when travelling is probably better achieved by learning how to convey meaning, without needing to get all of the specifics right. This is where the AI’s shortcomings can be advantageous, even if its use of language is not entirely correct (given the available resources, having natural conversations with an AI may be ideal but not realistic). Nonetheless, sufficient training of the AI was still necessary. The result enabled progression through interaction, rather than a natural conversation, between the user and the AI within one of many scenes in the application. As this was our intention, we were able to meet our learning objectives.

The ‘self-contained’ design prompt was useful for the development of the VR scenario. In particular, it paved the way for the realization that it was not necessary for the language to be entirely accurate. The student is able to practice fluency solely within the VR application without needing to reference external sources of information.

A second design prompt that sits alongside this approach is the ‘evaluable’ prompt, in which we became concerned with the ‘digestion’ of learning rather than its ‘ingestion’ (as with the ‘self-contained’ prompt) and the types of methods applicable to retaining the value gained from practicing fluency in this natural way. In order to fulfil the need suggested by the ‘evaluable’ prompt, we built a voice recorder into the application which captured the entire conversation between the user and the AI. This recording was accessible within the application and could also be exported to a user’s smartphone or PC, where they could review their performance, so as to enable self-reflection and refine their future attempts.

### 3.2 Case Study 2: AI and VR in Phlebotomy

Building on our first use of AI, we next included it in a VR phlebotomy application (phlebotomy is the process of drawing blood samples from medical patients). This application aims to teach users the process behind venepuncture (making an incision in a vein with a needle) with a focus on the human interactions - something which is sometimes neglected in current training approaches.

Much in the same way as in our first case, the AI was used to facilitate a conversation between the user and an avatar. However, different toolkit design prompts encouraged us to think about what kinds of learning objectives would best make use of the way in which the AI conversational functions worked. In this instance, we used the

AI as an alternative to prescribed dialogue trees, which typically are used to present users with timely on-screen options from which to choose in order to progress the dialogue. The user is the instigator of these actions, just as they would be in real life. However, giving users multiple dialogue options to choose from can lead to false positives, because even if the user has no prior knowledge, they are able to use common sense to successfully navigate them. This goes for both the right thing to say and do, and knowing the correct timing and order of the process. Introducing the AI enables users to initiate conversations and actions without prompting, by being prepared to respond accordingly. Similarly to the language case, we were not especially interested in the AI being able to converse fluently. Rather, we wanted the AI and the user to share an understanding so that the user could progress through the multiple scenes that make up the correct venepuncture scenario.

The ‘immersive’ design prompt was particularly useful for our development of this VR scenario, because of the way in which a user would expect to act in the real world. The intention is to train phlebotomy practitioners using this application, so many facets need to be true to real life – the key question was which ones.

On the other hand, the need for free speech here is not the same as the language-learning case. Instead, the onus is on choosing the right actions at the right time, which is arguably one of the most difficult things to simulate well in VR. Finally, questions raised by the ‘experiential’ design prompt encouraged the exploration of an experience-based VR like this, as experiential VR can be well-suited at a fundamental level. With this phlebotomy application, users are not substituting real world practice. Instead, they are looking to gain practical confidence, and to reinforce and consolidate their understanding of both the procedure and the etiquette before working with real patients.

## 4 Discussion and Future Work

VR technology shows promise in educational contexts because of its uniquely immersive properties. When combined with AI and voice interfaces, even more immersive mixed-reality experiences become possible. However, this increased level of immersion alone—without consideration for the learning objectives, content and scenario—is unlikely to result in better pedagogical outcomes.

As our understanding of the pedagogic affordances of VR emerge, learning design guidance will need to evolve. This paper makes an early contribution to this work by presenting our suitability toolkit design prompts, which emerged from, and can be used to inform, the development of effective VR pedagogical scenarios. We also recognize the plurality of technologies, software and disciplines, and their rapid developments, all of which give rise to additional complexity. It may well be the case that a plurality of design practices and additional design prompts are needed to respond to this complexity.

## References

1. Fuchs, P., Moreau, G., Guitton, P. (eds.): *Virtual Reality: Concepts and Technologies*. CRC Press, Boca Raton (2011)
2. Sanchez-Vives, M.V., Slater, M.: From presence to consciousness through virtual reality. *Nat. Rev. Neurosci.* **6**(4), 332–339 (2005)
3. Slater, M.: Implicit learning through embodiment in immersive virtual reality. In: Liu, D., Dede, C., Huang, R., Richards, J. (eds.) *Virtual, Augmented, and Mixed Realities in Education*. SCI, pp. 19–33. Springer, Singapore (2017). [https://doi.org/10.1007/978-981-10-5490-7\\_2](https://doi.org/10.1007/978-981-10-5490-7_2)
4. Dede, C., Richards, J.: Strategic planning for R&D on immersive learning. In: Liu, D., Dede, C., Huang, R., Richards, J. (eds.) *Virtual, augmented, and mixed realities in education*, pp. 237–244. Springer, Singapore (2017). [https://doi.org/10.1007/978-981-10-5490-7\\_13](https://doi.org/10.1007/978-981-10-5490-7_13)
5. Milgram, P., Kishino, F.: A taxonomy of mixed reality visual displays. *IEICE Transact. Inf. Syst.* **77**, 1321–1329 (1994)
6. Cowling, M., Birt, J.: Pedagogy before technology: a design-based research approach to enhancing skills development in paramedic science using mixed reality. *Information* **9**, 29 (2018)
7. Parong, J., Mayer, R.E.: Learning science in immersive virtual reality. *J. Educ. Psychol.* **110** (6), 785–797 (2018)



# Assessing and Improving Learning Outcomes for Power Management Experiments Using Cognitive Graph

Yi Kuang<sup>1</sup>, Bin Duan<sup>1</sup>(✉), Shuyang Zhong<sup>2</sup>, and Mengping Lv<sup>1</sup>

<sup>1</sup> Xiangtan University, Xiangtan 411105, Hunan, China  
ky\_0814@outlook.com, db61850@163.com

<sup>2</sup> Texas Instruments Semiconductor Technologies (Shanghai) Co., Ltd,  
Shanghai 200122, China

**Abstract.** The series of Power Management Lab Kits (PMLK), released by Texas Instruments (TI), have been globally adopted in classroom settings. We propose a cognitive graph-based method to assist better adoption of TI-PMLK in Chinese power electronics education and specifically assessment of the experiment-based learning experience. First, construct a power management cognitive graph. Then, identify knowledge weaknesses using a Deterministic Inputs Noisy And Gate (DINA) model based cognitive diagnosis method. An Automatic Items Generation System (AIG) is then developed to generate personalized experiment items for any given student. Finally, learning outcomes evaluation is generated from the AIG experimental items using Bayesian psychometric models.

**Keywords:** Learning outcomes · Power electronics · Cognitive graph · Automatic item generation · Bayesian network

## 1 Introduction

In Chinese engineering education system, Texas Instruments (TI) University Program is widely adopted and has provided over 600 universities in China with analog and embedded processing technology learning tools, labs, and teaching materials.

Despite its prevalent adoption, there are a few issues with TI-PMLKs application in the Chinese power electronics experimental education. First of all, TI-PMLK based experiments have predominately been used for validating students knowledge points rather than for enabling and encouraging open-ended exploration and discovery. Secondly, TI-PMLK has commonly used in group experiments which multiple students participate in. This introduces a homogeneity problem in that students cannot effectively identify their individual academic

---

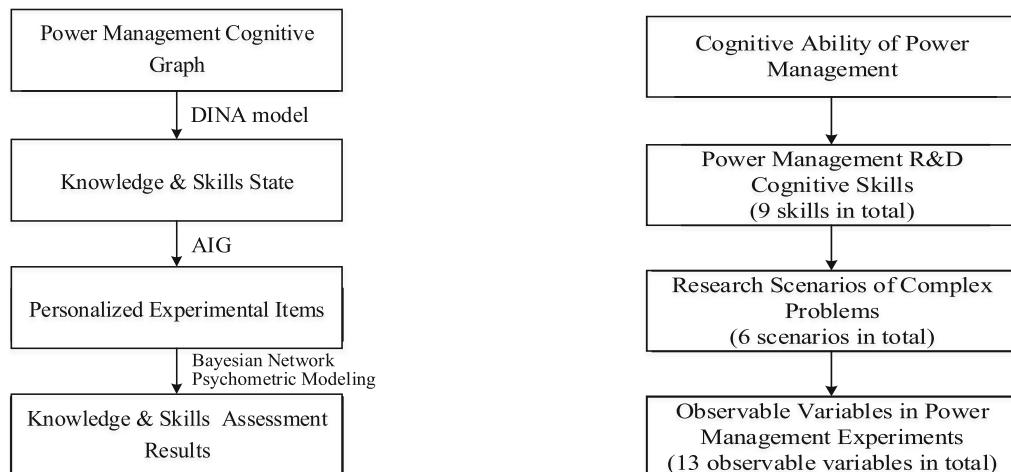
Supported by Development of Innovative Practice Platform Based on TI PMLK, Texas Instruments Semiconductor Technologies (Shanghai) Co., Ltd, Reform of the Teaching Content and the Course System.

weaknesses. Moreover, assessment of these power electronics experiments is often overly subjective and one-sided, making it difficult to evaluate the effectiveness of the study or identify areas for improvement [10].

In this work, to improve the effectiveness of TI-PMLKs application in Chinese power electronics engineering education system, a method is proposed for assessing and improving the learning outcomes of power management experiments using cognitive graph. Firstly, a power management cognitive graph is constructed. Secondly, based on the cognitive graph, a cognitive questionnaire for Buck regulator is conducted. The DINA model is adopted to analyze the mastery of knowledge. Then, AIG for power management experiment is developed. Finally, learning outcomes evaluation is carried out, with the example of investigation skills, by using Bayesian psychometric model methods with the personalized experimental items generated by AIG. The flowchart of the proposed procedure is illustrated in Fig. 1.

## 2 Power Management Cognitive Graph

Firstly, we analyzed the relationships between physical parameters and power supply performances in TI-PMLK Buck experiments [2]. Then, a cognitive graph for power management is constructed based on the domain knowledge and correlation analysis of parameters and factors, as shown in Fig. 2.



**Fig. 1.** Flowchart of the proposed method

**Fig. 2.** Power management cognitive graph

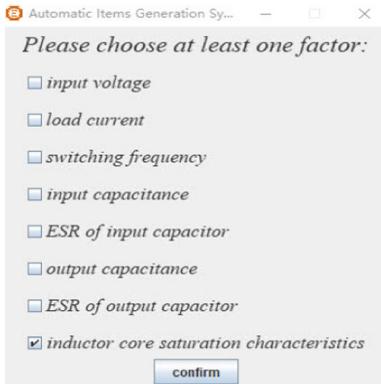
## 3 Cognitive Diagnosis Based on DINA Model

To assess examinees' mastery of Buck regulator knowledge, we designed a cognitive questionnaire, containing 16 single-choice questions and 9 knowledge points. The DINA cognitive model is used to estimate multidimensional knowledge mastery degree [1]. Using examinees' response matrix and Q-matrix, the conditions of their knowledge points are calculated using the maximum a posteriori algorithm [12, 13].

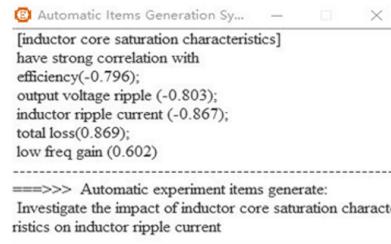
## 4 AIG Based on Cognitive Graph

On the basis of power management cognitive graph and simulation data, Pearson correlation analysis is used to calculate the correlation coefficient  $r$  between factors and performance indexes.

Based on the knowledge weaknesses of particular examinee evaluated by the DINA model, AIG will list the factors that are strongly related ( $|r| > 0.6$ ) to the selected one according to the correlation coefficient matrix, and generate personalized experimental items from them, as shown in Figs. 3 and 4.



**Fig. 3.** Interface of AIG



**Fig. 4.** The results of AIG

## 5 Learning Outcomes Assessment

### 5.1 Evidence-Centered Design Framework

On the account of investigation skill, the TI-PMLK mainly training about, it is necessary to do cognitive diagnosis assessment, to measure this ability after examinees finish the items generated by AIG, as well as evaluate the examinees' weaknesses and strengths [3, 8]. The evidence-centered design (ECD) assessment framework is used [5, 7]. The Bayesian network is adopted to do the cognitive assessment [11]. In the layers of ECD, conceptual assessment framework mainly contains a proficiency model, task models, and evidence models [6, 9].

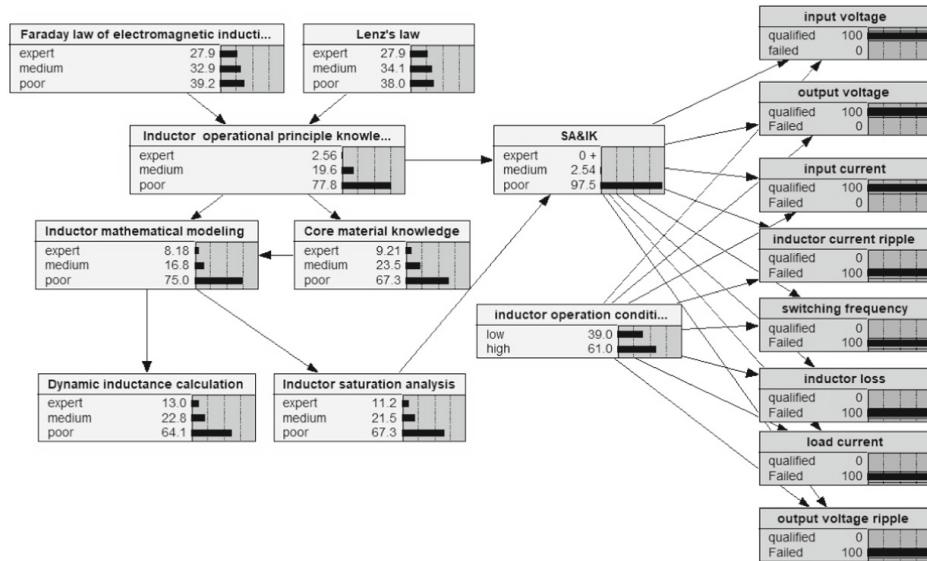
A proficiency model defines one or more variables related to the knowledge, skills, and abilities we expect to measure in evidence model. We built an inductor proficiency model by domain modeling and analysis based on cognitive graph. The proficiency model can be assessed by the feedback of students' evaluation results. Task models describe many situations which extract evidences to support the evidence models. With the example of investigating impact of inductor core saturation characteristics on inductor current ripple, the examinee is required to design Buck regulator circuit and analysis the saturation characteristics of ferrite and powder core inductor in different operating conditions. The relevant calculation formulas and knowledge should be used to interpret experimental data,

and make valid conclusions. Evidence models bridge the proficiency model and the task models, which specifies detailed guidance on how evidence from assessment tasks to refine our information about the proficiency model variables [4]. Two evidence models, dynamic inductance calculation and inductor saturation analysis evidence model, are built, corresponding to different aspects of inductor mathematical modeling. The combination of inductor proficiency model and inductor saturation analysis evidence model is illustrated in Fig. 5.

## 5.2 Case Study

Three graduate students major in Electrical Engineering in a certain university were selected randomly from 27 students to do investigation skills assessment. Examinees were supposed to investigate the impact of inductance saturation on current and voltage ripple. By comparing the theoretical and calculated values, the results show that examinees completed tasks with 8 observed data points, in which 8, 6 and 3 indicators were qualified respectively for examinees A, B, and C.

By using software Netica, the prior and posterior probability of the inductor saturation condition analytic investigation ability for these three examinees are obtained. The assessment results of examinee C is shown in Fig. 5. We can conclude that examinee C has insufficient knowledge of the core material knowledge.



**Fig. 5.** Inductor material investigation ability assessing system of examinee C

## 6 Conclusions and Future Work

This paper addresses a key issue in today's Chinese power electronics education. A method has been proposed for this field but may be transferrable to other disciplines. As one potential future work, we will study examinees' work process in order to perform formative assessment and refine parameters in the Bayesian network. By leveraging data generated in the process, we can further improve the assessments' validity and accuracy.

## References

1. Culpepper, S.A.: Bayesian estimation of the DINA model with Gibbs sampling. *J. Educ. Behav. Stat.* **40**(5), 454–476 (2015)
2. Femia, N.: Power Management Lab Kit Buck Experiment Book. Texas Instruments, Dallas (2016)
3. Guzmán, E., Conejo, R.: Measuring misconceptions through item response theory. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 608–611. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_73](https://doi.org/10.1007/978-3-319-19773-9_73)
4. Levy, R.: Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educ. Assess.* **18**(3), 182–207 (2013)
5. Levy, R., Mislevy, R.J.: Bayesian Psychometric Modeling. Chapman and Hall/CRC, Boca Raton (2016)
6. Millán, E., Jiménez, G., Belmonte, M.-V., Pérez-de-la-Cruz, J.-L.: Learning Bayesian networks for student modeling. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 718–721. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_100](https://doi.org/10.1007/978-3-319-19773-9_100)
7. Min, W., Frankosky, M.H., Mott, B.W., Wiebe, E.N., Boyer, K.E., Lester, J.C.: Inducing stealth assessors from game interaction data. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 212–223. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_18](https://doi.org/10.1007/978-3-319-61425-0_18)
8. Mislevy, R.J.: Sociocognitive Foundations of Educational Measurement. Routledge, Abingdon (2018)
9. Mislevy, R.J., Haertel, G., Riconscente, M., Rutstein, D.W., Ziker, C.: Evidence-centered assessment design. Assessing Model-Based Reasoning using Evidence-Centered Design. SS, pp. 19–24. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-52246-3\\_3](https://doi.org/10.1007/978-3-319-52246-3_3)
10. Peng, Y., Luo, Z., Yu, X., Gao, C., Li, Y.: The optimization of test design in cognitive diagnostic assessment. *Acta Psychologica Sinica* **48**(12), 1600–1611 (2016)
11. Rosen, Y., et al.: Adaptive learning open source initiative for MOOC experimentation. In: International Conference on Artificial Intelligence in Education (2018)
12. Song, L., Wang, W., Dai, H., Ding, S.: Comparing two classification methods based on the attribute hierarchy method and the DINA model. In: International Conference on Consumer Electronics (2012)
13. Sorrel, M.A., Abad, F.J., Olea, J., de la Torre, J., Barrada, J.R.: Inferential item-fit evaluation in cognitive diagnosis modeling. *Appl. Psychol. Meas.* **41**(8), 614–631 (2017)



# Does Choosing the Concept on Which to Solve Each Practice Problem in an Adaptive Tutor Affect Learning?

Amruth N. Kumar<sup>(✉)</sup> 

Ramapo College of New Jersey, Mahwah, NJ 07430, USA  
amruth@ramapo.edu

**Abstract.** We conducted a controlled study to investigate whether having students choose the concept on which to solve each practice problem in an adaptive tutor helped improve learning. We analyzed data from an adaptive tutor used by introductory programming students over three semesters. The tutor presented code-tracing problems, used pretest-practice-post-test protocol, and presented line-by-line explanation of the correct solution as feedback. We found that choice did not increase the amount of learning or pace of learning. But, it resulted in greater improvement in score on learned concepts, and the effect size was medium.

**Keywords:** Choice of practice problem · Adaptive tutor · Controlled study

## 1 Introduction

An adaptive tutor presents practice problems on the concepts students have not yet mastered. If the tutor requires students to select the concept on which they would like to solve each practice problem, it would give students choice, and thereby, a sense of agency [2], which is known to improve performance on learning tasks (e.g., [3]). It would give them a sense of control over their path through the learning material, the type of choice typically associated with enhanced learning [3, 13]. In response, if the tutor honors the student's choice when presenting the next practice problem, it facilitates self-directed learning [6]. If it does not, i.e., it presents the next problem on a concept other than the one chosen by the learner, it promotes cognitive dissonance [4], which is known to help learning [1]. So, the act of choosing the practice concept might itself enhance learning, whether or not the tutor subsequently honors the choice. We investigated whether having the student choose the concept of each practice problem helped improve learning in an adaptive tutor.

For this study, we used a problem-solving software tutor on loop, a programming concept. The tutor presents code-tracing problems, wherein, the student is asked to read a program and identify its cumulative output, one output at a time along with the line in the program that produces that output. The tutor provides line-by-line explanation of the correct answer as feedback, which has been shown to improve learning in prior evaluations [8]. This explanation is in the style of a worked example [12]. The tutor is

adaptive [9] and covers 10 concepts in C++, Java or C#. It is part of a suite of problem-solving tutors for introductory programming topics called problets ([www.problets.org](http://www.problets.org)).

The tutor is accessible over the web. Instructors of introductory programming courses use the tutor typically for after-class assignments. So, typically, students use the tutor on their own time, and often, multiple times till they master all the concepts. In this study, the institutions that used the tutor were randomly assigned to control or experimental group each semester. Data was collected over 3 semesters from Fall 2017 - Fall 2018. During that time, the number of students who used the tutor and granted permission for the use of their data in the study was 202 in control group and 179 in experimental group.

Every time the tutor was used, it administered pretest-practice-post-test protocol [7]. During pretest, the tutor presented one problem per concept to prime the student model. If the student solved a problem partially correctly, incorrectly, or opted to skip the problem without solving it, the tutor presented line-by-line explanation as feedback. Once the student had solved all the pretest problems, the tutor presented practice problems on only the concepts on which the student had skipped solving the problem or solved the problem partially/incorrectly during pretest. For each such concept, the tutor presented multiple problems until the student had mastered the concept, i.e., solved a minimum percentage (e.g., 60%) of the problems correctly. After each incorrectly solved problem, the tutor presented line-by-line explanation of the correct answer. Finally, during adaptive post-test, which was interleaved with practice, the tutor presented a test problem on each and only the concepts that had already been mastered by the student during practice. Pretest, practice and post-test were administered by the tutor back-to-back without interruptions, entirely over the web. The entire protocol was limited to 30 min.

The only difference in treatment between control and experimental groups was during adaptive practice. Before each practice problem, the tutor presented a list of all the concepts and the percentage of problems the student had solved correctly on each concept. The tutor used the same pre-determined order of concepts for both the groups. Experimental subjects were asked to pick the next concept that they wanted to practice, *but only when at least two concepts remained un-mastered*. In contrast, control subjects just viewed the list of concepts before moving on to the next problem. As a result, the sequence of practice problems solved by experimental subjects differed from that of control subjects. For control subjects, the tutor presented problems in the pre-determined order of concepts using round-robin algorithm, taking care not to present more than two problems back to back on any one concept. For experimental subjects, it used the subject's choice as the seed to pick the next concept not yet mastered in the pre-determined order of concepts, and presented a problem on it. *The resulting problem may or may not have been on the concept chosen by the student*. According to learning theory, the student could benefit whether the practice problem matched the chosen concept or not: when the two matched, the student could benefit from cueing [11] and self-directed learning [6]. When the two mismatched, the student could benefit from cognitive dissonance [1].

The concepts covered by the tutor can be classified as known, attempted, practiced or learned by the student. A concept is **known** if the student solves the pretest problem on the concept correctly. A concept is **learned** if the student solves the pretest problem

on the concept partially/incorrectly or skips solving it, solves enough problems during practice to master it, and solves the post-test problem correctly. On the other hand, if the student solves the post-test problem incorrectly, the concept is **practiced**, not yet learned. If so, the tutor schedules additional practice problems on the concept for the student. If the student runs out of time because of the 30-min. limit placed on the duration of the tutoring session and does not complete mastering the concept during practice, the concept is categorized as **attempted**.

During grade calculation of code-tracing problems, the outputs identified in the correct sequence (*c*) were credited and any incorrectly identified outputs thereafter (*i*) were penalized. The grade was calculated as  $\max((c - i)/n, 0)$ , where *n* was the total number of outputs in the program. Therefore, the score on each problem was normalized to  $0 \rightarrow 1.0$ .

If a student used a tutor multiple times, we considered data from the session when the student had learned the most number of concepts. If the student did not learn any concepts, we considered data from the first session when the student had solved the most number of problems. Since the only difference in treatment between control and experimental groups was during practice stage, and that too, when students solved problems on two or more concepts, only students who had solved practice problems on two or more concepts were retained for the study in both control and experimental groups. As a result, 98 students each remained in control and experimental groups.

We considered 7 variables for the study: (1) The **score per pretest problem** to verify that control and experimental groups were comparable; (2) The **number of practice problems** solved during the session. This included practice problems solved on learned concepts as well as concepts merely practiced or attempted. This and the next two variables were used to evaluate the impact, if any, of having to choose the concept before each practice problem; (3) The **score per practice problem**; (4) The **time spent per practice problem**; (5) The **number of concepts learned** as a measure of the amount of learning; (6) The **number of practice problems solved per learned concept**, as a measure of the pace of learning. It was calculated by dividing the number of practice problems solved on all the learned concepts by the number of concepts learned; and (7) **Pre-post change in grade per problem on learned concepts** as a measure of improvement in learning. The fixed factor was treatment: experimental group students had to choose the concept underlying each practice problem, whereas control group students did not.

## 2 Results and Discussion

We conducted one-way ANOVA of each of the variables with treatment as the fixed factor.

We found no significant main effect for treatment on the **score per pretest problem** [ $F(1,195) = 1.841, p = 0.176$ ] or the time spent per pretest problem [ $F(1,195) < 0.001, p = 0.991$ ]. So, the two groups were comparable. We did not find any significant difference in the **number of practice problems** solved [ $F(1,195) = 0.991, p = 0.321$ ], but the **score per practice problem** was significantly different [ $F(1,195) = 7.897, p = 0.005$ ]: control group subjects scored a mean of  $0.835 \pm 0.029$  points per practice

problem whereas experimental subjects scored  $0.776 \pm 0.029$  points. *So, control subjects scored significantly more per practice problem.* One explanation for this difference is that experimental subjects suffered cognitive dissonance when the problem they were presented was not on the concept they chose. Hence, they scored significantly less on practice problems. We found no significant main effect for treatment on the **time spent per practice problem**.

We did not find a significant difference between control and test groups on **the number of concepts learned**. *So, the treatment did not affect the amount of learning.* We found a significant main effect for treatment on the **number of practice problems solved per learned concept** [ $F(1,109) = 4.965$ ,  $p = 0.028$ ]: control subjects solved a mean of  $2.92 \pm 0.249$  problems to learn each concept whereas experimental subjects solved  $3.30 \pm 0.223$  problems. *So, the pace of learning was significantly slower for experimental subjects than control subjects.* This may also be ascribed to cognitive dissonance: the practice problem presented by the tutor matched the concept chosen by the student in only 40.65% of the cases. *So, the treatment of merely providing choice without always honoring it did not benefit the pace of learning.*

We also found a significant main effect for treatment on **pre-post change in grade per problem on learned concepts** [ $(F1,109) = 5.028$ ,  $p = 0.027$ ]: the change was  $0.716 \pm 0.056$  for control subjects compared to  $0.802 \pm 0.051$  for experimental subjects. *So, the improvement in score was significantly greater for experimental group than control group.* The effect size (Cohen's d) is 0.43, corresponding to medium effect.

To summarize, even though the two groups were comparable to begin with, experimental group needed significantly more practice problems to learn each concept, but had significantly greater improvement in score on the learned concepts. We found the same results when we considered only less-prepared students, i.e., those whose score per pretest problem was 0.9 or less. Insofar as choice enhances intrinsic motivation and engenders agency, the results of this study warrant the provision of choice, given that we did not find any negative cognitive effects of choice.

Experimental subjects chose the first concept in the list presented to them 65.25% of the time, and second concept 22.88% of the time. Given that the list always contained at least two concepts, this lopsided distribution of choice suggests that students more often than not chose concepts in the order in which they were presented.

Cueing has been shown to improve transfer of learning [11], in particular, when multiple problem-solving examples are presented to the learner [5, 10]. In our experimental setup, the name of the concept could have served as a consistent, valid cue if the practice problem had always matched the student's choice. We speculate that providing choice, combined with consistent, valid cueing might lead to better learning outcomes than were observed in this study. This will be the subject of a future study.

**Acknowledgments.** Partial support for this work was provided by the National Science Foundation under grant DUE-1432190.

## References

1. Aïmeur, E., Frasson, C., Lalonde, M.: The role of conflicts in the learning process. *SIGCUE Outlook* **27**(2), 12–27 (2001)
2. Bandura, A.: Social cognitive theory: an agentic perspective. *Annu. Rev. Psychol.* **52**(1), 1–26 (2001)
3. Cordova, D., Lepper, M.: Intrinsic motivation and the process of learning: beneficial effects of contextualization, personalization, and choice. *J. Educ. Psychol.* **88**, 715–730 (1996)
4. Festinger, L., Carlsmith, J.M.: Cognitive consequences of forced compliance. *J. Abnorm. Soc. Psychol.* **58**(2), 203–210 (1959)
5. Gentner, D., Loewenstein, J., Thompson, L.: Learning and transfer: a general role for analogical encoding. *J. Educ. Psychol.* **95**(2), 393–408 (2003)
6. Hammond, M., Collins, R.: *Self-Directed Learning: Critical Practice*. Kogan Page Limited, London (1991)
7. Kumar, A.N.: A model for deploying software tutors. In: Proceedings IEEE 6th International Conference on Technology for Education (T4E), Amritapuri, India, pp. 3–9 (2014)
8. Kumar, A.N.: Explanation of step-by-step execution as feedback for problems on program analysis, and its generation in model-based problem-solving tutors. *Technol. Instr. Cogn. Learn. (TICL)* **4**(1), 65–107 (2006). *J. Special Issue on Problem Solving Support in Intelligent Tutoring Systems*
9. Kumar, A.: A scalable solution for adaptive problem sequencing and its evaluation. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) AH 2006. LNCS, vol. 4018, pp. 161–171. Springer, Heidelberg (2006). [https://doi.org/10.1007/11768012\\_18](https://doi.org/10.1007/11768012_18)
10. Norman, G., Dore, K., Krebs, J., Neville, A.J.: The power of the plural: effect of conceptual analogies on successful transfer. *Acad. Med.* **82**(10), S16–S18 (2007)
11. Speicher, T., Bell, A., Kehrhahn, M., Casa, D.: Effect of cueing on learning transfer among health profession students engaged in a case-based analogical reasoning exercise. *Internet J. Allied Health Sci. Pract.* **12**(3), Article no. 4 (2014)
12. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. *Cogn. Instr.* **2**, 59–89 (1985)
13. Zuckerman, M., Porac, J., Lathin, D., Smith, R., Deci, E.: On the importance of self-determination for intrinsically-motivated behavior. *Pers. Soc. Psychol. Bull.* **4**, 443–446 (1978)



# Measuring Content Complexity of Technical Texts: Machine Learning Experiments

M. Zakaria Kurdi<sup>(✉)</sup>

University of Lynchburg, Lynchburg, USA  
kurdi\_m@lynchburg.edu

**Abstract.** Classifying texts by their content complexity is important for applications like adaptive foreign language reading recommender systems and information retrieval. The goal of this paper is to propose a computational model of technical texts' content complexity based on three criteria: knowledge depth, required knowledge, and content focus. To implement this model, 28 features of content and lexical complexity were extracted from 1702 texts of three types: general blogs, science journalistic texts and research papers. The machine learning experiments showed that content features alone can provide high classification accuracy.

**Keywords:** Text content complexity · Text mining ·  
Reading recommender systems · Intelligent tutoring systems

## 1 Introduction

The goal of this paper is to build a system that can provide a classification of the texts based on their content complexity. Such a system is useful within a reading recommender system for foreign language learners as it can help match a set of candidate texts to a reader not only according to his language mastery but also according to his intellectual level. Furthermore, this system can provide useful feedback to writers about the difficulty of the content of their texts. Finally, such a system can be integrated within the ranking component of an information retrieval system.

## 2 A Model of the Content Complexity of Technical Texts

Inspired by the Depth of Knowledge (DoK) model that was introduced by Norman Webb [1–3], the proposed model measures the complexity of technical texts with the following three criteria: knowledge depth, required prior knowledge, and content focus. **Knowledge depth** is about the specialization of the content covered by a text. The more specialized the content of a text, the harder it is to understand by common readers. **Required knowledge** of a text is about the number of concepts that the reader needs to know, prior to reading the text, to be able to understand its content. **Content focus** is about the focus on a limited number of topics with an in-depth presentation and discussion.

### 3 Corpus

A corpus of 1702 texts, grouped into three categories, is used. The first group is a collection of 600 blogs<sup>1</sup> with at least 700 words per blog. This corpus is used as an example of semi-formal texts by non-professional writers. Given the shallow way the subjects are presented and discussed and that the targeted audience of the blogs are common readers of different age ranges and intellectual levels, this is considered as the lowest degree in terms of content complexity among the three collected groups of texts. A group of 502 full scientific research papers is selected by hand from different free available resources on the web such as CORE<sup>2</sup>. The third group is made of 600 full scientific press papers. Although these papers present scientific content, they target a general audience of *well-informed* readers but who are not necessarily specialized in the paper's field.

## 4 Feature Extraction

The features used in this paper are grouped into four categories, where the first three categories echo the elements of the content complexity model proposed in Sect. 2.

### 4.1 Knowledge Depth Features

To measure the depth of knowledge in a text, the semantic relationships of their words are used as an indicator. Are proposed here two measures of relationships based on the horizontal and vertical axes. The horizontal axis is about the synonymous relationships between the words. The vertical axis is about the position of a word within the hyponymic/hypernymic hierarchy measured by the number of its hyponyms.

### 4.2 Prior Knowledge Requirement Features

In this paper, Prior Knowledge Requirement (PKR) is measured as the number of difficult words or abbreviations mentioned in the text without definition.

The number of illustrations (figures and tables) used is also considered as an indication of PKR. This gives two features: the Illustration Mention per Words (IMW) and the Illustration Mention per Difficult Words (IMDW).

The Percentage of Abbreviations per Words (PAW), as well as the Percentage of Defined Abbreviations per Words (PDAW) are also used as features.

To calculate the Ratio of Defined Difficult Words to the total number of difficult words in a text (RDDW), it is necessary to build a module that can detect if a sentence is a definition of a word or not. The approach adopted here consist of finding the first occurrence of a difficult word in the text and decide if the sentence in which it is used or the next one is a definition of this word. Several works in the literature targeted building definition detection modules like [4]. Two groups of criteria are typically used

---

<sup>1</sup> <https://www.kaggle.com/rtatman/blog-authorship-corpus>.

<sup>2</sup> CORE (COnecting REpositories) is an aggregation of papers from open access journals <https://www.jisc.ac.uk/core>.

independently or in combination: syntax and semantics. There are different syntactic layouts of definitions. This is processed with syntactic patterns like in [5]. In this paper, an approach based on ngram profiles of POS tags is adopted. Hence, profiles of real definitions and sentences that are not definitions are built. The distance between a candidate sentence and both profiles is measured to decide if the sentence's layout is closer to a definition or to a general sentence. Semantically, definitions use more specific terms than average words. It is commonly assumed that the key has a hyponym relationship with the words of its definition [5] and [6]. In this paper, in addition to hyponymic relations, other semantic relations such holonym, synonym, antonym, and semantic field are used as indicators of definitions.

### 4.3 Content Focus Features

A text is focused when it is about a single or at least a small number of topics. The focus is an indication of content complexity as it signals in-depth coverage. To measure the focus of a text, the first step consists of extracting the key nouns from the texts. The nouns are extracted using the Rapid Automatic Keyword Extraction (RAKE) algorithm [7]. The distances between the keywords are then calculated with the Leacock-Chodorow (LCH) Similarity<sup>3</sup>.

### 4.4 Miscellaneous Lexical Features (MLF)

As lexicon and content are inseparable, several classic lexicon complexity features are also considered. These features can be viewed as being both content and linguistic features. Hence, seventeen lexical features are examined here as candidates for classifying the texts like lexical density and lexical sophistication. Lexical diversity is also considered with features like Type Token Ratio (TTR) and Guiraud's corrected TTR (GTTR) [8] and Carroll's corrected TTR (CTTR) as well as the Continuous Lexical Frequency Score (CLFS) [9]. Furthermore, several psycholinguistic measures are also extracted such as Kucera-Francis Written Frequency (KFWF), Kucera-Francis number of categories (KFnC), Kucera-Francis number of samples (KFnS) [10], Brown verbal frequency, Familiarity rating, Concreteness rating, Imageability rating, Meaningfulness (Colorado Norms) [11], Meaningfulness (Paivio Norms), as well as age of acquisition rating. For every lexical item, these psycholinguistic features are extracted from the MRC<sup>4</sup> database.

## 5 Text Classification by Content Complexity Experiments and Results

To evaluate the considered features, two models are built. The first is made of all the 28 features. The second model includes only the 11 content features proposed in Sects. 4.1, 4.2 and 4.3. This helps judge how decisive are the content features in the

---

<sup>3</sup> Based on the shortest path that connects the senses and the maximum depth of the hierarchy in which the senses occur.

<sup>4</sup> [http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa\\_mrc.htm](http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm).

classification. Several Machine Learning Algorithms (MLA) are used in preliminary experiments. The results of the two best MLA, Neural Network and Random Forest, are reported in Table 1. Cross-validation with 20 folds is performed.

**Table 1.** Content difficulty classification results

Model	Classifier	AUC	F1 <sup>a</sup>	Precision	Recall
All features	Random forest	0.99	0.98	0.98	0.98
	NN	1	0.98	0.98	0.98
Content features only	Random forest	0.98	0.97	0.97	0.97
	NN	0.99	0.98	0.98	0.98

<sup>a</sup>See [12] for a definition of these measures.

## 6 Discussion

The results show that content features are as effective as all the features combined. To show how effective is the classifier in distinguishing between the three types of texts, a confusion matrix is provided in Table 2.

The confusion matrix shows that some limited confusion occurs between research papers and scientific press papers. Given the proximity of the two types, this limited confusion (about 3% of the texts in these two categories are misclassified) is a good indication of the performance of the classifier. Blogs are completely distinct from the two other categories. This is possibly the result of clear boundaries drawn between blogs and the scientific texts both in terms of depth of words, focus, and usage abbreviations.

**Table 2.** Confusion matrix, random forest MLA with the content features only

		Predicted			Total
		Blog	Research	Sc. Press	
Actual	Blog	<b>600</b>	0	0	600
	Research	0	<b>480</b>	23	503
	Scientific	0	13	<b>587</b>	600

## 7 Conclusion and Perspectives

This paper is about modeling and classifying the content complexity of technical texts. First, a model is proposed based on knowledge depth, prior knowledge requirement, and content focus. This led to propose or adopt 28 features that were extracted from the texts. Experiments were carried on a set of 1702 texts of three types: blogs (lower content complexity), scientific journalism papers (middle content complexity), and research papers (highest level of content complexity). The results showed that the implemented system with the proposed features is effective in distinguishing the three types of texts.

## References

1. Webb, N.: Alignment of science and mathematics standards and assessments in four states, Washington, D.C. CCSSO. Research Monograph No. 18: August 1999. [https://www.researchgate.net/publication/239925507\\_Alignment\\_of\\_science\\_and\\_mathematics\\_standards\\_and\\_assessments\\_in\\_four\\_states](https://www.researchgate.net/publication/239925507_Alignment_of_science_and_mathematics_standards_and_assessments_in_four_states)
2. Webb, N.: 28 March, Depth-of-Knowledge Levels for four content areas, unpublished paper (2002)
3. Wise, S.L., Kingsbury, G.G., Webb, N.L.: Evaluating content alignment in computerized adaptive testing. *Educ. Measur. Issues Pract.* **34**(4), 41–48 (2015)
4. Fahmi, I., Bouma, G.: Learning to Identify Definitions using Syntactic Features, Workshop of Learning Structured Information in Natural Language Applications, EACL, Italy (2006)
5. Fiser, D., Pollak S., Vintar S.: Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, pp. 2932–2936 (2010)
6. Pollak, S., Vavpetic, A., Kranjc, J., Lavrac N., Vinta, S.: NLP workflow for on-line definition extraction from English and Slovene Text Corpora. In: Proceedings of KONVENTS, Vienna, 19 September (2012)
7. Rose, S., Dave, E., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Berry, M., Kogan, J. (eds.) *Text Mining: Applications and Theory*. Wiley, Hoboken (2010). ISBN 978-0-470-74982-1
8. Guiraud, P.: *Problèmes et Méthodes de la Statistique Linguistique*. D. Reidel, Dordrecht (1960)
9. Kurdi, M.Z.: Lexical and syntactic features selection for an adaptive reading recommendation system based on text complexity. In: Proceedings of the 2017 International Conference on Information System and Data Mining, ICISDM 2017, pp. 66–69 (2017)
10. Francis, W.N., Kucera, H.: *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston (1982)
11. Nickerson, C.A., Cartwright, D.S.: Behavior Research Methods. *Instrum. Comput.* **16**, 355 (1984). <https://doi.org/10.3758/BF03202462>
12. Kurdi, M.Z.: *Natural Language Processing and Computational Linguistics 2: Semantics, Discourse, and Applications*, ISTE. ISTE-Wiley, London (2017)



# Should Students Use Digital Scratchpads? Impact of Using a Digital Assistive Tool on Arithmetic Problem-Solving

Minji Kwak and Gahgene Gweon<sup>(✉)</sup>

Department of Transdisciplinary Studies, Seoul National University,  
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea  
[{indigo218, ggweon}@snu.ac.kr](mailto:{indigo218, ggweon}@snu.ac.kr)

**Abstract.** An excessive cognitive load may reduce a student’s problem-solving performance by preventing effective learning. Using an assistive tool, such as a notepad, can reduce such extraneous cognitive while solving a problem, thereby improving a student’s performance. In this paper, we collected game log data from an educational game, called “Double digit”, which has a digital scratchpad as an assistive tool that can be used to reduce student’s cognitive load. We examine whether a correlation exists between the amount of “digital scratchpad usage” and a student’s “game performance”. Game log data, which consisted of 1,440,000 actions, was collected from 418 students in kindergarten to grade 2. Our data analysis using person-correlation shows a significant positive relationship between digital scratchpad usage and game performance for all three game difficulty levels. Interestingly, the correlation increases as the game difficulty level increases. This suggests that as game level difficulty increases, which requires a higher cognitive load of a student, students who used the digital scratchpad had higher game performance.

**Keywords:** Digital assistive tool · Cognitive load · Digital scratchpad · Game performance · Arithmetic addition

## 1 Introduction

A student’s problem-solving performance can be improved by minimizing his cognitive load, which is the total amount of mental efforts required in learning [2]. Sweller states that a learner’s cognitive load is divided into 3 categories; intrinsic, extraneous, and germane [1]. Intrinsic cognitive load is associated with the difficulty of a task itself. Extraneous cognitive load is associated with effectiveness of teaching methods. Germane cognitive load is a learner’s mental effort left within the capacity of a limited working memory. According to Sweller’s Cognitive Load Theory, a learner has to process new information in a limited amount of working memory, and the sum of these three types of cognitive load categories should not exceed the total amount that working memory can cover and process. Among these categories, the extraneous cognitive load can be reduced by using a more effective educational tool. In turn, working memory capacity freed by extraneous cognitive reduction is allocated to

germane cognitive load. Then the learner can then put more mental efforts to tasks, so that schema acquisition and automation become easier [3].

In this study, we hypothesize that in the context of solving math problems in a digital environment, using a digital scratchpad as an assistive tool will be associated with an improved student's arithmetic problem-solving performance due to reduced cognitive load. We conjecture that a digital scratchpad can be useful in reducing extraneous cognitive load as it imitates the paper and pencil in the real world. In line with our hypothesis, there have been some studies which suggest usage of digital assistive tools would reduce extraneous cognitive load when conducting various tasks. For example, Ando and Ueno [5] measured learner comprehension and memory retention during e-learning, making annotations and writing notes in Japanese using assistive tools such as keyboard or tablet PCs. Their study showed that writing on a tablet PC, compared to using a keyboard, increases learners' comprehension and memory retention as writing annotations on tablet PC helps to reduce the extraneous cognitive load. Oviatt's study [4] compared performance characteristics such as memory retention and math errors while solving math problems using different assistive tools such as a pen tablet and a digital paper. High performing students made fewer errors when using digital tools than when using paper and pen because using digital tools caused less extraneous cognitive load. These studies support the effectiveness of digital assistive tools during problem-solving activities by reducing cognitive load. However, these studies did not examine whether the amount of assistive tool usage affects students' performance.

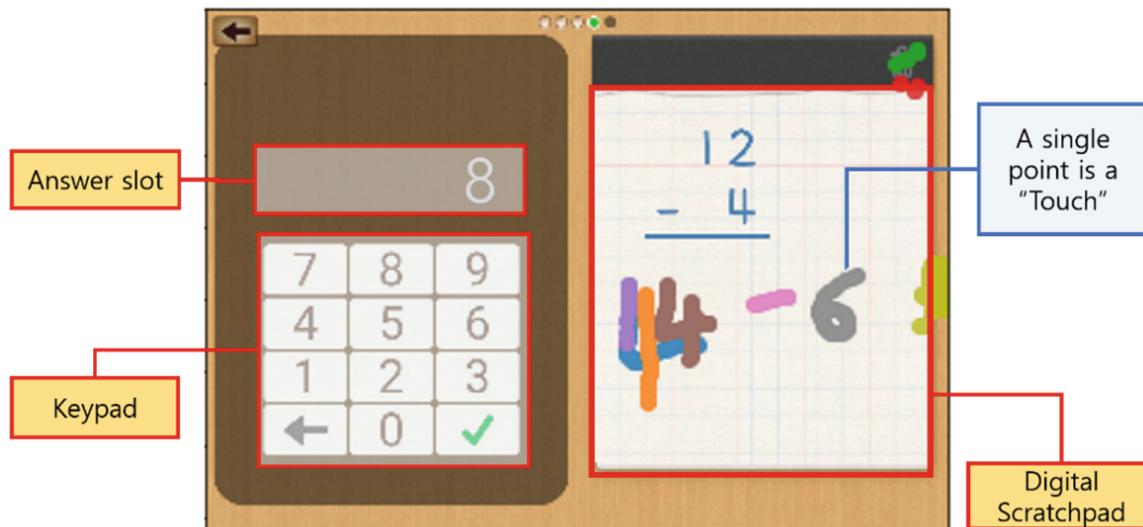
If an intrinsic cognitive load is low, total cognitive load is unlikely to exceed working memory capacity. However, extraneous cognitive load becomes more critical as intrinsic cognitive load increases [3]. Therefore, in our study, we examined if correlations between digital scratchpad usage and game performance increases at different game difficulty levels that require a different amount of intrinsic cognitive load. Thus, our hypothesis is as follows: The correlation between digital scratchpad usage and game performance is higher in games with higher difficulty that require more cognitive load.

To examine our hypothesis, we used game log data to explore the relationship between a digital scratchpad usage and game performance. Existing work mostly examined the efficacy of digital assistive tools using user testing data with a limited number of students. For example, Couse and Chen [6] observed videos of children using a stylus and a tablet and interviewed them, concluding that the usage of the assistive tools helps support drawing activities. Although such qualitative observations yield valuable insights, analyzing such data requires much time and efforts. By using game data log as a data source, analyzing a larger number of student data is possible.

## 2 Study Design

“Kitkit school” is an educational game, which is designed to help children to practice math and literacy skills in developing countries such as Tanzania and Kenya. The game is designed for children in Kindergarten to grade 2 levels. In this study, the data used for analysis are game log data of children playing “Double digit” game, which is one of

20 math games in Kitkit school. In this game, two numbers are randomly generated and students are asked to add or subtract these numbers. Data was collected from 418 students since May 2018 until November 2018, with an average of about 1,350 number of problems solved per student, totaling 564,251 problems.



**Fig. 1.** Game interface of the Double digit game

The start screen of the Double digit game shows 28 levels with differing difficulty. Figure 1 shows the Double digit game interface after selecting a game level. As a game level increases, students calculate numbers with higher digits. The left part of the screen displays a keypad with an answer slot, and the right part displays an arithmetic problem on a digital scratchpad where students can write or draw. A “touch” event occurs every time pressure on the screen changes as a student touches the digital scratchpad.

In order to compare game performance depending on different cognitive loads, we categorized these 28 game levels into 3 difficulty levels depending on differing cognitive load required in solving a problem. Table 1 shows the 3 difficulty levels with the criteria for each. Note that we labeled the two numbers that are given for addition or subtraction as N1 and N2. Since computation of larger numbers requires more cognitive load, with increased processing needed for more digits and regrouping of numbers, the game levels with larger numbers are labeled as belonging to a higher

**Table 1.** Difficulty level criteria. N1 and N2 are the two numbers in a given problem

Difficulty level: game level	Condition	Sample problem
1: 1 <=Game level < 13	Max(N1) <=20 and Max(N2) <=20	14 + 12
2: 13 <=Game level < 23	50 <=Max(N1) <=99 or 50 <=Max(N2) <=99	78 – 55
3: 23 <=Game level <=28	Max(N1) >=100 and Max(N2) >=100	209 – 100

difficulty level.

## 2.1 Measurement

To test our hypothesis, the correlations between the digital scratchpad usage and game performance are calculated for each of the three difficulty levels. First, a student's digital scratchpad usage and a student's game performance are calculated using formula (1) and (2), respectively. Both variables were calculated for each of the 418 students. Next, a Pearson's correlation between these variables is calculated for difficulty levels 1, 2, and 3.

$$\text{Digital scratchpad usage (\%)} = \frac{\#\text{of touches on the digital scratch pad}}{\#\text{of total answers entered}} \times 100 \quad (1)$$

$$\text{Game performance (\%)} = \frac{\#\text{of correct answers}}{\#\text{of total answers entered}} \times 100 \quad (2)$$

## 3 Results

The Pearson correlations between the digital scratchpad usage and game performance was calculated for three game difficulty levels. Across all three game difficulty levels, the correlation between the digital scratchpad usage and the game performance shows a positive correlation ( $r = 0.152$ ,  $p = 0.005$ ). Thus, we see a generally positive trend showing that a digital scratchpad usage increases as a game performance increases.

The Pearson correlations between two variables for each game difficulty level is consistent with our hypothesis. Namely, for problems in higher game difficulty level that has a higher cognitive load, the correlation between digital scratchpad usage and game performance is also higher. For difficulty level 1, the correlation significant at  $r = 0.122$  ( $p = 0.024$ ). For difficulty level 2, the correlation is also significant at  $r = 0.193$  ( $p = 0.002$ ). Finally, for difficulty level 3, where students have to compute numbers over 100, the correlation between the digital scratchpad usage and game performance is highest at  $r = 0.480$  ( $p = 0.00000$ ).

## 4 Discussions and Conclusion

In this study, we explored whether increased usage of digital scratchpad has positive correlation with game performance as game difficulty level increases. As hypothesized, our data collected from 418 students, show that the relationship between digital scratchpad usage and game performance shows a more positive correlation in games with a higher difficulty level that require a higher cognitive load. The study has limitations in that although we examined the amount of students' digital scratchpads usage, we did not check for the content of the writings, i.e. whether a student solved

math problems or drew meaningless drawings. For future work, user interviews on user experiences and decision-making process will yield additional insights that can explain such content of writings. Despite such limitation, the study provides insights on using large-scale log data to explore the impact of using digital assistive tools on game performance.

**Acknowledgements.** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT): 2017R1D1A1B03034511 & Enuma, Inc.

## References

1. Sweller, J.: Cognitive load during problem solving: effects on learning. *Cogn. Sci.* **12**(2), 257–285 (1988)
2. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* **4** (4), 295–312 (1994)
3. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* **38**(1), 1–4 (2003)
4. Oviatt, S.: Human-centered design meets cognitive load theory: designing interfaces that help people think. In: MM 2006 Proceedings of the 14th ACM International Conference on Multimedia, pp. 871–880. ACM, Santa Barbara (2006)
5. Ando, M., Ueno, A.: Analysis of the advantages of using tablet PC in e-Learning. In: 10th IEEE International Conference on Advanced Learning Technologies, Sousse, pp. 122–124 (2010)
6. Couse, L., Chen, D.: A tablet computer for young children? Exploring its viability for early childhood education. *J. Res. Technol. Educ.* **43**(1), 75–96 (2010)



# What Does Time Tell? Tracing the Forgetting Curve Using Deep Knowledge Tracing

Amar Lalwani<sup>(✉)</sup> and Sweety Agrawal

funtoot, Bangalore, India

[amarlalwani1707@gmail.com](mailto:amarlalwani1707@gmail.com), [sweety.v.agrawal@gmail.com](mailto:sweety.v.agrawal@gmail.com)

**Abstract.** Recurrent Neural Network (RNN) based Deep Knowledge Tracing (DKT) can extract a complex representation of student knowledge just using the historical time series of correct-incorrect responses given as input and can predict the student’s performance on the next problem. funtoot is a personalized and adaptive learning system used by students to practice problems in school and at home. Our analysis of students’ interaction with funtoot showed a time-gap as high as 1 h, 1 day and also 1 week between two problems attempted by a student in a task. In this work, along with the time series of previous correct-incorrect responses, we also encode the time-gap as a feature to investigate its effect on predictions. We call this variant of DKT as DKT-t. We test these models on our dataset and two major publicly available datasets from - Assistments and Carnegie Learning’s Cognitive Tutor and analyze the predicted student knowledge by both the models and report our findings. We also show that DKT-t can help us trace the forgetting curve given various response sequences and their knowledge states.

**Keywords:** Deep Knowledge Tracing · Intelligent tutoring systems · Time · Time-gap · Forgetting curve · funtoot

## 1 Introduction

Models, like Deep Knowledge Tracing (DKT) [6], encode student knowledge as a latent variable along with its temporal dynamics. Based on the progression of a student’s work on the tutor, these models update their estimates of predicted knowledge based on the correctness of student responses.

On tutor *funtoot*, used in school and at home, students might not complete a topic in one session. Even in one session, various factors might hinder students’ engagement with the tutor. For instance, talking to the neighbors/friends. Analysis of the data of students’ interaction on tutor *funtoot* showed that there is a time difference between submitting one problem and generating the next problem. We call this time difference as a *time-gap* on the task. The time-gap between two practice opportunities is as low as 0–2 s and also as high as a week.

Researchers in [3, 9] have extended the Bayesian knowledge tracing [1] (BKT) model and authors of [10] have extended DKT by adding lots of features and improve its predictions. However, none of these three studies have considered the time-gap we are interested in. Authors of [7] have modeled time-gap of a day or more, but in Bayesian Knowledge Tracing.

In this study, we leverage the time-gap in DKT to enable us to predict the delayed performance after the time-gap. We attempt to improve the prediction accuracy of DKT with time-gap and compare it with the DKT model without the time-gap. Using simulation, we also analyze the predicted knowledge following a time-gap and trace the forgetting curves.

## 2 Experiments

We have trained the DKT model as explained in paper [5] with bloom's cognitive level - bloom's taxonomy learning objective (btlo) as a feature (skill). A problem might involve more than one skills. This can also be encoded in DKT as shown in [4].

The DKT model explained in the above papers considers only the skill of a problem and its correctness. In the proposed variant of DKT, we also consider time-gap as a feature. We call this variant as DKT-t. We have identified 9 time-gaps which we model as a feature in DKT. The 9 time-gaps are as follows: Gap#1 - < 2 s; Gap#2 - [2 s – 5 s); Gap#3 - [5 s – 10 s); Gap#4 - [10 s – 30 s); Gap#5 - [30 s – 1 min); Gap#6 - [1 min – 5 min); Gap#7 - [5 min – 1 h); Gap#8 - [1 h – 1 week); Gap#9 - > 1 week. Please note that '[' denotes the inclusion and ')' denotes the exclusion of the respective point in the interval.

We test the performance of DKT and DKT-t on three datasets: funtoot dataset, Cognitive Tutor dataset and Assisments dataset. We used the publicly available **Assisments** dataset<sup>1</sup> [2] which has start-time and end-time of a problem attempt. We chose 12 highest used skills from this dataset and only considered answers to the original problems. It contains 8,97,971 data-points from 7,856 students. One dataset<sup>2</sup> we choose comes from the **Cognitive Tutor** called Algebra I 2005-2006 [8]. We chose 114 units with the prefix 'CTA1' and 'ES'. It contains 72 skills and 5,62,103 data-points generated by 560 students. **funtoot** dataset contains 17 skills and 4,66,212 data-points generated by 8,000 students.

To study the effect of time-gap on the predicted knowledge immediately following a time-gap, we perform simulation using DKT-t model. We pick five most used skills from all the three datasets. For the chosen skills, we predict the response to the next problem for all the time-gaps using DKT-t after solving five problems correctly of the skill.

---

<sup>1</sup> Downloaded from: <https://sites.google.com/site/assistsmentsdata/home/2012-13-school-data-with-affect>.

<sup>2</sup> Downloaded from: <https://pslcdatasetshop.web.cmu.edu/KDDCup/downloads.jsp>.

### 3 Results

The results of the learned models are evaluated and compared by *AUC*, the square of Pearson correlation ( $R^2$ ) and mean error (me). Mean error is the residual error computed as: the mean of the actual performance subtracted by the predicted performance [7].

**Table 1.** Model statistics

Parameter	Model	funtoot			Assistments			Algebra I 2005-2006		
		AUC	$R^2$	me	AUC	$R^2$	me	AUC	$R^2$	me
Overall	DKT	0.762	0.183	-0.003	0.717	0.131	-0.008	0.815	0.272	0.003
	DKT-t	0.76	0.178	-0.01	0.719	0.133	0.002	0.829	0.299	-0.004
Gap#1	DKT	0.745	0.142	0.011	0.732	0.15	0.002	0.821	0.283	0.003
	DKT-t	0.74	0.143	-0.006	0.731	0.15	0.002	0.839	0.318	-0.006
Gap#2	DKT	0.747	0.145	0.001	0.719	0.131	-0.004	0.791	0.216	0.017
	DKT-t	0.743	0.138	-0.007	0.717	0.128	0.004	0.79	0.204	-0.007
Gap#3	DKT	0.756	0.166	-0.002	0.708	0.126	-0.029	0.811	0.234	0.019
	DKT-t	0.753	0.159	-0.018	0.705	0.125	-0.018	0.808	0.231	-0.003
Gap#4	DKT	0.754	0.176	0.01	0.693	0.098	-0.04	0.804	0.253	0.021
	DKT-t	0.75	0.17	-0.016	0.694	0.105	0.001	0.8	0.251	0.008
Gap#5	DKT	0.724	0.146	0.003	0.692	0.101	-0.037	0.773	0.212	0.001
	DKT-t	0.717	0.136	-0.012	0.693	0.104	-0.004	0.769	0.207	0.001
Gap#6	DKT	0.707	0.129	-0.003	0.688	0.097	-0.031	0.748	0.177	-0.002
	DKT-t	0.704	0.125	0.011	0.691	0.102	-0.007	0.753	0.191	0.002
Gap#7	DKT	0.74	0.162	-0.049	0.677	0.086	-0.024	0.73	0.144	-0.016
	DKT-t	0.729	0.157	-0.011	0.677	0.086	0.001	0.731	0.149	-0.006
Gap#8	DKT	0.721	0.141	-0.03	0.675	0.084	-0.017	0.786	0.217	-0.067
	DKT-t	0.717	0.137	-0.017	0.681	0.09	0.012	0.78	0.23	-0.001
Gap#9	DKT	0.741	0.156	-0.056	0.663	0.069	-0.014	0.714	0.078	-0.117
	DKT-t	0.741	0.164	-0.028	0.673	0.078	0.012	0.747	0.172	-0.041

We report these three metrics per gap in DKT and DKT-t and overall to analyze the difference with the time-gap parameter. Table 1 shows the results.

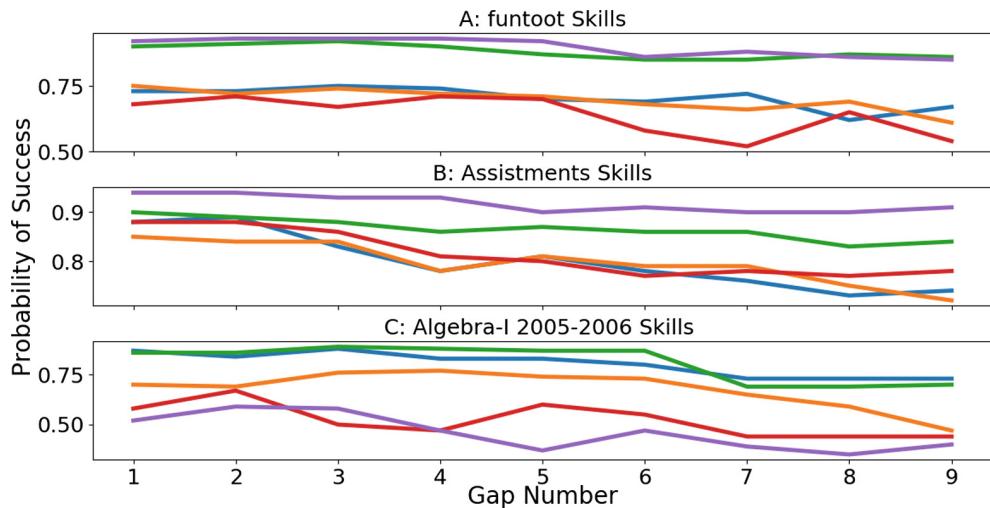
The overall AUC of DKT and DKT-t remained almost same for Assistments and funtoot. There is a minor improvement of 1.72% AUC with DKT-t for dataset Algebra I 2005-2006. For funtoot, considering the time-gaps also the AUC's remain similar for DKT and DKT-t. There is a clear decrease in AUC from gap#1–gap#9 with DKT, and also with DKT-t for datasets - Assistments and Algebra I 2005-2006.

We also observe that there is a decrease in  $R^2$  from gap#1–gap#9 with DKT, and also with DKT-t for datasets - Assistments and Algebra I 2005-2006. For funtoot dataset, DKT highly over-predicts (negative mean error) for gap#7–gap#9 which is reduced to almost half by DKT-t. In Assistments dataset, DKT heavily over-predicted for gap#3–gap#7 while DKT-t almost reduced the mean error to zero. For gap#8–gap#9, DKT moderately over-predicted, whereas, DKT-t moderately under-predicted (positive mean error). DKT under-predicted for gap#2–gap#4, while it is close to zero with DKT-t for Algebra I 2005-2006

dataset. DKT heavily over-predicted for gap#7–gap#9, but the mean error is close to zero with DKT-t.

**Forgetting Curve.** Since we have the estimates of how the students might do following each time-gap through simulations, we can plot a curve of predictions for every time-gap. If the hypothesis that post higher time-gaps, students might forget the learned material, there should be a decline in predictions as the time-gap increases which is called as forgetting curve.

Figure 1 shows the forgetting curves for five most used skills from all the datasets. For datasets funtoot and Algebra I 2005–2006, as shown in Figs. 1A and 1C, either there is a slight increase in predictions for gap#3–gap#5 or they remain steady. However, for Assistments dataset shown in Fig. 1B, there is a steady decline in predictions as the time-gap increases. Across all the three datasets, there is a slight increase in prediction following gap#8.



**Fig. 1.** Forgetting curves for skills from Algebra I 2005-2006 dataset

## 4 Discussion and Conclusion

This work attempts to incorporate and model time-gap into Deep Knowledge Tracing. The predictability of the student performance decreases systematically with the increase in time-gaps for DKT (indicated by AUC and  $R^2$ ) which remains the same even after modeling the time-gap in DKT-t. However,  $R^2$  is comparatively higher with DKT-t than DKT for larger time-gaps. For Algebra I 2005-2006 dataset, the predictability with DKT-t improved by 0.14 AUC units.

Since DKT considers only the ordering of the student responses, we observe that it heavily over-predicts the next response performance following larger time-gaps. DKT-t reduces these residuals between the actual and predicted performances.

The forgetting curves across all the three datasets demonstrate the decay of knowledge as time progresses. The predicted performance for gap#1 is slightly lower than the gap#2–gap#5 for some skills. In gap#1, students might be gaming the system or moving on to the next problem too quickly with little introspection. We need to study and validate this hypothesis in the future work. Additionally, there is a rise in the predicted performance following gap#8. We are not clear about the reason behind this.

One of the main contributions of this work is the unique approach to trace the forgetting curve using the historical student responses generated on the digital tutors. These models have the potential to empower the researchers to simulate various learning scenarios and theories and get the sense of their effects on learning and forgetting.

## References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**(4), 253–278 (1994)
2. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. *User Model. User-Adap. Inter.* **19**(3), 243–266 (2009)
3. Joseph, E.: Engagement tracing: using response times to model student disengagement. In: Artificial intelligence in education: Supporting learning through intelligent and socially informed technology **125**, 88 (2005)
4. Lalwani, A., Agrawal, S.: Few hundred parameters outperform few hundred thousand? In: Educational Data Mining (2017)
5. Lalwani, A., Agrawal, S.: Validating revised bloom’s taxonomy using deep knowledge tracing. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 225–238. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_17](https://doi.org/10.1007/978-3-319-93843-1_17)
6. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems. pp. 505–513 (2015)
7. Qiu, Y., Qi, Y., Lu, H., Pardos, Z.A., Heffernan, N.T.: Does time matter? modeling the effect of time with Bayesian knowledge tracing. In: EDM, pp. 139–148 (2011)
8. Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G., Koedinger, K.: Algebra I 2005–2006. In: Challenge data set from KDD Cup 2010 Educational Data Mining Challenge (2010)
9. Wang, Y., Heffernan, N.T.: Leveraging first response time into the knowledge tracing model. In: International Educational Data Mining Society (2012)
10. Zhang, L., Xiong, X., Zhao, S., Botelho, A., Heffernan, N.T.: Incorporating rich features into deep knowledge tracing. In: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, pp. 169–172. ACM (2017)



# Evaluating the Transfer of Scaffolded Inquiry: What Sticks and Does It Last?

Haiying Li<sup>(✉)</sup>, Janice Gobert, and Rachel Dickler

Rutgers University, New Brunswick, NJ 08901, USA

{haiying.li, janice.gobert,  
rachel.dickler}@gse.rutgers.edu

**Abstract.** The Next Generation Science Standards [1] expect students to master disciplinary core ideas, crosscutting concepts, and scientific practice. In prior work, we showed that students benefited from real time scaffolding of science practices such that students' inquiry competencies both improved over time and transferred to new science topics. The present study examines the robustness of adaptive scaffolding by evaluating students' inquiry performances at a very fine-grained level in order to investigate *what* aspects of inquiry are robust over time once scaffolding was removed. 108 middle school students in grade 6 used Inq-ITS and received adaptive scaffolding for three lab activities in the first inquiry topic they completed (i.e. Animal Cell); they then completed 10 activities without scaffolding across three new topics. Results showed that after removing scaffolding, student's inquiry performance generally improved with slight variations in performance across driving questions and over time. Overall, these findings suggest that adaptive scaffolding may support students' inquiry learning and transfer of inquiry practices over time and across topics.

**Keywords:** Science inquiry · Growth in inquiry performance · Scaffolding

## 1 Introduction

In science inquiry contexts, students require support in order to effectively engage in inquiry investigations [2–4]. Supports provided to students can be in the form of scaffolds designed to help students reach a level of performance that would not be possible if they were to do a task independently [5, 6]. The types of scaffolds students receive within online science environments may vary from fixed [7] to faded [8] to adaptive scaffolds [9]. Fixed scaffolds are supports that are provided to all students consistently, regardless of student performance [7, 10]. Faded scaffolds, on the other hand, are supports that are gradually removed with increasing use of a particular system [8, 10, 11]. Another form of scaffolds are adaptive scaffolds, which are supports that are provided to students in real-time based on students' performance in a system [12, 13]. While fixed [7], faded [8], and adaptive scaffolds [12] have benefited student learning in science environments to some extent, adaptive scaffolds show the greatest promise in terms of promoting transfer of inquiry practices [13, 14] because they provide students with the information they need when they need it most [15].

In the context of science inquiry, transfer of inquiry practices may be assessed in terms of near transfer (i.e. transfer to similar inquiry tasks presented briefly after the initial inquiry task; [16]) or far transfer (i.e. transfer to inquiry tasks in different contexts and after extended periods of time; [16]). Studies have demonstrated how engagement in computer-supported learning environments can promote transfer of science content understandings [17, 18] and practices such as scientific reasoning [16]. In the intelligent tutoring system, Inq-ITS [9], researchers have demonstrated transfer of multiple scientific practices across topics and over time [14] including: hypothesizing [12, 19], collecting data [20, 21], and interpreting data/warranting claims with evidence [22, 23]. Each of these practices can be operationalized into different finer-grained sub-practices. Studies have yet to investigate the transfer of inquiry at the sub-practice level over time and across topics. The present study examines whether adaptive scaffolding of inquiry practices in the first three Inq-ITS activities (i.e. driving questions) leads to transfer of inquiry practices across topics at varying time intervals at the sub-practice level.

## 2 Method

### 2.1 Participants and Materials

The participants in the present study were 108 6<sup>th</sup> grade students from a middle school in the northeastern United States who completed the following Inq-ITS [9] lab activities: *Animal Cell* (three driving questions: (1) how can you increase the transfer or protein in an animal cell?, (2) how you can decrease the production of ribosomes?, and (3) how you can reduce the production of protein?), *Plant Cell* (three driving questions: (1) how can you increase the transfer or protein in a plant cell?, (2) how you can decrease the production of ribosomes?, and (3) how you can reduce the production of protein?), *Genetics* (three driving question activities: how does changing a mother monster's (1) F, (2) L, and (3) H alleles impact the traits of the babies?), and *Natural Selection* (four driving questions: what is the optimal foliage for (1) the green, long furred and (2) the red, short furred monsters?, what is the optimal temperature for (3) the green, short furred and (4) the red, long furred monsters?).

Each of these Inq-ITS activities contained four stages where students first formed a question/hypothesis, carried out an investigation/collected data, analyzed and interpreted data, and finally communicated their findings [9, 10]. Currently, adaptive, real-time scaffolding is available within the first three stages of the microworlds [19–23] (scaffolding is being developed for communicating findings [24]) based on automated scoring in Inq-ITS ([25]; see Measures section). The only difference between adaptive scaffolded and unscaffolded Inq-ITS activities is the presence of the pedagogical agent, Rex. For example, in the scaffolded animal cell activities in the present study, if a student was evaluated as having difficulty on a particular practice, then Rex would pop up on the student's screen with different types of information depending on the student's specific difficulty [26, 27]. Rex would first provide students with an orienting hint reminding the students of the inquiry practice/sub-practice that they were engaging in [28]. If the students continued to have difficulty with the practice, Rex would provide

a procedural hint (explaining the steps involved in the practice/sub-practice) followed by a conceptual hint (explaining the inquiry practice/sub-practice) and finally an instrumental hint (explaining the exact steps).

## 2.2 Measures

In the present study, the dependent variables were four inquiry practices. Each inquiry practice in Inq-ITS is operationalized at a fine-grained level (i.e., broken down into different sub-practices/sub-components). The *hypothesizing practice* was measured by: identifying an independent variable (IV) and dependent variable (DV). The *collecting data practice* was measured by: testing the hypothesis and running targeted and controlled trials. The *interpreting data practice* was measured by: correctly selecting the IV and DV for a claim, correctly interpreting the relationship between the IV and DV, and correctly interpreting the hypothesis/claim relationship. The *warranting claims practice* was measured by: warranting the claim with more than one trial, warranting with controlled trials, correctly warranting the relationship between the IV and DV, and correctly warranting the hypothesis/claim relationship. Each inquiry sub-practice was automatically scored as 0 points if incorrect or 1 point if correct using the knowledge engineering and educational data mining techniques in Inq-ITS, validated in prior studies [9].

This study had a time variable with four levels: Time 1 (i.e., Animal Cell in month 0), Time 2 (i.e., Plant Cell in month 1.3), Time 3 (i.e., Genetics in month 2.7), and Time 4 (i.e., Natural Selection in month 5.7). Moreover, this study had a variable of the number of driving questions that students completed over time: driving questions 1 to 3 in month 0 (i.e., Animal Cell), 4 to 6 in month 1.3 (i.e. Plant Cell), 7 to 9 in month 2.7 (i.e., Genetics), and 10 to 13 in month 5.7 (i.e., Natural Selection).

## 3 Results and Discussion

We used linear mixed models (LMMs) to investigate whether there was evidence of transfer by evaluating students' inquiry competencies across driving questions over time after removing the adaptive scaffolding. We performed four sets of LMM analyses where we focused on the pattern within each inquiry practice.

### 3.1 Model Selection

For the analysis of the data, we followed the “top-down” modeling strategy and selected the models that best fit the data. We ran an unconditional model with intercepts only, and then added each variable independently as well as in combination. Each type of added variable(s) generated three models based on the variation of random effects: subjects only (Intercept), the number of driving questions and/or time variable(s) only (Slope), or both subjects and the number of driving questions and/or time variable(s). We compared the models using the  $-2$  Restricted Log Likelihood ( $-2\text{RLL}$ ) [29] and selected the full models in this study due to their best fit for a greater number of practices (namely, hypothesis, data collection, and warranting claims).

### 3.2 Performance Across Driving Questions and Over Time

We then examined inquiry scores across driving questions and time for each practice. Results showed that the fixed effects for the hypothesizing practice were significant,  $F(1, 108.25) = 24.39, p < .001$  for driving questions and  $F(1, 107.25) = 11.32, p = .001$  for time. Fixed-effects parameters were significant for hypothesizing ( $\beta = 0.03, p < .001$  for driving question;  $\beta = -0.04, p = .001$  for time), collecting data ( $\beta = 0.05, p < .001$  for driving question;  $\beta = -0.06, p < .001$  for time), and warranting claims practices ( $\beta = 0.04, p < .001$  for driving question;  $\beta = -0.05, p < .001$  for time). These results indicate that students improved their performance on these three inquiry practices with the increasing use of Inq-ITS, but that the long-time intervals between usage resulted in a slight decrease in performance. This pattern was not found for the practice of interpreting data, potentially due to students starting with relatively high performance (Mean = 0.79) or interactions with topic complexity [30].

The random effects showed a significant intercept for the hypothesizing ( $\beta = 0.03, Z = 3.17, p < .01$ ), collecting data ( $\beta = 0.05, Z = 3.87, p < .001$ ), and warranting claims practice ( $\beta = 0.06, Z = 4.18, p < .001$ ). Results also showed a significant driving question random effect for hypothesizing ( $\beta = 0.001, Z = 1.97, p < .05$ ) and collecting data ( $\beta = 0.002, Z = 2.00, p < .05$ ). Additionally, in hypothesizing, we found a significant driving question and time random effect ( $\beta = -0.002, Z = -2.03, p < .05$ ) and significant time effect ( $\beta = 0.004, Z = 2.26, p < .05$ ). We also found a significant covariance between the intercept and the driving question coefficient for collecting data ( $\beta = -0.01, Z = -2.01, p < .05$ ). The findings of these random effects confirmed a fair amount of student-to-student variation in the starting performance for practices of hypothesizing, collecting data, and warranting claims, but varied patterns for driving question, time, and both driving question and time effects. This demonstrates that transfer of learning was different for different inquiry practices for students.

## 4 Conclusions, Future Directions, and Implications

In this study we investigated the robustness of our scaffolding using students' performances on various inquiry practices across driving questions at different time intervals, thereby addressing near (across driving questions at each time) and far transfer (over time). Our results showed, in general, that our scaffolding was robust for practices of hypothesizing, collecting data, and warranting claims. A limitation of the present study is that there was no control condition, which makes it challenging to distinguish between effects of external factors such as teacher instruction between usage of the system. In the future it will be valuable to examine differences between students in a scaffolded and unscaffolded condition to more fully understand the influence of the adaptive scaffolds in Inq-ITS on students' inquiry performance.

Overall, the findings in the present study inform assessment designers and researchers that, if properly designed, scaffolding aimed at supporting students' competencies at various inquiry practices can greatly benefit students' deep learning of, transfer of, and performance on inquiry practices over time.

## References

1. Next Generation Science Standards Lead States: Next Generation Science Standards: For States, by States. National Academies Press, Washington (2013)
2. Hmelo-Silver, C.E., Duncan, R.G., Chinn, C.A.: Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* **42**, 99–107 (2007)
3. Kang, H., Thompson, J., Windschitl, M.: Creating opportunities for students to show what they know: the role of scaffolding in assessment tasks. *Sci. Educ.* **98**, 674–704 (2014)
4. McNeill, K.L., Krajcik, J.S.: Supporting grade 5–8 students in constructing explanations in science: the claim, evidence, and reasoning framework for talk and writing. Pearson (2011)
5. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1978)
6. Quintana, C., et al.: A scaffolding design framework for software to support science inquiry. *J. Learn. Sci.* **13**, 337–386 (2004)
7. Tabak, I., Reiser, B.J.: Software-realized inquiry support for cultivating a disciplinary stance. *Pragmat. Cogn.* **16**, 307–355 (2008)
8. van Joolingen, W.R., de Jong, T., Lazonder, A.W., Savelsbergh, E.R., Manlove, S.: Co-Lab: research and development of an online learning environment for collaborative scientific discovery learning. *Comput. Hum. Behav.* **21**, 671–688 (2005)
9. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S.: From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *J. Learn. Sci.* **22**, 521–563 (2013)
10. McNeill, K.L., Lizotte, D.J., Krajcik, J., Marx, R.W.: Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *J. Learn. Sci.* **15**, 153–191 (2006)
11. Martin, N.D., Tissenbaum, C.D., Gnesdilow, D., Puntambekar, S.: Fading distributed scaffolds: the importance of complementarity between teacher and material scaffolds. *Instr. Sci.* **47**, 1–30 (2018)
12. Gobert, J.D., Moussavi, R., Li, H., Sao Pedro, M., Dickler, R.: Real-time scaffolding of students' online data interpretation during inquiry with Inq-ITS using educational data mining. In: Auer, M.E., Azad, A.K.M., Edwards, A., de Jong, T. (eds.) *Cyber-Physical Laboratories in Engineering and Science Education*, pp. 191–217. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-76935-6\\_8](https://doi.org/10.1007/978-3-319-76935-6_8)
13. Noroozi, O., Kirschner, P.A., Biemans, H.J., Mulder, M.: Promoting argumentation competence: extending from first-to second-order scaffolding through adaptive fading. *Educ. Psychol. Rev.* **30**, 1–24 (2017)
14. Li, H., Gobert, J., Dickler, R.: Testing the robustness of inquiry practices once scaffolding is removed. Submitted to: Intelligent Tutoring Systems (submitted)
15. Koedinger, K.R., Anderson, J.R.: Illustrating principled design: the early evolution of a cognitive tutor for algebra symbolization. *Interact. Learn. Environ.* **5**, 161–180 (1998)
16. Chen, Z., Klahr, D.: Remote transfer of scientific-reasoning and problem-solving strategies in children. In: *Advances in Child Development and Behavior*, pp. 419–470. JAI (2008)
17. Borek, A., McLaren, B.M., Karabinos, M., Yaron, D.: How much assistance is helpful to students in discovery learning? In: Cress, U., Dimitrova, V., Specht, M. (eds.) *EC-TEL 2009. LNCS*, vol. 5794, pp. 391–404. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04636-0\\_38](https://doi.org/10.1007/978-3-642-04636-0_38)
18. Tao, P.K., Gunstone, R.F.: The process of conceptual change in force and motion during computer-supported physics instruction. *J. Res. Sci. Teach.* **36**, 859–882 (1999)

19. Gobert, J.D., Sao Pedro, M.A., Baker, R.S., Toto, E., Montalvo, O.: Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *J. Educ. Data Min.* **4**, 111–143 (2012)
20. Sao Pedro, M.: Real-Time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills Within Physical Science Simulations. Worcester Polytechnic Institute, Worcester (2013)
21. Sao Pedro, M., Baker, R., Gobert, J.: Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In: Proceedings of the 6th International Conference on Educational Data Mining, pp. 185–192. EDM Society (2013)
22. Moussavi, R.: Design, Development, and Evaluation of Scaffolds for Data Interpretation Practices During Inquiry. Worcester Polytechnic Institute, Worcester (2018)
23. Moussavi, R., Gobert, J., Sao Pedro, M.: The effect of scaffolding on the immediate transfer of students' data interpretation skills within science topics. In: Proceedings of the International Conference of the Learning Sciences, pp. 1002–1005. Scopus, Ipswich (2016)
24. Li, H., Gobert, J., Dickler, R.: Automated assessment for scientific explanations in on-line science inquiry. In: Hu, X., Barnes, T., Hershkovitz, A., Paquette, L. (eds.) *Proceedings of the Conference on Educational Data Mining*, pp. 214–219. EDM Society, Wuhan (2017)
25. Gobert, J.D., Baker, R.S., Sao Pedro, M.A.: Inquiry skills tutoring system. U.S. Patent No. 9,373,082. U.S. Patent and Trademark Office, Washington, DC (2016)
26. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. *J. Learn. Sci.* **4**(2), 167–207 (1995)
27. Koedinger, K.R., Corbett, A.: Cognitive tutors: technology bringing learning sciences to the classroom. In: *The Cambridge Handbook of the Learning Sciences*, pp. 61–77. Cambridge University Press, New York (2006)
28. de Jong, T.: Computer simulations: technological advances in inquiry learning. *Science* **312**, 532–533 (2006)
29. West, B.T., Welch, K.B., Galecki, A.T.: Linear mixed model. Chapman Hall/CRC, Boca Raton (2007)
30. Li, H., Gobert, J., Dickler, R., Moussavi, R.: The impact of multiple real-time scaffolding experiences on science inquiry practices. In: Nkambou, R., Azevedo, R., Vassileva, J. (eds.) *ITS 2018. LNCS*, vol. 10858, pp. 99–109. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91464-0\\_10](https://doi.org/10.1007/978-3-319-91464-0_10)



# Automatic Short Answer Grading via Multiway Attention Networks

Tiaoqiao Liu<sup>1</sup>, Wenbiao Ding<sup>1</sup>, Zhiwei Wang<sup>2</sup>, Jiliang Tang<sup>2</sup>,  
Gale Yan Huang<sup>1</sup>, and Zitao Liu<sup>1(✉)</sup>

<sup>1</sup> TAL AI Lab, Beijing, China

{liutianqiao,dingwenbiao,galehuang,liuzitao}@100tal.com

<sup>2</sup> Data Science and Engineering Lab, Michigan State University, East Lansing, USA  
{wangzh65,tangjili}@msu.edu

**Abstract.** Automatic short answer grading (ASAG), which autonomously score student answers according to reference answers, provides a cost-effective and consistent approach to teaching professionals and can reduce their monotonous and tedious grading workloads. However, ASAG is a very challenging task due to two reasons: (1) student answers are made up of free text which requires a deep semantic understanding; and (2) the questions are usually open-ended and across many domains in K-12 scenarios. In this paper, we propose a generalized end-to-end ASAG learning framework which aims to (1) autonomously extract linguistic information from both student and reference answers; and (2) accurately model the semantic relations between free-text student and reference answers in open-ended domain. The proposed ASAG model is evaluated on a large real-world K-12 dataset and can outperform the state-of-the-art baselines in terms of various evaluation metrics.

## 1 Introduction

Assessing the knowledge acquired by students is one of the most important aspects of the learning process as it provides feedback to help students correct their misunderstanding of knowledge and improves their overall learning performance. Traditionally, the assessing paradigm is often conducted by instructors or teachers. However, this access paradigm is not suitable in many cases especially when teaching resources are not readily available. To address this gap, many computer-assisted assessment approaches are developed to automate the assessment process [1].

One specific task, automatic short answer grading (ASAG), whose objective is to automatically score the free-text answers from students according to the corresponding reference answer [9], has attracted great attentions from a variety of research communities and some promising results have been already obtained [5, 7–10]. However, ASAG still remains challenging mainly for two reasons. Firstly, the student answers are expressed in different ways of free texts.

---

Z. Wang—Work was done when the authors did internship in TAL AI Lab.

Thus, it requires the ASAG approach to have a deep semantic understanding of the student answers. Secondly, the questions or assessments (and the corresponding reference answers) usually are open-ended and across different domains. The ASAG approach should be general and applicable into different scenarios.

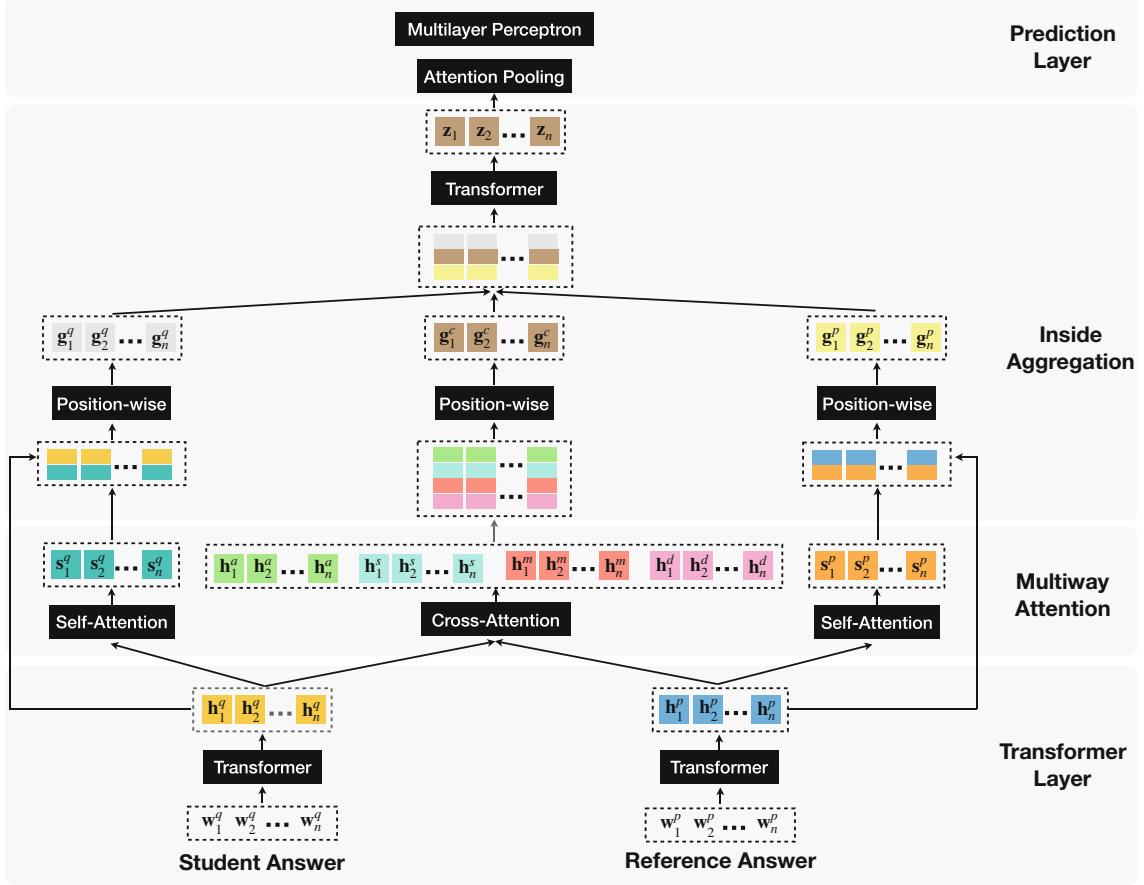
In this paper, to address challenges above, we take the advantage of recent advances in natural language processing field [2, 12] and propose a deep learning framework to tackle the ASAG problem in an end-to-end approach. Specifically, our framework utilizes attention mechanisms to understand the semantics of student and reference answers with most relevant information and is very flexible and efficient as it can be easily extended with extra neuron layers while still maintaining fast training speed thanks to its high parallelization ability. Our main contributions are summarized as follows: (1) We propose an end-to-end approach that does not require any feature engineering effort to tackle the short answer grading problem; (2) We develop a novel framework that is able to modeling the relation between student and reference answers by accurately identifying matching information and understanding the semantic meaning; and (3) The proposed framework can be used in a wide range of domains and is easily scalable for large-scale datasets. It is demonstrated on a large-scale real-world dataset collected from millions of K-12 students.

## 2 Our Approach

In this section, we introduce our proposed framework, the overall structure is shown in Fig. 1. Before detailing each component next, we first introduce the notations. We use bold lower case letters for vectors and bold upper case letters for matrices. We use subscript to represent the vector index, which is the index of word in each sentence in most cases. We also use superscript to represent the category of vectors.

**Transformer Layer.** The input of the transformer layer is the student and reference answer, which are two sequences of words and denoted as  $\{\mathbf{w}_1^q, \mathbf{w}_2^q, \dots, \mathbf{w}_n^q\}$  and  $\{\mathbf{w}_1^p, \mathbf{w}_2^p, \dots, \mathbf{w}_n^p\}$ , respectively, where  $\{\mathbf{w}_i^q\}$  and  $\{\mathbf{w}_i^p\}$  are the pre-trained word embeddings. Next, the *transformer* [12] model is applied as:  $\{\mathbf{h}_1^*, \mathbf{h}_2^*, \dots, \mathbf{h}_n^*\} = \text{transformer}(\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_n^*)$ , where  $* \in \{p, q\}$  and each  $\{\mathbf{h}_i^q\}$  and  $\{\mathbf{h}_i^p\}$  are the word embeddings that contain its contextual sentence information in the student and reference answers, respectively.

**Multiway Attention.** We design the multiway attention layer to capture the relations between student and reference answers. Specifically, it consists of two blocks. The first is self-attention block where each  $\mathbf{h}_i^*$  will attend every  $\mathbf{h}_j^*, j \in \{1, 2, \dots, n\}$  to obtain new representation  $\mathbf{s}_i^*, * \in \{p, q\}$ . The second is cross-attention block in which each  $\mathbf{h}_i^q$  will attend every  $\mathbf{h}_j^p, j \in \{1, 2, \dots, n\}$  to obtain another set of new representations  $\mathbf{h}_i^t, t \in \{a, s, m, d\}$ , where  $a, s, m, d$  are additive, subtractive, multiplicative, and dot-product attention mechanisms, respectively [11].



**Fig. 1.** The overview of our model (better viewed in color).

**Inside Aggregation.** This layer is designed to aggregate multiway attention layer outputs to a single representation  $\mathbf{z}$ . Specifically, we first concatenate the outputs from cross-attention and self-attention blocks by positions respectively and feed them to different position-wise feed forward networks to obtain the compressed representations  $\mathbf{g}_i^*, * \in \{p, q, c\}$ , where  $p, q, c$  represent student answer sequence, reference answer sequence, and cross-attention sequence, respectively. We concatenate the outputs  $\mathbf{g}_i^*$  by positions and after another Transformer block, we get new sequence representation  $\mathbf{Z} = \text{transformer}([\mathbf{g}_1^p, \mathbf{g}_1^q, \mathbf{g}_1^c]), i \in \{1, 2, \dots, n\}$  which contains the information in student and reference answers and the relations between them.

**Prediction Layer.** The evaluation of student answer will be produced by this layer. Specifically, we first convert the aggregated sequence representation  $\mathbf{Z}$  to a fixed-length vector with self-attention pooling layer. This transformation is defined as:  $\mathbf{x} = \text{softmax}(\mathbf{w}_1^{\mathbf{z}} \tanh(\mathbf{W}_2^{\mathbf{z}} \mathbf{Z}^T)) \mathbf{Z}$ , where  $\mathbf{w}_1^{\mathbf{z}}$  and  $\mathbf{W}_2^{\mathbf{z}}$  are learned parameters during training step. Then we build a feed forward network that takes  $\mathbf{x}$  as input and outputs a two-dimensional vector. The output vector is sent to a softmax function to obtain the final probabilistic evaluation vector. The first entry gives the probability of wrong answer while the second entry

gives right answer probability. The objective is to minimize the cross entropy of the relevance labels.

### 3 Experiments

In this section, we conduct experiments on a large real-world educational data, which contains 120,000 pairs of student answers and question analysis from an online education platform, each labeled with binary value indicating whether the student has the right answer. The positive and negative instances are balanced and we randomly select 30,000 samples as our test data and use the rest for validation and training. The hyperparameters of our model are selected by internal cross validation. We use both AUC and accuracy as our evaluation metrics and for both metrics, a higher value indicates better performance.

We compare our model with several state-of-the-art baselines. More specifically, we choose: (1) Logistic regression (LR). (2) Gradient boosted decision tree (GBDT) [3, 13]. (3) Multichannel convolutional neural networks (TextCNN) [4]. (4) Sentence embedding by Bidirectional Transformer block (Bi-Transformer) [12]. (5) Multiway Attention Network (MAN) [11]. And (6) Manhattan LSTM with max pooling (MaLSTM) [6].

#### 3.1 Experimental Results

We report the experimental results in Table 1. From the table, we observe that our model outperforms all of the baselines. We argue that this is because our model is able to effectively capture the semantic information between student and reference answers. This is confirmed by the fact that MAN shows the superior performance among all baselines, as it not only aggregates sentence information within Transformer block, but matches words in both query sentence and answer sentence from multiple attention functions.

**Table 1.** ASAG performance comparison on a real-world K-12 dataset.

	LR	GBDT	TextCNN	Bi-Transformer	MaLSTM	MAN	Our
Accuracy	0.8297	0.8628	0.8772	0.8813	0.8825	0.8808	<b>0.8899</b>
AUC	0.8808	0.9287	0.9312	0.9335	0.9375	0.9365	<b>0.9444</b>

### 4 Conclusion

In this paper we present our multi-way attention network for automatic short answer grading. We use transformer blocks and attention mechanisms to extract answer matching information. To comprehensively capture the semantic relations between the reference answer and the student answers, we apply multiway

attention functions instead of single attention channel. Experiment results on a large real-world education dataset demonstrate the effectiveness of the proposed framework. There are several directions that need further exploration. We may use one attention mechanism with multiple heads instead of multiple attention mechanisms and we may replace transformer block with other type of sentence encoder like self-attention network or hierarchical attention network.

## References

1. Daradoumis, T., Bassi, R., Xhafa, F., Caballé, S.: A review on massive e-learning (MOOC) design, delivery and assessment. In: 2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pp. 208–213. IEEE (2013)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
3. Friedman, J.H.: Stochastic gradient boosting (1999)
4. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
5. Mitchell, T., Russell, T., Broomhead, P., Aldridge, N.: Towards robust computerised marking of free-text responses (2002)
6. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: AAAI, vol. 16, pp. 2786–2792 (2016)
7. Nielsen, R.D., Ward, W., Martin, J.H.: Recognizing entailment in intelligent tutoring systems. Nat. Lang. Eng. **15**(4), 479–501 (2009)
8. Ramachandran, L., Cheng, J., Foltz, P.: Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 97–106 (2015)
9. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhwatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: use both. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 503–517. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_37](https://doi.org/10.1007/978-3-319-93843-1_37)
10. Sultan, M.A., Salazar, C., Sumner, T.: Fast and easy short answer grading with high accuracy. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1070–1075 (2016)
11. Tan, C., Wei, F., Wang, W., Lv, W., Zhou, M.: Multiway attention networks for modeling sentence pairs. In: IJCAI, pp. 4411–4417 (2018)
12. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
13. Ye, J., Chow, J.H., Chen, J., Zheng, Z.: Stochastic gradient boosted distributed decision trees. In: CIKM 2009 (2009)



# Automatic Classification of Error Types in Solutions to Programming Assignments at Online Learning Platform

Artyom Lobanov<sup>1,2</sup>(✉), Timofey Bryksin<sup>1,3</sup>, and Alexey Shpilman<sup>1,2</sup>

<sup>1</sup> JetBrains Research, Saint Petersburg, Russia  
[alexey@shpilman.com](mailto:alexey@shpilman.com)

<sup>2</sup> Higher School of Economics, Saint Petersburg, Russia  
[avlobanov@edu.hse.ru](mailto:avlobanov@edu.hse.ru)

<sup>3</sup> Saint Petersburg State University, Saint Petersburg, Russia  
[t.bryksin@spbu.ru](mailto:t.bryksin@spbu.ru)

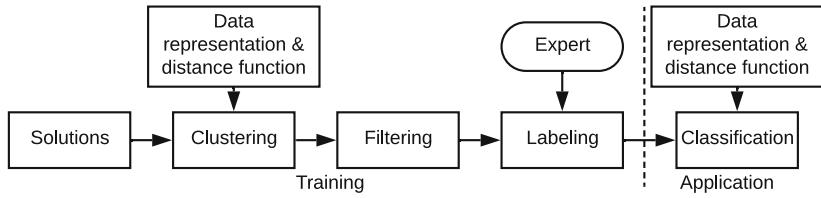
**Abstract.** Online programming courses are becoming more and more popular, but they still have significant drawbacks when compared to the traditional education system, e.g., the lack of feedback. In this study, we apply machine learning methods to improve the feedback of automated verification systems for programming assignments. We propose an approach that provides an insight on how to fix the code for a given incorrect submission. To achieve this, we detect frequent error types by clustering previously submitted incorrect solutions, label these clusters and use this labeled dataset to identify the type of an error in a new submission. We examine and compare several approaches to the detection of frequent error types and to the assignment of clusters to new submissions. The proposed method is evaluated on a dataset provided by a popular online learning platform.

**Keywords:** MOOC · Automatic evaluation · Clustering · Classification · Programming

## 1 Introduction

Recently more and more people get additional education through massive online open courses (MOOC), including programming courses. They are very convenient for students, but you get less feedback on what you are doing wrong since the solutions are usually checked using an automated verification system. In our study, we propose an automatic data-driven method for error type classification that can be used to provide hints for students, rather than just inform them on whether or not their submission has passed all the necessary tests.

The main idea of the proposed approach is to automatically identify and recognize the most common errors through identifying **edit scripts** and analyze these edits through clustering. We use expert evaluation to assign error types to clusters. The general pipeline for the process can be seen in Fig. 1.



**Fig. 1.** General pipeline of the proposed approach.

## 2 Related Work

Several papers have investigated topics related to our task, namely representing code changes in a vector form, their clustering, and classification.

Falleri et al. [1] introduced an approach and a tool they called GumTree to generate **edit scripts**: sequences of atomic abstract syntax tree (AST) modifications that turn a source tree into a target tree. Today, GumTree is considered to be a state-of-the-art tool for generating edit scripts and is widely used in different research tasks.

The task of clustering source code changes based on edit scripts was studied in [5]. One of the considered approaches is similar to the one used in our work: edit scripts were generated using ChangeDistiller [2] (a predecessor of GumTree) and a similarity-based clustering algorithm was applied to them.

Another related task is the classification of code changes. In [8], the authors try to detect code changes that are likely to be specific types of refactorings. They define heuristic rules to define each refactoring type. This kind of approach could not be applied to our case because the number of possible errors and their types are not known beforehand. In [4], the authors used an SVM classifier and features such as commit's metadata, complexity metrics and bag-of-words of the changed code to identify commits that are likely to introduce new bugs. The authors treated code as text using bag-of-words models, whereas working with an AST usually gives more useful information. The somewhat similar idea was implemented in [3], where features of AST changes were used for bug prediction. Their experiments also confirmed that using AST features rather than text-based ones yields better results.

One recent study [9] provides an alternative way to represent edit scripts. The authors employ a deep learning approach to generate a vector of features (embedding) for the edit scripts.

## 3 Overview of the Approach

### 3.1 Dataset

The dataset is provided by Stepik<sup>1</sup> and consists of submitted solutions in Java with their metadata, including verification result. We follow the assumption that

---

<sup>1</sup> Stepik MOOC platform: <http://stepik.org/>.

between the first correct solution and the previous (incorrect) one the user fixed a mistake, therefore changes between these versions contain information about a correctable error.

The dataset consists of  $\{\text{incorrect submission}, \text{correct submission}\}$  pairs for 2 tasks: 1472 pairs for the problem A (a Java Stream API problem) and 8294 pairs for the problem B (checking double values for equality). The dataset was divided into train/validation/test subsets: 1176/148/148 pairs for the problem A and 6588/200/200 pairs for the problem B respectively.

### 3.2 The Pipeline

The pipeline is divided into two stages: training and application. At the training stage, we try to find the most common errors in the incorrect solutions database. To achieve this goal, we cluster edit scripts for incorrect solutions. Edit scripts for the same error are expected to fall into the same cluster. At the next stage, labeled clusters are used to create a classifier. This classifier outputs a type for a new error if it falls into one of the identified clusters, and labels this error as “unknown” otherwise.

To generate edit scripts we used the GumTree library. Since users can fix errors differently, at the clustering step we calculate edit scripts between an incorrect solution and all the correct solutions in the dataset and select the shortest edit scripts. The same procedure is used when we try to identify edit scripts for new incorrect solutions at the classification step.

In this paper, we used classification and clustering algorithms that require only a distance function defined between data points. We considered the following distance metrics for edit scripts:

1. several modifications of the Jaccard similarity coefficient depending on the definition of equality of atomic changes in edit scripts;
2. cosine similarity for the bag-of-words model;
3. cosine similarity for the autoencoder embeddings of the edit scripts.

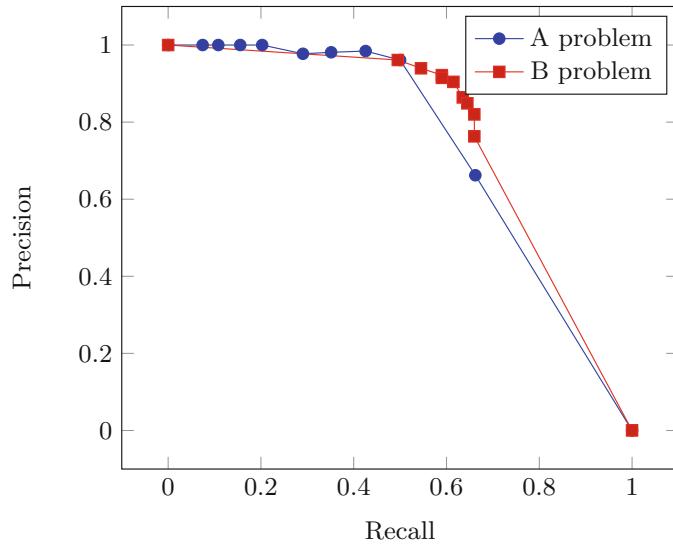
We use Hierarchical Agglomerative Clustering (HAC) [10] to cluster all solutions edit scripts using one of the distance functions described above. After the clustering is complete, clusters smaller than a certain threshold are removed and others are presented to experts, who label them according to the error type.

When classifying a new incorrect solution, we find the nearest correct one and classify the obtained edit script. The easiest way to choose the right cluster is to find the nearest one. As an alternative, we used the k-nearest neighbors (kNN) [6] method with weighted voting. Since the new object may not belong to any cluster, we provide a user with a hint only if we are sure of the classification accuracy.

## 4 Evaluation

To evaluate various clustering and classification algorithms and their parameters we used the area under the precision-recall curve (PR-AUC) since classification should be tuned according to the particular goals.

In all our experiments we used the same clustering algorithm, but, depending on the values of the hyperparameters, we obtained 96 different clustering patterns for each problem. We don't have a proper way to evaluate the quality of clusters themselves, so we compare the quality of the final classification. The validation dataset was used to compare the quality of approaches and find the best one. All in all, we evaluated 27648 combinations of different parameters. Then, best configurations were applied to the test dataset to get an independent assessment. Approaches based on the cosine similarity of the bag-of-words model and the autoencoder embeddings of the edit scripts demonstrated the best results on our dataset for problems A and B respectively. PR-curves for these classifiers are shown in Fig. 2.



**Fig. 2.** Precision-recall curve for best classifiers for test problems A and B.

## 5 Conclusion

In this paper, we present a method for the automatic classification of error types in solutions to programming assignments at an online learning platform. It is based on a notion of an edit script: we cluster these edit scripts at the training stage and classify newly submitted incorrect solutions according to these clusters. Manual labeling of these clusters allows us to provide users with a hint containing an error description. We provide an extensive evaluation of the proposed approach using various clustering methods, edit scripts representations and distance metrics. The evaluation shows that this approach could be successfully implemented in online programming courses at scale.

For the future work, we consider more advanced techniques for embedding atomic changes and edit scripts, e.g., RNN [7]. It is also worthwhile to study the proposed method on a larger number of different problems and to identify characteristics of the dataset that could improve the quality of this approach.

## References

1. Falleri, J.R., Morandat, F., Blanc, X., Martinez, M., Monperrus, M.: Fine-grained and accurate source code differencing. In: ASE (2014)
2. Gall, H.C., Fluri, B., Pinzger, M.: Change analysis with evolizer and changedistiller. *IEEE Softw.* **26**, 26–33 (2009)
3. Giger, E., Pinzger, M., Gall, H.C.: Comparing fine-grained source code changes and code churn for bug prediction. In: Proceedings of the 8th Working Conference on Mining Software Repositories, MSR 2011, pp. 83–92. ACM, New York (2011). <https://doi.org/10.1145/1985441.1985456>
4. Kim, S., James Whitehead, E., Zhang, Y.: Classifying software changes: clean or buggy? *IEEE Trans. Softw. Eng.* **34**, 181–196 (2008)
5. Kreutzer, P., Dotzler, G., Ring, M., Eskofier, B.M., Philippse, M.: Automatic clustering of code changes. In: 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR), pp. 61–72 (2016)
6. Larose, D.T., Larose, C.D.: Discovering Knowledge in Data: An Introduction to Data Mining. Wiley, Hoboken (2014)
7. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning. arXiv preprint [arXiv:1506.00019](https://arxiv.org/abs/1506.00019) (2015)
8. Weißgerber, P., Diehl, S.: Identifying refactorings from source-code changes. In: 21st IEEE/ACM International Conference on Automated Software Engineering (ASE 2006), pp. 231–240 (2006)
9. Yin, P., Neubig, G., Allamanis, M., Brockschmidt, M., Gaunt, A.L.: Learning to represent edits. In: International Conference on Learning Representations (2019). <https://openreview.net/forum?id=BJl6AjC5F7>
10. Zaki, M.J., Meira, W.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, Cambridge (2014)



# Using Recurrent Neural Networks to Build a Stopping Algorithm for an Adaptive Assessment

Jeffrey Matayoshi<sup>(✉)</sup>, Eric Cosyn, and Hasan Uzun

McGraw-Hill Education/ALEKS Corporation, Irvine, CA, USA  
[{jeffrey.matayoshi,eric.cosyn,hasan.uzun}@aleks.com](mailto:{jeffrey.matayoshi,eric.cosyn,hasan.uzun}@aleks.com)

**Abstract.** ALEKS (“Assessment and LEarning in Knowledge Spaces”) is an adaptive learning and assessment system based on knowledge space theory. In this work, our goal is to improve the overall efficiency of the ALEKS assessment by developing an algorithm that can accurately predict when the assessment should be stopped. Using data from more than 1.4 million assessments, we first build recurrent neural network classifiers that attempt to predict the final result of each assessment. We then use these classifiers to develop our stopping algorithm, with the test results indicating that the length of the assessment can potentially be reduced by a large amount while maintaining a high level of accuracy.

**Keywords:** Recurrent neural networks · Adaptive assessment · Knowledge space theory · Deep learning

## 1 Introduction and Background

ALEKS (“Assessment and LEarning in Knowledge Spaces”) is a web-based, artificially intelligent system [17] based on knowledge space theory (KST) [5–7]. The foundation of ALEKS is an adaptive assessment that aims to precisely and efficiently identify the topics in an academic course that a student knows. ALEKS Placement, Preparation and Learning (ALEKS PPL) is a specialized product that has been developed to offer recommendations for placing students in post-secondary mathematics courses.

Deep learning has recently achieved dramatic successes in various fields [14] and is beginning to move into the education domain. In particular, because of the sequential nature of many types of educational data, recurrent neural networks (RNNs) are appearing more frequently in the educational literature [1, 11–13, 15, 18, 21]. Our goal is to augment the performance and efficiency of the KST-powered adaptive assessment algorithm of ALEKS PPL with the classification strengths of RNN models.

In KST, an *item* is a problem type that tests a discrete unit of the curriculum. A *knowledge state* is a set of items that a student masters, and a *knowledge space* is the collection of all such feasible knowledge states. At all times in an ALEKS

PPL assessment, the 314 items under consideration are partitioned into the following categories:

- items that are most likely in the student’s knowledge state (in-state);
- items that are most likely not in the student’s knowledge state (out-of-state);
- the remaining items (uncertain).

The assessment terminates when either (a) there are no remaining “uncertain” items, or (b) the predetermined limit of 29 questions is reached.<sup>1</sup> The assessment then returns the in-state items as its best estimate of the student’s knowledge state. Most ALEKS PPL assessments reach the maximum limit of 29 questions and thus end with a number of “uncertain” items. The *percentage score* of the student is simply the percentage of the 314 items that are categorized as being in-state. Based on the value of the percentage score at the end of the assessment, ALEKS PPL recommends placement in one of six different mathematics courses (see [4] for further details and background on ALEKS PPL).

## 2 Experimental Setup and Models

The data for our experiments consist of 1,449,625 full-length (i.e., 29 question) ALEKS PPL assessments, with each assessment being taken by a unique student for placement purposes in a college or university setting. We use 50,000 assessments for a held-out test set, another 50,000 for a validation set to tune hyperparameters and compare several models, and the remainder (1,349,625) for training our models. Each assessment generates a sequence of inputs, and the target (ground truth) label for each sequence is determined by the course placement recommendation made by the ALEKS system using all 29 questions from the assessment. Thus, the results of the ALEKS PPL assessment can be viewed as a multiclass classification problem with six different class labels, one for each of the possible course placement recommendations.

For our RNN models, we use two different recurrent units: gated recurrent units (GRU) [2] and long short-term memory (LSTM) units [9]. We include both models in our experiments since there currently is not a consensus that one architecture or the other gives superior performance, as several studies have not revealed a clear winner; these include studies both within the education domain [1, 12], as well as from the broader AI community [3, 22]. Additionally, as a comparison, we also build a set of logistic regression classifiers.

Our models will use the actual item categorizations of the ALEKS assessment as features. Thus, we require  $3 \times 314 = 942$  independent variables to represent all

---

<sup>1</sup> Students actually answer up to 30 questions when accounting for a randomly chosen question that is used for validation and other statistics. This number of questions balances the need to gather enough information about the student’s knowledge state against the possibility of overwhelming the student. Regarding the latter concern, see [16] for evidence of a “fatigue effect” experienced by students in ALEKS assessments.

possible combinations of assessment categories (in-state, out-of-state, and uncertain) and items. The  $n$ -th vector of each sequence contains the categorization of the items by the assessment after question  $n$ .

For the LSTM and GRU models, the number of hidden layers, the sizes of the hidden layers, and the learning rate are tuned on the validation set. We also use batch normalization [10] and, to help prevent overfitting, early stopping [19] and dropout [8, 20]. For the logistic regression models, the only tuned hyperparameter is the strength of the L2 regularization.

### 3 Stopping Algorithm and Model Evaluation

The best performing models on the validation set are used to implement our stopping algorithm for the ALEKS assessment. As shown in Algorithm 1, our first criterion is that the most confident predicted class label is above a certain threshold,  $\alpha$ . Additionally, we require that the course placement recommendation at the current question (as determined by the student's percentage score at that point in the assessment) matches the classifier's predicted class label, and we also require that the assessment has asked at least 10 questions (to ensure that our classifier has a minimal amount of data to work with).

---

#### Algorithm 1. Assessment stopping algorithm

---

**Inputs:**

$\alpha$ , stopping threshold probability

$P(k | \mathbf{x}_n)$ , predicted probability of class  $k$ ,  $k = 1, \dots, 6$ , after question  $n$

$K_n = \arg \max_{k=1, \dots, 6} P(k | \mathbf{x}_n)$ ; i.e., the most likely class after question  $n$

$C_n$ , the current recommended course placement after question  $n$

**Iterations:**

**for**  $n = 10$  to 29 **do**

    Compute  $K_n$  and  $C_n$  using information from questions 1 to  $n$

**if**  $n == 29$  or  $(P(K_n | \mathbf{x}_n) > \alpha \text{ and } K_n == C_n)$  **then**

        Stop the assessment

**end if**

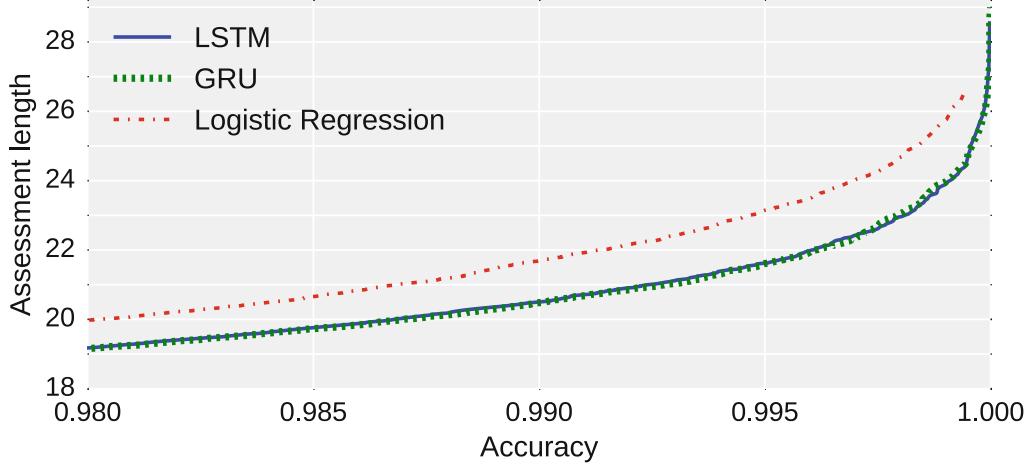
**end for**

**Output:**

$C_n$ , the (predicted) course placement recommendation

---

The results from applying Algorithm 1 to the held-out test data are shown in Fig. 1, where we plot the average assessment length versus the accuracy of the predicted course placement recommendation, for various probability thresholds (i.e., various values of  $\alpha$ ). The plot shows that at any accuracy rate of 0.995 or higher, the RNN models are a minimum of 1.5 questions better than the logistic regression, with the maximum difference being about 2.2 questions.



**Fig. 1.** Average assessment length vs. accuracy on held-out test data.

Next, Table 1 shows the results for the LSTM RNN model partitioned by the actual (ground truth) classification label, using a value of  $\alpha = 0.99$ . The best results are for the extreme labels 1 and 6; on the other hand, while still being acceptable, the gains are not nearly as large for labels 4 and 5. It is worth mentioning that these results closely parallel what was found in [4], where it was shown that ALEKS PPL has the greatest variability for labels 4 and 5, and it seems likely that this variability is a major reason for the weaker performance of the stopping algorithm with these labels.

**Table 1.** Stopping statistics by ground truth label for the LSTM RNN model on held-out test data, using a threshold of  $\alpha = 0.99$ .

Class label	1	2	3	4	5	6
Sample size	4357	8680	11108	7640	8259	9956
Average length	17.87	21.74	22.25	24.75	25.8	16.54
Accuracy	0.9963	0.9955	0.9959	0.9921	0.9921	0.9971

## 4 Discussion

The results from applying our stopping algorithm on a held-out test set show a large potential reduction in the average length of the ALEKS PPL assessment. For example, Fig. 1 shows that at an accuracy of 0.995 the average number of questions for the RNN models is about 21.6, a roughly 25% reduction from the full-length assessment of 29 questions. Additionally, the GRU and LSTM models perform equally well, with both outperforming the logistic regression model, adding further evidence to the growing literature supporting the benefits of applying RNN models to educational data. Of note is that we use a relatively general approach, in that the features are obtained simply by taking the output

of the assessment and feeding it to an RNN. The effectiveness of this technique here motivates the need for further studies involving other adaptive assessments; at the moment, it is not clear if this approach can be successful more generally, or if it is some peculiarity of ALEKS PPL that allows it to work so well.

## References

1. Botelho, A.F., Baker, R.S., Heffernan, N.T.: Improving sensor-free affect detection using deep learning. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017. LNCS (LNAI)*, vol. 10331, pp. 40–51. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_4](https://doi.org/10.1007/978-3-319-61425-0_4)
2. Cho, K., van Merriënboer, B., Gülcühre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR* abs/1406.1078 (2014). <http://arxiv.org/abs/1406.1078>
3. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014)
4. Doble, C., Matayoshi, J., Cosyn, E., Uzun, H., Karami, A.: A data-based simulation study of reliability for an adaptive assessment based on knowledge space theory. *Int. J. Artif. Intell. Educ.* (2019). <https://doi.org/10.1007/s40593-019-00176-0>
5. Doignon, J.P., Falmagne, J.C.: Spaces for the assessment of knowledge. *Int. J. Man-Mach. Stud.* **23**, 175–196 (1985)
6. Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (eds.): *Knowledge Spaces: Applications in Education*. Springer, Heidelberg (2013). <https://doi.org/10.1007/978-3-642-35329-1>
7. Falmagne, J.C., Doignon, J.P.: *Learning Spaces*. Springer, Heidelberg (2011). <https://doi.org/10.1007/978-3-642-01039-2>
8. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: *Advances in Neural Information Processing Systems*, vol. 29 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997)
10. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, pp. 448–456 (2015)
11. Jiang, W., Pardos, Z.A., Wei, Q.: Goal-based course recommendation. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 36–45 (2019)
12. Jiang, Y., et al.: Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10947, pp. 198–211. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_15](https://doi.org/10.1007/978-3-319-93843-1_15)
13. Khajah, M., Lindsey, R., Mozer, M.: How deep is knowledge tracing? In: *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 94–101 (2016)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)
15. Lin, C., Chi, M.: A comparisons of BKT, RNN and LSTM for learning gain prediction. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017. LNCS (LNAI)*, vol. 10331, pp. 536–539. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_58](https://doi.org/10.1007/978-3-319-61425-0_58)

16. Matayoshi, J., Granziol, U., Doble, C., Uzun, H., Cosyn, E.: Forgetting curves and testing effect in an adaptive learning and assessment system. In: Proceedings of the 11th International Conference on Educational Data Mining, pp. 607–612 (2018)
17. McGraw-Hill Education/ALEKS Corporation: What is ALEKS? [https://www.aleks.com/about\\_aleks](https://www.aleks.com/about_aleks)
18. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
19. Prechelt, L.: Early stopping — but when? In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 53–67. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-35289-8\\_5](https://doi.org/10.1007/978-3-642-35289-8_5)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *Neural Comput.* **15**, 1929–1968 (2014)
21. Xiong, X., Zhao, S., Vaninwegen, E., Beck, J.: Going deeper with knowledge tracing. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 545–550 (2016)
22. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint [arXiv:1702.01923](https://arxiv.org/abs/1702.01923) (2017)



# Participatory Design to Lower the Threshold for Intelligent Support Authoring

Manolis Mavrikis, Sokratis Karkalas, Mutlu Cukurova<sup>(✉)</sup>,  
and Emmanouela Papapesiou

UCL Knowledge Lab, University College London,  
23-29 Emerald Street, London WC1N3QS, UK  
[m.mavrikis@ucl.ac.uk](mailto:m.mavrikis@ucl.ac.uk)

**Abstract.** One of the fundamental aims of authoring tools is to provide teachers with opportunities to configure, modify and generally appropriate the content and pedagogical strategies of intelligent systems. Despite some progress in the field, there is still a need for tools that have low thresholds in terms of the users' technical expertise. Here, we demonstrate that designing systems with lower entry barrier can potentially be achieved through co-design activities with non-programmers and carefully observing novices. Following an iterative participatory co-design cycle with teachers who have little or no programming expertise, we reflect on their proposed enhancements. Our investigations focus on Authelo, an authoring tool that has been designed primarily for Exploratory Learning Objects, but we conclude the paper by providing transferable lessons, particularly the strong preference for visual interfaces and high-level pedagogical predicates for authoring analysis and feedback rules.

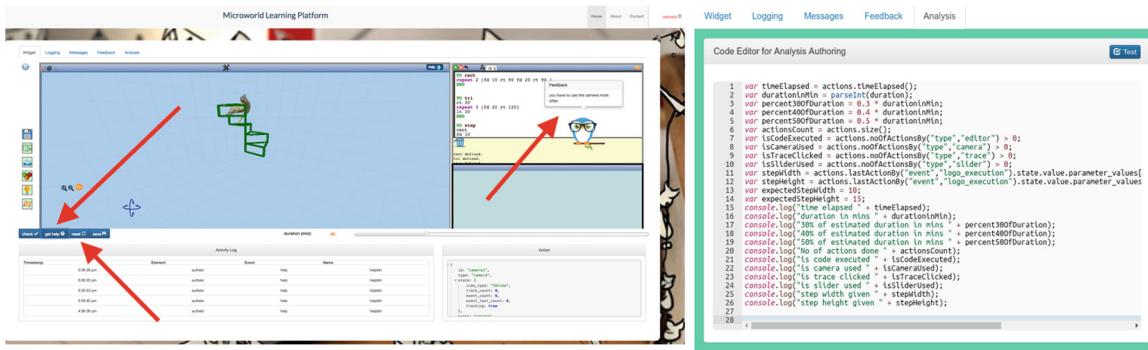
**Keywords:** Intelligent systems · Authoring tools · Participatory design

## 1 Introduction

The aspirational goal behind the development of authoring tools for many years has been to enable users, and teachers in particular, with low technical expertise to create or modify the content and ideally the adaptivity of AIED system, according to their preferred pedagogical strategies [2,3,9]. However, the usability of such tools and particularly the time required to invest in learning them, are factors that affect teachers' adoption and engagement in the design process of authoring [6]. It is important to understand that teachers have different expertise, needs and motivations and authoring tools should aim to meet those. In this paper, we present our approach to better appropriate authoring tools for teachers through participatory co-design activities.

Our case study is on AuthELO [5] that has been specifically designed for authoring intelligent support for Exploratory Learning Objects (ELOs) i.e. open-ended environments such as simulators, microworlds and other inquiry learning environments [8]. AuthELO's design is inspired by the example-tracing approach [1] that encourages authors to develop feedback by executing the activity like a student and the FRAME approach [4] that requires separating the different concerns of feedback, reasoning, analysis and raw data from the model/events. This provides the author with data in a log window that represent the various states of the student interaction throughout the learning activity. Based on this evidence, the author can then perform analysis to derive the facts that drive feedback decisions in real time. Then the variables that correspond to those facts are used to set up rules for the generation of formative and summative feedback (see Fig. 1 and our previous work [4,5] for more details).

This paper reflects on a participatory design study aiming to inform further development of AuthELO towards lowering the entry threshold for teachers who have little or no programming expertise.



**Fig. 1.** Parts of the AuthELO interface. After configuring the ‘logging’ in the corresponding tab, authors can start doing the activity like a student. That will immediately start generating data in the log that can be used for analysis and authoring feedback rules as in this example. The owl is the chosen feedback agent and here it is being tested with a 3D Logo activity.

## 2 Participatory Design for the Authoring UI

For the purposes of this study participation of teachers in the design process was of paramount importance since the main aim is to lower the entry barrier. This is a clear case for participatory design, a well-accepted method for attempting to solve a complicated design problem with the active involvement of people from different backgrounds and different expertise [10].

Based on a non-random sampling strategy and a design-thinking approach, we carefully selected 6 newly qualified teachers who were studying at the UCL masters in Education and Technology and had a range of expertise in using technology in pre-school and primary education settings, but no programming

background (*non-programmers group*). The main goal was to provide rich and in-depth data and so the participants were further divided into two focus groups that were facilitated by one of the authors (EP) going through ideation, sketching brainstorming, and thinking aloud around the interaction with a prototype.

We also selected 3 more experienced computing teachers with enough programming background to teach computing but not necessarily professional programmers to develop applications. They are all skilled in basic JavaScript (*novice group*). They were supported to develop 15 different activities in a 3D Logo environment called MALT [7], and the corresponding support in AuthELO. We recorded the support that one of the authors (SK) had to provide. The objective was to see how the authoring is used and identify difficulties, commonalities and patterns in their solutions that can provide the basis of a higher-level language.

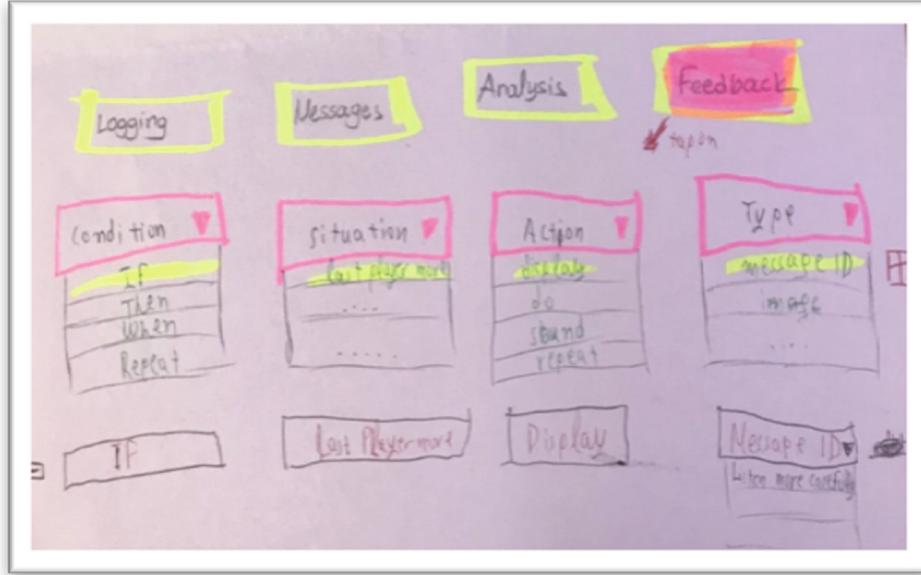
### 3 Key Findings and Discussion

Due to space limitations, we focus only on two key themes that emerged.

**The Influence of Block Coding.** One of the participants of the first group (familiar with block coding user interfaces) spontaneously proposed to introduce blocks of code with pre-defined “variables already written on, so we can drag and drop the blocks and connect them”. Building on this idea, another participant drew a sketch with custom select lists “from where you can click on to see all the variables and choose one”. Ensuing a conversation and brainstorming, the group sketched their final idea that involved use four custom select lists as shown in Fig. 2. They named the first list ‘condition’, and from this list, they could pick the words ‘if’, ‘then’ and ‘others’. The second one was named ‘situation’, and when clicked all the previously set variables would appear on the list so they are able to pick the one that they need. The participants named the third list ‘action’, and with this list, they set what the variable should do, e.g. ‘display’, ‘do’, ‘play sound’. The rest of the discussion was pragmatic and involved including a list that the participants named ‘type’ to choose from the list of resources that should be displayed and other aspects such as a delete button. The key contribution here was the idea that the interaction would result with the constructed rule clearly visibly below that would be added in the list of rules.

The second focus group, led by a one member of the group who volunteered to sketch their thoughts, steered towards an idea of ‘board’ with a standard structure where the words ‘if’, ‘then’, ‘else’, ‘or’ were written i.e. a custom select list for the various conditional statements (called ‘conditions’). This idea became the centrepiece around which two other boards were proposed (‘actions’ and ‘reactions’) as well as a ‘construction’ area for dragging and dropping the various choices to make the feedback rules.

**Situation and Actions as Predicates.** Analysing the brainstorming of the non-programmer groups, a dominant theme was their concern on how to trigger



**Fig. 2.** Low fidelity paper prototype proposed by the non-programmers group

feedback—something that they came up with unprompted. Without any prior knowledge in authoring rules or functional programming, they referred to ‘situations’ or ‘actions’ in a similar way to predicates in programming. Generalising even from a simple task they had, the participants referred to “all previously set variables to appear on the list so they are able to pick the one that they need”.

Analysing the support we had to provide to the novice users, the majority also revolved around the type of actions logged. Observing and generalising the solutions, we managed to reduce the code into a very small set of functions that seem to be a common requirement in all the activities. In particular, some of the high level constructs that were used in analysis and feedback and seem general to other situations are: (1) Number of actions since a particular trigger point (e.g. the start of the activity, or a particular event such as enabling the camera view), (2) The number of actions of a particular type (e.g. the number of times a button was clicked), (3) A list of actions of a particular type and references mostly to the first and last of those. Furthermore, other high-level constructs were the time elapsed from the beginning of the activity, a pre-defined expected duration of the activity and a way to refer to the potential feedback messages and their types directly in a simplified way.

## 4 Conclusions

In this paper, we described our approach to potentially lowering the threshold for intelligent support authoring. We have incorporated the high-level constructs that emerged in the design of AuthELO to ease the authoring declarative statements. The work described here also paves the way for a new user interface that takes advantage of the prevalence of block coding among computing teachers. Of course, it remains to be seen whether busy teachers with low technical

expertise would be inclined to engage deeply with an authoring system even if it has a lower threshold. Our sample is not the most representative of teachers but the participatory design process indicated that while full development of an intelligent system would not necessarily be of interest to our participants, they value the possibility to easily modify the pre-designed feedback and occasionally add to suit their needs. Although, the results are from a small sample, we think that the designs proposed can reduce the amount of initial knowledge required from an author, increase readability, testability and maintainability of the code generated, and allow the analysis to be communicated easier between authors in collaborative projects.

## References

1. Aleven, V., et al.: Example-tracing tutors: intelligent tutor development for non-programmers. *Int. J. Artif. Intell. Educ.* **26**(1), 224–269 (2016). <https://doi.org/10.1007/s40593-015-0088-2>
2. Dağ, F., Durdu, L., Gerdan, S.: Evaluation of educational authoring tools for teachers stressing of perceived usability features. *Procedia - Soc. Behav. Sci.* **116**, 888–901 (2014). <https://doi.org/10.1016/j.sbspro.2014.01.316>. <http://www.sciencedirect.com/science/article/pii/S1877042814003334>
3. Gaffney, C., Wade, V.P., Dagger, D.: Authoring and delivering personalised simulations an innovative approach to adaptive eLearning for soft skills (2010). <http://www.tara.tcd.ie/handle/2262/67212>
4. Gutierrez-Santos, S., Mavrikis, M., Magoulas, G.D.: A separation of concerns for engineering intelligent support for exploratory learning environments. *J. Res. Pract. Inf. Technol.*, 103–116 (2012). [http://www.acs.org.au/\\_data/assets/pdf\\_file/0020/16706/JRPIT44.3.347.pdf](http://www.acs.org.au/_data/assets/pdf_file/0020/16706/JRPIT44.3.347.pdf)
5. Karkalas, S., Mavrikis, M.: Feedback authoring for exploratory learning objects: Authelo. In: CSEDU 2016 - Proceedings of the 8th International Conference on Computer Supported Education, Volume 1, Rome, Italy, 21–23 April 2016, pp. 144–153 (2016). <https://doi.org/10.5220/0005810701440153>
6. Karoui, A., Marfisi-Schottman, I., George, S.: Mobile learning game authoring tools: assessment, synthesis and proposals. In: Bottino, R., Jeuring, J., Veltkamp, R.C. (eds.) GALA 2016. LNCS, vol. 10056, pp. 281–291. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-50182-6\\_25](https://doi.org/10.1007/978-3-319-50182-6_25)
7. Kynigos, C., Grizioti, M.: Programming approaches to computational thinking: integrating turtle geometry, dynamic manipulation and 3D space. *Inform. Educ.* **17**(2), 321–340 (2018). <https://eric.ed.gov/?id=EJ1195612>
8. Mavrikis, M., Gutierrez-Santos, S., Geraniou, E., Noss, R.: Design requirements, student perception indicators and validation metrics for intelligent exploratory learning environments. *Pers. Ubiquitous Comput.* **17**(8), 1605–1620 (2013). <https://doi.org/10.1007/s00779-012-0524-3>
9. Murray, T.: Coordinating the complexity of tools, tasks, and users: on theory-based approaches to authoring tool usability. *Int. J. Artif. Intell. Educ.* **26**(1), 37–71 (2016). <https://doi.org/10.1007/s40593-015-0076-6>
10. Vines, J., Clarke, R., Wright, P., McCarthy, J., Olivier, P.: Configuring participation: on how we involve people in design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 429–438. ACM, New York (2013). <https://doi.org/10.1145/2470654.2470716>



# Finding Relevant e-Learning Materials

Blessing Mbipom<sup>(✉)</sup>

Department of Computer Science, University of Reading, Reading RG6 6AY, UK  
[b.e.m bipom@reading.ac.uk](mailto:b.e.m bipom@reading.ac.uk)

**Abstract.** Learners in e-Learning environments often have difficulty finding and retrieving relevant learning materials to support their learning goals because they lack sufficient domain knowledge to craft effective queries that convey what they wish to learn. In addition, the unfamiliar vocabulary often used by domain experts makes it difficult to map learners' queries to relevant documents. Hence the need to develop a suitable method that would support finding and recommending relevant learning materials to learners. These challenges are addressed by exploiting a knowledge-rich method that automatically creates custom background knowledge in the form of a set of rich concepts related to the selected learning domain. A method is developed which allows the background knowledge to influence the refinement of queries during the recommendation of learning materials. The effectiveness of this approach is evaluated on a dataset of Machine Learning and Data Mining documents and it is shown to outperform benchmark methods. The results confirm that adopting a knowledge-rich representation within e-Learning recommendation improves the ability to find and recommend relevant e-Learning materials to learners.

**Keywords:** Knowledge representation and reasoning · e-Learning · Knowledge discovery · Artificial intelligence · Recommender systems

## 1 Introduction

Have you ever struggled when trying to type a query into a search engine? You try a couple of times, and then refine your query based on the search results you receive. This example describes how difficult searching for relevant documents can be. Besides, research has shown that users find it difficult when searching for relevant information [1, 2], and so is finding relevant learning materials online.

At a conference I attended during my PhD, each research student was assigned to a mentor. The students discussed their research with their mentors and received some feedback. After I described my project, my mentor said “this is a very relevant project, I believe that my new PhD students can benefit from this project when finding their research papers. Often, when they try to look for relevant literature, they report challenges in finding relevant papers, but when I search, I am able to find those papers”. Well, this story highlights a challenge in e-Learning recommendation. The vocabulary used by teaching experts is often

different from the vocabulary used by learners. So, when developing a solution to address this challenge, there is a need to bridge the semantic gap between the teaching experts and learners, to find documents that address learners' queries as illustrated in Fig. 1. Related work adopts external knowledge sources for query refinement [3, 4]. This paper draws insight from such methods with a focus on exploiting knowledge from teaching experts for refining queries.

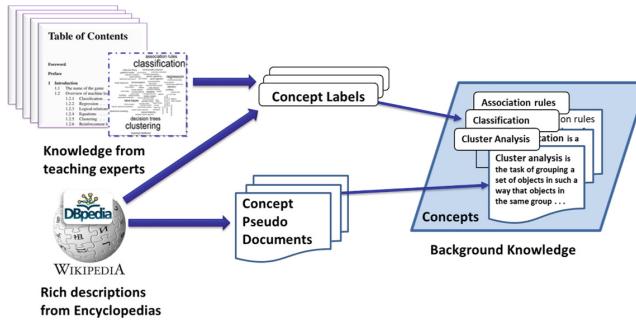


**Fig. 1.** Bridging the semantic gap between learners and teaching experts

In this paper, the challenge is addressed by exploiting background knowledge harnessed from the knowledge of teaching experts contained in e-Books. The e-Books are used as a guide to identify important domain topics. The identified topics are then enriched with discovered text from an encyclopedia source, DBpedia and this helps to increase the richness of the background knowledge. So, when a query is received from a learner, the vocabulary from the background knowledge is used to refine a learner's query. Then the refined query is used to search for relevant learning materials. The e-Learning recommender system developed in [5] is employed to evaluate the performance of the developed method. The evaluation is performed by experts in the domain, and the results show the developed method to outperform standard techniques.

## 2 Refining Queries Using Background Knowledge

Background knowledge refers to information about a domain that is useful for general understanding and problem-solving [6]. Domain knowledge has been leveraged to define problem-solving techniques in intelligent tutoring systems [7], as well as for answering medical queries in WatsonPaths [8]. The quality of learners' queries are assessed in [9], to enable relevant feedback to be provided. In e-Learning, background knowledge can be employed to influence the refinement of learners' queries [4]. For example, in a domain such as Machine Learning, one would find topics such as Classification, and Clustering. Each of these topics would be represented by a concept, in the form of a concept label and a pseudo-document which provides a rich description for the concept. The background knowledge developed in [10] is employed to influence the refinement of learners' queries. Figure 2 shows the method for creating background knowledge.



**Fig. 2.** The background knowledge creation method

The output from this process is the background knowledge containing a set of 150 domain concepts each comprising a concept label and an associated pseudo-document. The terms  $t_1$  to  $t_c$  from the pseudo-documents of concepts,  $C_1$  to  $C_m$  provide the concept vocabulary that is used for refining learners' queries. A concept term matrix using Term Frequency-Inverse Document Frequency (TF-IDF) weighting is created from the concepts. TF-IDF is useful for identifying concepts that are relevant to queries [11]. A collection of potentially useful concept terms is selected from the concept vocabulary to scale up the representation. The background knowledge is represented using the top 10% of concept terms that have the highest TF-IDF values. The selected concept terms are the set of potential terms that will be used by the CONCEPTBASED (CB) method for query refinement to find relevant documents.

When a new query is received from a learner, a search is performed on the domain concepts. A ranked list based on the similarity of each domain concept to the query is retrieved. The terms from the term-vectors of the most similar concepts are combined to create a potential refined query. Terms with the highest weights are then selected from the potential refined query and added to the initial query to create a refined query. The refined query can be used to search on a document collection, and documents would be retrieved and presented to the learner. The retrieved documents should be relevant because the query used for the search has been generated using domain concepts related to the initial query.

### 3 Evaluation and Results

The e-Learning recommender system developed in [12] is employed to evaluate the performance of the different query refinement methods in an e-Learning recommendation task. Three methods are evaluated. First, the CONCEPTBASED query refinement method which uses the most similar concepts to create a concept based representation of a query. Second, the benchmark Bag-Of-Words (BOW) method, which is a standard Information Retrieval method, where a learner's query is represented using the terms in the query only. Finally, a HYBRID method which exploits query features to dynamically choose when to apply the CONCEPTBASED or BOW method to refine a query. The concept label in a query was found to be a dominant feature. So, in the HYBRID method, BOW

is used for queries that contain a concept label such as well-formed queries while CONCEPTBASED is used otherwise for vague queries.

The dataset contains 504 chapters from 32 Machine Learning and Data Mining e-Books. The query collection has 70 learner-focused queries generated from students and online sources such as MOOCs and Quora. The 70 queries were run on all the methods, and the top 3 recommendations from each method was stored. The evaluation system was deployed using Microsoft Azure [13], and it was available to users online for 8 weeks. The evaluation is not a standard one where the users are learners. Instead, the users are employed for the purpose of judging the relevance of recommendations made by the different methods. There were 22 users, 16 PhD students, 3 Researchers, and 3 Lecturers or Professors. All the users had experience in ML/DM. This is useful because the judgements made should be from people who know the domain. The user profile confirmed that most users are competent or expert in the subject. Hence, the confidence in the judgements provided.

The evaluation uses the ratings given by the users across all the query-recommendation pairs to compute the performance of the CONCEPTBASED, BOW and HYBRID methods. The *rating* is the average of the ratings from those users who have evaluated the recommendation,  $r$  for the query,  $q$ .

$$\text{rating}(q, r) = \sum_{u \in U_q} \frac{R_u(q, r)}{|U_q|} \quad (1)$$

where  $(q, r)$  is a query-recommendation pair,  $R_u$  is the rating a user,  $u$  has given to a  $(q, r)$  pair, and  $U_q$  are the users that have evaluated a query,  $q$ . Performance of a method is computed by taking the average *rating* across the queries.

The users provided ratings for 521 query-recommendation  $(q, r)$  pairs. On average users evaluated 4.8 queries and provided ratings for 23.7  $(q, r)$  pairs. For all users, the average ratings for CB is 3.54, HYBRID is 3.45, and BOW is 3.33. So,  $\text{CB} > \text{HYBRID} > \text{BOW}$ . Hence, using the CONCEPTBASED and HYBRID representations of a query to find learning materials is better than when the standard BOW representation is used.

## 4 Conclusion

The growing availability of e-Learning materials on the Web provides opportunities for learners to access new and valuable information. However, finding relevant e-Learning materials can be challenging. This is because learners are often new to the topic they are researching, and so are unable to create effective queries in a search engine. An e-Learning recommender system is employed to demonstrate a method that exploits background knowledge from teaching experts to influence the refinement of learners' queries. The refined queries are then used to focus the search on relevant documents.

Evaluation results demonstrate the effectiveness of the method to support finding and recommending relevant e-Learning materials. In future, the method

can be extended by exploring a Machine Learning approach for choosing when to refine a query. The impact of the adoption of the method presented in this paper can enable increased engagement of learners with e-Learning materials.

## References

1. Liu, J., Kim, C.S., Creel, C.: Why do users feel search task difficult? In: Proceedings of the Association for Information Science and Technology, vol. 50, no. 1, pp. 1–4 (2013)
2. Wood, E., et al.: Exploration of the relative contributions of domain knowledge and search expertise for conducting internet searches. *Ref. Libr.* **57**(3), 182–204 (2016)
3. Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Mapping queries to the linking open data cloud: a case study using DBpedia. *Web Semant.: Sci. Serv. Agents World Wide Web* **9**(4), 418–433 (2011)
4. Bendersky, M., Metzler, D., and Croft, W. B.: Effective query formulation with multiple information sources. In: 5th ACM International Conference on Web Search and Data Mining, pp. 443–452 (2012)
5. Mbipom, B., Massie, S., and Craw, S.: An e-Learning recommender that helps learners find the right materials. In: Proceedings of the 8th Symposium on Educational Advances in Artificial Intelligence, pp. 7928–7933. AAAI Press (2018)
6. Zhang, X., Liu, J., Cole, M.: Task topic knowledge vs. background domain knowledge: impact of two types of knowledge on user search performance. In: Rocha, Á., Correia, A., Wilson, T., Stroetmann, K. (eds.) *Adv. Inf. Syst. Technol.*, pp. 179–191. Springer, Heidelberg (2013)
7. Heeren, B., Jeuring, J.: An extensible domain-specific language for describing problem-solving procedures. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017. LNCS (LNAI)*, vol. 10331, pp. 77–89. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_7](https://doi.org/10.1007/978-3-319-61425-0_7)
8. Lally, A., et al.: WatsonPaths: scenario-based question answering and inference over unstructured information. *AI Mag.* **38**(2), 59–76 (2017)
9. Kopp, K.J., Johnson, A.M., Crossley, S.A., McNamara, D.S.: Assessing question quality using NLP. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017. LNCS (LNAI)*, vol. 10331, pp. 523–527. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_55](https://doi.org/10.1007/978-3-319-61425-0_55)
10. Mbipom, B., Craw, S., Massie, S.: Harnessing background knowledge for e-Learning recommendation. In: Bramer, M., Petridis, M. (eds.) *Research and Development in Intelligent Systems XXXIII*, pp. 3–17. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-47175-4\\_1](https://doi.org/10.1007/978-3-319-47175-4_1)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**(5), 513–523 (1988)
12. Mbipom, B.: Knowledge driven approaches to e-Learning recommendation (2018)
13. Copeland, M., Soh, J., Puca, A., Manning, M., Gollob, D.: *Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud*. Apress, New York (2015)



# Predicting Dialogue Breakdown in Conversational Pedagogical Agents with Multimodal LSTMs

Wookhee Min<sup>1</sup>(✉), Kyungjin Park<sup>1</sup>, Joseph Wiggins<sup>2</sup>,  
Bradford Mott<sup>1</sup>, Eric Wiebe<sup>1</sup>, Kristy Elizabeth Boyer<sup>2</sup>,  
and James Lester<sup>1</sup>

<sup>1</sup> Center for Educational Informatics, North Carolina State University,  
Raleigh, NC 27606, USA

{wmin, kpark8, bwmott, wiebe, lester}@ncsu.edu

<sup>2</sup> Department of Computer and Information Science and Engineering,  
University of Florida, Gainesville, FL 32601, USA

{jbwiggi3, keboyer}@ufl.edu

**Abstract.** Recent years have seen a growing interest in conversational pedagogical agents. However, creating robust dialogue managers for conversational pedagogical agents poses significant challenges. Agents' misunderstandings and inappropriate responses may cause breakdowns in conversational flow, lead to breaches of trust in agent-student relationships, and negatively impact student learning. Dialogue breakdown detection (DBD) is the task of predicting whether an agent's utterance will cause a breakdown in an ongoing conversation. A robust DBD framework can support enhanced user experiences by choosing more appropriate responses, while also offering a method to conduct error analyses and improve dialogue managers. This paper presents a multimodal deep learning-based DBD framework to predict breakdowns in student-agent conversations. We investigate this framework with dialogues between middle school students and a conversational pedagogical agent in a game-based learning environment. Results from a study with 92 middle school students demonstrate that multimodal long short-term memory network (LSTM)-based dialogue breakdown detectors incorporating eye gaze features achieve high predictive accuracies and recall rates, suggesting that multimodal detectors can play an important role in designing conversational pedagogical agents that effectively engage students in dialogue.

**Keywords:** Conversational pedagogical agent · Multimodal · Dialogue breakdown detection · Natural language processing · Gaze

## 1 Introduction

Recent years have seen the emergence of increasingly robust conversational agents paralleling significant advances in natural language processing [1]. A particularly important line of research on conversational agents investigates conversational pedagogical agents [2, 3]. They have demonstrated significant potential in intelligent

tutoring systems as an effective approach to engaging students in tutorial dialogue [4], assessing student knowledge [5], and supporting learning [6]. Conversational pedagogical agents can play a central role in student interactions in game-based learning environments by enhancing students' engagement and facilitating learning through customized narratives and adaptive problem-solving support [7–9].

It is critical that conversational pedagogical agents effectively prevent dialogue breakdown, which is a conversational phenomenon in which a dialogue cannot easily proceed [10]. Dialogue breakdown occurs when an agent misunderstands what a human intends to communicate and, as a result, responds inappropriately. A robust dialogue breakdown detection (DBD) framework could inform conversational pedagogical agents of the need to adaptively modify their dialogue strategies to prevent breakdowns and implement a dialogue recovery strategy [11], and also could enable researchers to examine causes of breakdown in the context of error analysis [12].

In this paper, with the objective of preemptively preventing dialogue breakdown, we investigate multimodal data streams to model human dialogue behaviors. Specifically, we examine four channels: natural language utterances, eye gaze traces, student gender, and task states. Gaze behaviors have been found to be related to cognitive [13] and affective [14] processes, and temporal patterns in eye movements are associated with humans' attention and engagement [15], boredom [14], and intention [16]. We hypothesize that these multimodal features will serve as strong predictors of DBD.

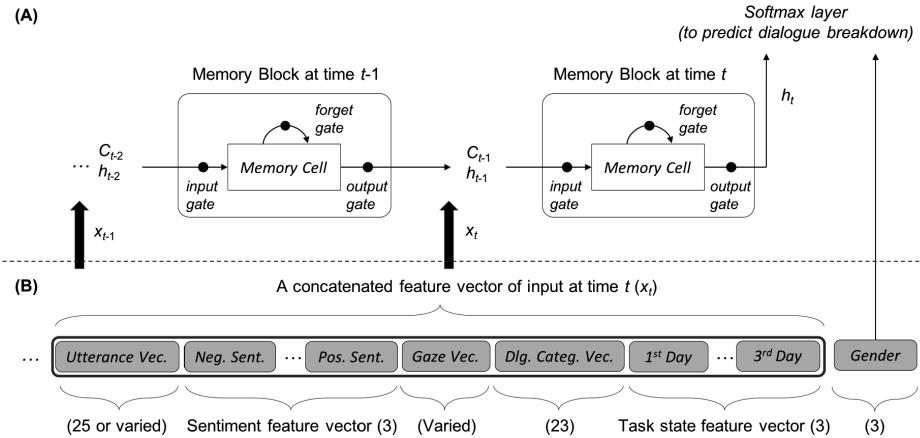
We present a multimodal DBD framework using long short-term memory networks (LSTMs) [17]. We examine 92 middle school students' interaction data with a conversational pedagogical agent in a game-based learning environment for science education [18]. We compare the LSTM-based DBD framework's predictive performance to linear chain conditional random fields (CRFs) as well as support vector machines (SVMs).

## 2 Dialogue Breakdown Detection in CRYSTAL ISLAND

CRYSTAL ISLAND is a game-based learning environment for middle school microbiology [18]. As an extension of the game-based learning environment, we incorporated a conversational pedagogical agent within the game to investigate both affective and cognitive influences on students' learning processes. We developed a state machine-based dialogue manager for this virtual agent, Alisha. Alisha's dialogue moves are made at the agent's initiative or responding to a student dialogue move. Alisha-initiated dialogue moves are triggered by student behaviors in the game, and Alisha-response dialogue moves are made in response to students' dialogue acts [19, 20].

Students played CRYSTAL ISLAND for up to three consecutive days of classroom periods or until they completed the game. Each day, they continued the game from where they ended in their prior session. We annotated dialogue data from 92 students who completed consent forms, conversed with Alisha during the study, and completed all of their surveys. Of these students, 38 identified as Female, 32 as Male, and 22 students did not report their gender. The mean age was 13.4 years ( $SD = 0.69$ ).

We defined a binary annotation scheme, *no breakdown* and *breakdown*, adapted from the labels defined in the Dialogue Breakdown Detection Challenge [11]. Two human annotators labeled the dialogue corpus. Both annotators labeled approximately



**Fig. 1.** (A) The LSTM-based dialogue breakdown detector. (B) An illustration of how the input at each time step is encoded. Each value in the parenthesis denotes the number of features.

20% of the entire corpus in common, achieving an inter-rater agreement of 0.765 (i.e., substantial agreement) using Cohen's kappa [21]. In summary, the number of *breakdown* and *no breakdown* instances are 282 (23.9%) and 897 (76.1%), respectively, from the 1,179 Alisha utterances that appear in the corpus.

We adopt LSTMs (Fig. 1) to model multimodal data streams for DBD. First, we investigate linguistic features. We compare two approaches: GloVe pre-trained word embeddings [22] and a bag-of-words method. In addition, we adopt an off-the-shelf sentiment analysis toolkit [23] to identify if the current conversation is flowing in a positive or negative manner, and sentiments of student dialogues serve as an explanatory variable for DBD. Second, we use traces of objects within the game world that students were looking at in CRYSTAL ISLAND [16]. Third, we use as predictive features the history of previous Alisha dialogue move categories. Fourth, we utilize students' gender as a variable for predictive models since we have observed that female students are more considerate to Alisha than male students, who often experience more breakdown. Finally, we use task states that encode the number of gameplay sessions the student has completed. In addition, we explore an automated post-processing method, which is inspired by work in text normalization [24], to refine model predictions of breakdown in a post-hoc manner.

### 3 Evaluation

We evaluate model performance using student-level ten-fold cross-validation. While predictive accuracy is an important metric, recall is particularly important in this work since the primary objective is to identify potential dialogue breakdown situations in advance and adapt the current policy to avoid them. Because the corpus has an imbalanced distribution in data (only 23.9% were labeled positive, i.e., *breakdown* instances), in each fold we randomly up-sample positive examples from the training set to have a 50–50 distribution between the two labels, and evaluate trained models with the test set for which no up-sampling was applied.

**Table 1.** Average accuracy rates over test examples in CV (**P** and **BoW** denote the post-processing technique applied and the bag-of-words method, respectively).

	Gaze+P	Gaze	NoGaze+P	NoGaze
LSTM (BoW)	<b>79.56</b>	78.29	79.22	78.20
LSTM (GloVe)	76.59	75.91	78.37	77.44
CRF (BoW)	71.25	70.40	<b>73.88</b>	73.03
CRF (GloVe)	70.23	68.53	72.01	70.48
SVM (BoW)	76.84	76.42	<b>77.10</b>	76.68
SVM (GloVe)	68.36	67.94	66.50	65.65

In this work, we investigate two baseline models, including linear CRFs and SVMs with a radial basis function. We evaluate the models' predictive accuracy across the three machine learning techniques. A different set of hyperparameters for each of the LSTMs, CRFs, and SVMs is explored, and only the highest accuracy rate among a set of hyperparameter configurations is reported per feature set variant in Table 1. Then, we further evaluate the recall, precision, and F1 of the models that achieve the highest predictive accuracy per machine-learning technique. The highest accuracy (79.56%), recall (0.67), precision (0.56), and F1 (0.61) are attained by multimodal LSTMs utilizing the eye gaze features and the bag-of-words method with the post-processing technique applied. Notably, these multimodal LSTMs outperform LSTMs not utilizing eye gaze traces with respect to all the metrics: predictive accuracy, recall, precision, and F1, as well as CRFs and SVMs. A sizable improvement was achieved by the with-gaze LSTMs in the recall rate over without-gaze LSTMs (0.674 vs. 0.642), which indicates multimodal LSTMs are more effective in detecting dialogue breakdowns. This difference accounts for a normalized gain of 8.94%.

## 4 Conclusion

Conversational pedagogical agents offer great potential for supporting students' problem solving and promoting engagement in game-based learning environments. However, dialogue breakdown between students and agents poses significant challenges and may impede student learning and diminish student engagement. This paper has presented a multimodal deep learning-based dialogue breakdown detection framework that utilizes natural language interactions, eye gaze traces, student gender, and tasks states. Results suggest that a multimodal LSTM-based DBD framework can achieve high predictive accuracies and recall rates, outperforming competitive baseline approaches. In future work it will be important to investigate the potential contribution of additional modalities for improving dialogue breakdown detection. For example, incorporating facial expression and other affective channels may lead to further improvements in dialogue breakdown detection, thereby increasing conversational pedagogical agents' capabilities to engage in even more effective dialogues with students during learning interactions.

**Acknowledgments.** This research was funded by the National Science Foundation under grants CHS-1409639 and DRL-1640141. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

1. Hirschberg, J., Manning, C.D.: Advances in natural language processing. *Science* **349**, 261–266 (2015)
2. Kim, Y., Baylor, A.L.: Research-based design of pedagogical agent roles: a review, progress, and recommendations. *Int. J. Artif. Intell. Educ.* **26**(1), 160–169 (2016)
3. Tegos, S., Demetriadis, S.: Conversational agents improve peer learning through building on prior knowledge. *Educ. Technol. Soc.* **20**, 99–111 (2017)
4. Graesser, A.C.: Conversations with AutoTutor help students learn. *Int. J. Artif. Intell. Educ.* **26**, 124–132 (2016)
5. Litman, D., et al.: Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 270–275 (2016)
6. Rus, V., Mello, S.D., Hu, X., Graesser, A.C.: Recent advances in conversational intelligent tutoring systems. *AI Mag.* **34**(3), 42–54 (2013)
7. Lester, J., Ha, E., Lee, S., Mott, B., Rowe, J., Sabourin, J.: Serious games get smart: intelligent game-based learning environments. *AI Mag.* **34**(4), 31–45 (2013)
8. Johnson, W.L., Lester, J.C.: Face-to-face interaction with pedagogical agents, twenty years later. *Int. J. Artif. Intell. Educ.* **26**(1), 25–36 (2016)
9. Pezzullo, Lydia G., et al.: “Thanks Alisha, keep in touch”: gender effects and engagement with virtual learning companions. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 299–310. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_25](https://doi.org/10.1007/978-3-319-61425-0_25)
10. Martinovsky, B., Traum, D.: The error is the clue: breakdown in human-machine interaction. In: Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems, pp. 11–17 (2003)
11. Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., Kaji, N.: Overview of dialogue breakdown detection challenge 3. In: Proceedings of Dialog System Technology Challenge 6 (2017)
12. Higashinaka, R., Funakoshi, K., Araki, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 87–95 (2015)
13. Steichen, B., Carenini, G., Conati, C.: User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 317–328. ACM (2013)
14. D'Mello, S., Olney, A., Williams, C., Hays, P.: Gaze tutor: a gaze-reactive intelligent tutoring system. *Int. J. Hum.-Comput. Stud.* **70**, 377–398 (2012)
15. Hutt, S., Mills, C., White, S., Donnelly, P.J., D'Mello, S.K.: The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 86–93 (2016)
16. Min, W., et al.: Multimodal goal recognition in open-world digital games. In: 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, pp. 80–86 (2017)

17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1–32 (1997)
18. Rowe, J.P., Shores, L.R., Mott, B.W., Lester, J.C.: Integrating learning, problem solving, and engagement in narrative-centered learning environments. *Int. J. Artif. Intell. Educ.* **21**, 115–133 (2011)
19. Stolcke, A., et al.: Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **26**(3), 339–373 (2000)
20. Min, W., et al.: Predicting dialogue acts of virtual learning companion utilizing student multimodal interaction data. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 454–459 (2016)
21. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960)
22. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
23. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics System Demonstrations, pp. 55–60 (2014)
24. Min, W., Mott, B.W.: NCSU\_SAS\_WOOKHEE: a deep contextual long-short term memory model for text normalization. In: Proceedings of the Workshop for the Normalization of Noisy User Text, pp. 111–119 (2015)



# Pique: Recommending a Personalized Sequence of Research Papers to Engage Student Curiosity

Maryam Mohseni<sup>1</sup>✉, Mary Lou Maher<sup>1</sup>, Kazjon Grace<sup>2</sup>, Nadia Najjar<sup>1</sup>, Fakhri Abbas<sup>1</sup>, and Omar Eltayeby<sup>1</sup>

<sup>1</sup> UNC Charlotte, Charlotte, USA

{mmohseni, m.maher, nanajjar, fabbasl, oeltayeb}@uncc.edu

<sup>2</sup> University of Sydney, Sydney, Australia

kazjon.grace@sydney.edu.au

**Abstract.** This paper describes Pique, a web-based recommendation system that applies word embedding and a sequence generator to present students with a sequence of scientific paper recommendations personalized to their background and interest. The use of natural language processing (NLP) on learning materials enables educational environments to present students with papers with content that is responsive to their knowledge history and interests. Instructors tend to focus on presentation of learning materials based on overall learning goals in a course rather than personalizing the presentation for each student. The ultimate goal of Pique is to provide learners with content that will encourage their curiosity to learn more by presenting sequences of papers with increasingly more novel content. We piloted Pique with students in a course and report on their responses to the recommended sequences. The next steps are to improve the identification of relevant keywords to represent content and the algorithm for the sequence generator.

**Keywords:** Personalized learning · Curiosity · Natural language processing · Educational recommendation

## 1 Introduction and Motivation

This paper presents a novel approach to personalized and adaptive learning that uses an AI model of similarity and surprise to recommend content to students. Personalization in the form of recommendations for learning materials is an area that has obtained significant interest from researchers in recent years [8]. The learning experiences of the student within a course can be personalized by considering student's profile which can consist of learner's prior knowledge, abilities, interests, and learning styles. Developing a learner's knowledge by recommending sequences of concepts which are relevant to her background and interest can be challenging when the concepts are presented in papers. We extend Surprise Walks, proposed by Grace et al. [5], a strategy for generating sequences of increasingly surprising concepts with a goal concept at the end of the sequence. The surprise walks study in [5] is in the domain of recipes while our study is in the domain of HCI research papers.

Unexpectedness can lead to surprise and curiosity [5–7, 10]. Recently surprise and unexpectedness have been proposed as components of a new kind of recommender system: one that explicitly attempts to expand its users’ preferences [1, 2, 9, 11]. Recommending surprising and valuable content may motivate users to read more broadly by stimulating the natural tendency to be curious. Curiosity plays an essential role in exploration and is desirable especially for learners engaging in information-seeking behavior [12]. Curiosity is recognized as a socio-emotional learning skill that leads to learning through constructing one’s own understanding, rather than “being told” or “instructed” what to think [12]. Curiosity is also a key intrinsic motivator in educational contexts, almost by definition: it is the desire to understand and discover [14].

Grace et al. [7] present a framework for PQE (Personalized Curiosity Engine) systems, based on the goal of encouraging curiosity in users by presenting or generating what they will find novel and valuable. Novelty is modeled as a likelihood of co-occurrence of topics in unstructured text documents. The prototype described in [7] represents each paper using its abstract and title and can identify abstracts that are highly novel as well as those that are highly conventional. Each paper is represented as a bag of words, and a Correlated Topic Model algorithm [3, 4] is used to generate the distribution and correlation of topics in the corpus of research papers. In this paper we describe Pique, an approach for recommending research papers that uses co-occurrence of topics [7] and the concept of surprise walks [5].

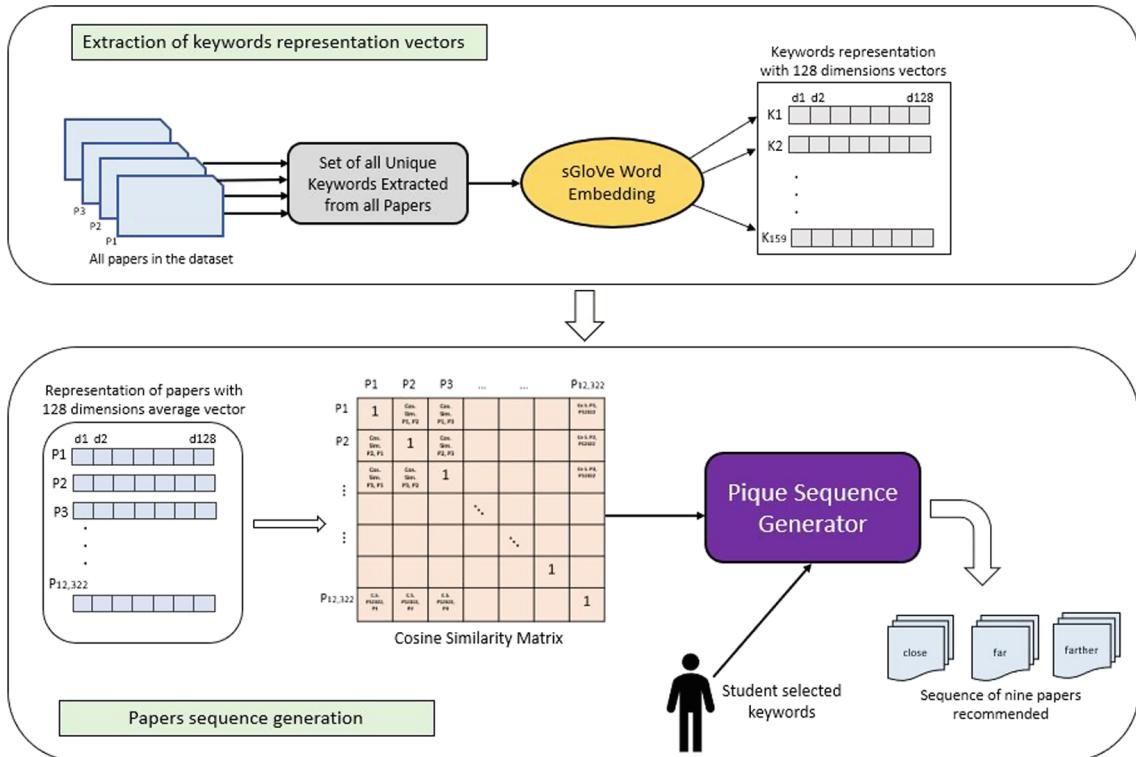
## 2 Model for Generating Sequences of Papers to Recommend

Our overall model for generating a personalized sequence of learning materials is shown in Fig. 1. The model has two major components: content preparation in which we extract a feature vector for each keyword associated with the learning materials (feature extraction/representation step) and a sequence generator that operates on the feature vector that represents each item in the corpus of learning material. The representation of the content of the learning materials is based on the keywords associated with the item. The sequence generator uses cosine similarity between feature vectors of keywords to measure the distance between any two papers in the corpus.

We selected a dataset to demonstrate Pique in a course titled “Interaction Design Studio”. The course has an interdisciplinary semester long project with students from two programs: HCI students with an interest in intelligent buildings, and architects with an interest in HCI. We collected papers from two sources with a total of 12,322 papers: the ACM Digital Library and the Cumincad Index. We extracted 9,452 conference, journal and magazine articles from ACM Digital Library (<http://dl.acm.org/>), each tagged according to the ACM Classification System [15] as belonging to the “Human-Centered Computing” topic. We also scraped 2,870 conference papers, journal articles, reports, and theses from the Cumincad Index (<http://papers.cumincad.org/>) related to the field of Architectural Computing.

Students using Pique are provided with two sets of keywords, and their responses are used to personalize the recommendation. The first set (source set) represents what the student already knows, and the second set (destination set) represents what the student wants to know. The output of the sequence generation algorithm is a sequence

of nine papers in three groups: close, far and farther. The first three papers, categorized as “close”, are closer to student familiarity based on her keywords selection. The next three papers, labeled as “far”, are farther from student familiarity and closer to what student is interested to know more about based on her keywords selection. The last three papers recommended, labeled as “farther” are even farther from the students familiarity and closer to her interests.



**Fig. 1.** Sequence generator using word embedding and cosine similarity

In the dataset each research paper has a set of author defined keywords, with a total of 159 unique keywords. We used sGlove [5], an extension of the GloVe (GLObal VEctors) model developed by Pennington et al. [13] to generate a word embedded model using the keywords in the papers in our dataset. Each paper is represented by a set of keywords:  $P_i : \{K_{i1}, K_{i2}, \dots, K_{in}\}$ , where paper i has n keywords. Using the sGlove model trained on the keywords' co-occurrence matrix, each keyword K is replaced with a vector of 128 dimensions as:  $K : \{d_1, d_2, \dots, d_{128}\}$ , where d is a number representing the value of the keyword K in each dimension in the sGlove vector representation. In order to transform the papers into the same vector space, we averaged the vectors of keywords in each paper, where the averaging was performed per dimension, so the paper representation also ended up with a vector of 128 dimensions:  $P_i : \{d_{i,1}, d_{i,2}, \dots, d_{i,128}\}$ , where d is a number representing the value of the paper Pi in each dimension in the sGlove vector representation. From the vector representations of each paper we constructed a cosine-similarity matrix to estimate the similarity between papers. Given two vectors of 128 dimension for any two papers Pi

and  $P_j$ , the cosine similarity between  $P_i$  and  $P_j$  is calculated using a dot product and magnitude as shown in Eq. 1 below.

$$\text{cosine similarity } P_i \& P_j = \frac{P_i \cdot P_j}{\|P_i\| \|P_j\|} = \frac{\sum_{t=1}^{128} (d_{i,t} * d_{j,t})}{\sqrt{\sum_{t=1}^{128} (d_{i,t})^2} \sqrt{\sum_{t=1}^{128} (d_{j,t})^2}} \quad (1)$$

The similarity matrix is created with each paper in the dataset listed as the index for rows and columns, where each element has the value of the cosine-similarity between the two average vectors of a pair of papers.

The sequence generator determines the origin and destination based on the source and destination sets of keywords. The paper with the highest number of shared keywords becomes the origin paper. For the destination paper the same process is repeated by using the destination keyword set. The selection process for the sequence uses the following rule: given any paper  $P$  the algorithm selects the most similar paper  $P_i$  such that  $P_i$  shares at least  $k$  keywords with the destination paper, where  $k$  is initially set to 1. The selection of papers that share  $k$  keywords with the destination ensures that at each step the algorithm progresses towards the student's stated learning objectives. The final sequence comprises the origin paper, seven papers in the sequence, and the destination paper.

### 3 Summary and Future Work

We observed 20 students using Pique to select three papers on a weekly basis for six weeks. Each time the student uses Pique, they are presented a sequence of nine papers in three categories of “close”, “far” and “farther”. For each paper, the students are presented with the title and keywords of the paper, and a thumbnail image of the first page. When they click on the thumbnail they can read the first page which provides them with the title, authors, abstract, and the publication title. Students were asked to download and read one paper in each of the three categories. After reading the three papers, students are asked two questions for each of the papers they have read. The first question asks about how familiar the content of the paper was to them, and the second asks about how surprising the paper was with the answers being: not at all, not much, neutral, some and very much.

Our observations indicated that we were able to personalize the sequence and students were eager to engage with systems like Pique, but improvements are required to address our goal of encouraging curiosity. We observed a decreasing trend in the students' answers for familiarity of papers in three categories of close, far, and farther, but we did not see any meaningful and expected result regarding surprise. We plan to develop a more robust way to identify keywords that represent the content of the papers and experiment with different models of expectation/surprise. For keyword identification, we plan to consider topic modeling and other NLP approaches. We plan to adapt the sequence generating algorithm to produce monotonically increasing surprise in the sequence of papers recommended. We also plan to collect more data as the preliminary results are not sufficient to determine statistical significance.

**Acknowledgements.** The research reported in this article is funded by NSF IIS1618810 CompCog: RI: Small: Pique: A cognitive model of curiosity for personalizing sequences of learning resources.

## References

1. Adamopoulos, P., Tuzhilin, A.: On over-specialization and concentration bias of recommendations: probabilistic neighborhood selection in collaborative filtering systems. In Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 2014, pp. 153–160. ACM, New York (2014)
2. Adamopoulos, P., Tuzhilin, A.: On unexpectedness in recommender systems: or how to better expect the unexpected. *ACM Trans. Intell. Syst. Technol.* **5**(4), 32 p. (2014). Article no. 54
3. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. *Ann. Appl. Stat.* **1**(1), 17–35 (2007). <https://doi.org/10.1214/07-AOAS114>
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
5. Grace, K., Maher, M.L., Davis, N., Eltayeby, O.: Surprise walks: encouraging users towards novel concepts with sequential suggestions. In: ICCC 2018 International Conference on Computational Creativity. ACC (2018)
6. Grace, K., Maher, M.L., Fisher, D., Brady, K.: Modeling expectation for evaluating surprise in design creativity. In: Gero, J.S., Hanna, S. (eds.) Design Computing and Cognition ‘14, pp. 189–206. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-14956-1\\_11](https://doi.org/10.1007/978-3-319-14956-1_11)
7. Grace, K., Maher, M.L., Mohseni, M., y Pérez, R.P.: Encouraging p-creative behaviour with computational curiosity. In: ICCC 2017 International Conference on Computational Creativity. ACC (2017)
8. Imran, H., Belghis-Zadeh, M., Chang, T.W., et al.: PLORS: a personalized learning object recommender system. *Vietnam J Comput Sci* **3**(3), 3–13 (2016)
9. Maher, M.L., Grace, K.: Encouraging curiosity in case-based reasoning and recommender systems. In: Aha, D.W., Lieber, J. (eds.) ICCBR 2017. LNCS (LNAI), vol. 10339, pp. 3–15. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61030-6\\_1](https://doi.org/10.1007/978-3-319-61030-6_1)
10. Niu, X., Abbas, F., Maher, M.L., Grace, K.: Surprise me if you can: serendipity in health information. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, paper No. 23. ACM, New York (2018)
11. Niu, X., Zadrożny, W., Grace, K., Ke, W.: Computational surprise in information retrieval. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, pp. 1427–1429. ACM, New York (2018)
12. Paranjape, B., Bai, Z., Cassell, J.: Predicting the temporal and social dynamics of curiosity in small group learning. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 420–435. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_31](https://doi.org/10.1007/978-3-319-93843-1_31)
13. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
14. Saraswat, N., Ghosh, H., Agrawal, M., Narayanan, U.: Contextual recommendation of educational contents. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 439–448. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_44](https://doi.org/10.1007/978-3-319-19773-9_44)
15. ACM Classification System. <https://dl.acm.org/ccs/ccs.cfm>



# Group Formation for Collaborative Learning

## A Systematic Literature Review

Chinasa Odo<sup>1</sup>(✉), Judith Masthoff<sup>2</sup>, and Nigel Beacham<sup>1</sup>

<sup>1</sup> University of Aberdeen, Aberdeen, UK

{r01cro17,n.beacham}@abdn.ac.uk

<sup>2</sup> Utrecht University, Utrecht, The Netherlands

j.f.m.masthoff@uu.nl

**Abstract.** This paper presents a systematic literature review (SLR) that investigates group formation, as a first step towards automated group formation for collaborative learning. Out of 105 papers selected for review, after using specific selection and a quality assessment method, a final list of 21 relevant studies was selected for analysis. The review revealed the current state of the art in group formation.

**Keywords:** Systematic review · Group formation · Collaboration

## 1 Introduction

A group is a structure that creates boundaries and sustains interaction amongst members. Grouping is essential in human nature specifically in learning. It is central to human lives, and it is hard to imagine human existence without a group [17]. Being in a group brings about collaboration in the achievement of set goals. Collaboration is rooted in the theory of [36] which explains the Zone of Proximal Development. Vygotsky's theory believes in the construction of knowledge through social interactions among peers in a community. Research has shown that learners feel more engaged when they are given the opportunity to be part of the learning process [7, 17] because it is student-centered learning.

Groups, are formed when two or more people interact and influence each other's discussion for learning and understanding learning contents more completely [10, 12]. In the case of online collaboration, it is based on the concept of knowledge construction and gradual building of knowledge through asynchronous online discussion among learners and the instructor. The instructor, acts as a facilitator who provides appropriate resources, learning activities and utilizes knowledge of learners' personality profiles.

This paper investigates the current state of the art in the ways collaborative learning groups are formed, with a particular interest in group types (homogeneous or heterogeneous), learner characteristics, group sizes and algorithms which should inform the automatic formation of learning groups.

## 2 Review Strategy

This research presents a review on group formation following the guideline by [32], which outlined a practical guide on how to carry out a systematic literature review in computer science, based on the work in [18]. Whilst we are interested in automated group formation, this review will be wider, given the limited work in this area, to inspire work on automated group formation.

**Research Questions.** The following research questions will be answered:

1. Do teachers consider homogeneity or heterogeneity when forming groups?
2. What learner characteristics are important criteria for forming groups?
3. What size is recommended as an ideal learning group for collaboration?
4. What are the various techniques in use for automated group formation?

**Literature Sources and Data Gathering.** The study included papers from online databases (EEE, Springer, ACM, Inspire, Crossref, ArXiv, GVK DBLP, Pubmed, PLOS, DOAJ) and search engines (Google Scholar, CiteSeerx) supported by Jabref. A search string was constructed using the method in [32], using synonyms for keywords. The search strings used were: (a) (group OR grouping OR team OR teaming) AND (forming OR formation); (b) (group OR grouping OR team) AND (collaboration OR collaborate); (c) (peer OR peering) AND (recommending OR recommendation OR recommender OR recommend); (d) (group OR grouping OR team Or teaming) AND (size OR sizing).

A search on the titles, abstract and keywords resulted in the first set of 105 papers published from 2002–2017. The identified papers included some that did not address the purposes of this study, stored in multiple databases, or published in many sources. For selection, we applied inclusion and exclusion criteria.

**Inclusion and Exclusion Criteria.** A paper is included only if: (1) it contains online/conventional collaboration, or group/team formation; (2) it is published between 2002 and 2017; (3) it is duplicated or stored in multiple sources, only one copy is selected; (4) it has multiple publications, the most recent one or full version is selected; (5) it has both a conference and journal version, the journal version is selected. A paper is excluded if it: (1) is not related to education; (2) is presented in a language other than English; (3) it is only available in the form of a presentation; (4) it does not address the problem of group collaboration. The application of these criteria reduced the study papers to 48.

**Quality Assessment.** To assess and analyze the selected papers, a 9 item quality assessment checklist was developed and assessed as follows:

1. Venue was evaluated depending on the paper source: (i) For conference and workshop papers, the Computing Research and Education rankings (CORE) were used [8], with values assigned as A = 1.5, B = 1, C = 0.5, No ranking = 0.

- (ii) For journal articles, the Journal Citation Report (JCR) was used which reports citation data [16]. Journals are ranked as Q1-Q4, with values assigned as Q1 = 2, Q2 = 1.5, Q3 = 1, Q4 = 0.5, No JCR = 0.
- 2. Other items such as if paper had been cited and whether it was relevant to each research question were rated as Yes = 1, No = 0.

Papers with an overall quality assessment score greater than 5 were included. At the end of this stage, a list of 21 papers were selected: S01 [3], S02 [5], S03 [6], S04 [9], S05 [20], S06 [21], S07 [25], S08 [29], S09 [31], S10 [34], S11 [37], S12 [38], S13 [1], S14 [2], S15 [4], S16 [13] S17 [15], S18 [22], S19 [23], S20 [26], S21 [30].

### 3 Results and Discussion

**Research Question 1 - Which Group Type Is Considered (homogeneous/heterogeneous) by Teachers/instructors When Forming Learning Groups?** S01, S02, S06, S13 S04, S05 S17 and S18 advocated the formation of heterogeneous groups which is in line with the study of [15, 39]. Only S01 discussed homogeneous group formation which focused on language preference. Study by [33] also noted that homogeneous groups are less stigmatized.

**Research Question 2 - What Learner Characteristics Are Considered as Important Criteria for Forming Collaborative Learning Groups?** Most reviewed papers focused on only one learner characteristic for group formation. The exceptions are S01, which considered gender and language preferences, S20 which considered interest and background for group formation and S17 which considered personality traits and performance feedback. While [27] advocated for group collaboration with interest in combination different learner characteristics. S02, S13 and S16 considered complementary skills of strong and weak learners. S02 suggested that: (i) All members should be expert in one of the identified complementary skills; (ii) Only one member can lead the team. This is like S07, which proposed to combine learners who are more knowledgeable with those who are less knowledgeable but did not mention how this knowledge will be determined. S04, S05 and S18 considered learning style as a measure to bring learners together. However, the use of learning styles is controversial as reported in [11]. S10 considered feedback as a criterion for group formation, maintaining that the quality of previous collaboration is important when forming a group. S11 and S12 mentioned diversity but were not specific on the area of diversification. S17 and S19 proposed personality traits as criteria for forming a group. Research by [24] shows that personality trait is an important factor in collaboration. S02, S03 and S13 proposed getting learners' team work profile using a questionnaire.

**Research Question 3 - What Group Size Is Recommended as an Ideal for Learning Collaboration?** S05, S11, S12 and S15 suggested small groups

without stating any number to constitute a group. S06 suggested that the size should depend on the tutor's choice. only S15 mentioned specific group sizes of 3, 5 and 7. This is supported by the Ringelmann effect (as mentioned by [19]) who noted that members become less productive as group size increases.

**Research Question 4 - What Are the Various Techniques in Use for Automatically Assigning Learners to Groups?** In S01 and S21, a binary integer technique which takes the values of 0 or 1 was proposed. Data mining was proposed by S04. There are many types of data mining techniques which are used to explore and analyze large data set in order to discover meaningful patterns as noted by [28, 35], but S04 did not specify the type they used for grouping. S05 and S11 used genetic algorithms in group formation problem. This consists a set of students S and a set of groups G. The goal is to allocate all learners in S to a group in G, such that the groups are as heterogeneous as possible. The genetic approach in [14] shows that learners are drawn from the population to produce the fittest groups by changing individuals to form better groups which takes different learner characteristics into account.

S06, S08, S09 used an approximation algorithm. In S06, group allocation is made by finding individuals to act as leaders for each group by minimizing a leadership cost function, and then adding individuals to the groups by minimizing the communication cost function (using Greedy Search). The user provides feedback on the resulting groups in terms of which learners to keep in the groups. The algorithm is run again, till the user is satisfied. In S08, a group allocation is made using learners activity ratings; learners with similar preferences are put together. This approximation is evolved into a final group allocation with the desired number of groups. S09 uses backtracking.

S07 used a semantic algorithm, which maximized the diversity in knowledge in the groups. Artifacts (such as essays) produced by learners were analyzed to extract knowledge of each learner. The learners concepts were aggregated into a unified data model, and used to calculate diversity. S10 is based on a group technological approach where similar characteristics are identified and grouped together to take advantage of the similarities. The input data is composed of two matrices: (1) learner characteristics compatibility and (2) assignments of the characteristics to learners. A clustering approach is then used to form groups based on these matrices. S14 used a Bayesian network. Initially learners were divided into disjoint teams. After every activity, learners evaluates their peers by stating the most predominant role of each teammates. At each iteration, Bayesian learning was employed to update the probability for a learner given the evaluation history, these probabilities are then used to form the next teams.

Finally, S17 and S20 used Ant colony optimization and Particle swarm optimization respectively. The first is inspired by the collective foraging behaviour of specific ant species. The objective of the algorithm is to maximize the heterogeneity of all groups based on the Goodness Heterogeneous values of all groups. In S20, the particle swarm optimization technique is used. In this technique, each particle has: (1) a current position in the search space, (2) a current velocity, and

a personal best position in the search space. During each iteration, each particle in the swarm is updated using (1) and (2). In S20, each particle represents a distribution of learners over groups.

## 4 Conclusions and Future Work

This paper provided a systematic literature review on group formation for collaborative learning, as a first step towards automated group formation by a computer agent for group collaborative learning. In the light of the findings, the research was able to identify that (1) The reviewed papers have not specifically considered which of the learner characteristics are considered important when forming a group but tended to focus on a particular characteristic. (2) The reviewed papers did not mention an ideal size to consider when forming a group. (3) The reviewed papers used a wide variety of algorithms with no studies to compare the relative effectiveness of such algorithms. Our future studies will use a mixed method research method with triangulation to determine which learner characteristics to combine to achieve effective collaborative learning groups.

## References

1. Al-Adrousy, W.M., Ali, H.A., Hamza, T.T.: A recommender system for team formation in MANET. *J. King Saud Univ. Comput. Inf. Sci.* **27**(2), 147–159 (2015)
2. Alberola, J.M., del Val, E., Sanchez-Anguix, V., Julian, V.: Simulating a collective intelligence approach to student team formation. In: Pan, J.-S., Polycarpou, M.M., Woźniak, M., de Carvalho, A.C.P.L.F., Quintián, H., Corchado, E. (eds.) HAIS 2013. LNCS (LNAI), vol. 8073, pp. 161–170. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-40846-5\\_17](https://doi.org/10.1007/978-3-642-40846-5_17)
3. Amarasinghe, I., Leo, D.H., Jonsson, A.: Intelligent group formation in computer supported collaborative learning scripts. In: Chang, M., Chen, N., Huang, R., Kinshuk, Sampson, D.G., Vasiu, R. (eds.) 17th IEEE International Conference on Advanced Learning Technologies, ICALT 2017, Timisoara, Romania, 3–7 July 2017, pp. 201–203. IEEE Computer Society (2017)
4. Babar, M.A., Kitchenham, B.: The impact of group size on software architecture evaluation: a controlled experiment. In: Proceedings of First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), pp. 420–429, September 2007
5. Battur, S., et al.: Enhancing the students project with team based learning approach: a case study. In: 2016 IEEE 4th International Conference on MOOCs Proceedings of the Innovation and Technology in Education (MITE), pp. 275–280, December 2016
6. Borges, J., Dias, T.G., Cunha, J.F.E.: A new group-formation method for student projects. *Eur. J. Eng. Educ.* **34**(6), 573–585 (2009)
7. Burke, A.: Group work: how to use groups effectively. *J. Eff. Teach.* **11**(2), 87–95 (2011)
8. Computer Research and Education: CORE Rankings Portal - Computing Research and Education (2017)

9. Costaguta, R., de los Angeles Menini, M.: An assistant agent for group formation in CSCL based on student learning styles. In: Proceedings of the 7th Euro American Conference on Telematics and Information Systems, EATIS 2014, pp. 24:1–24:4. ACM, New York (2014)
10. Dillenbourg, P.: What do you mean by ‘collaborative learning’?, no. 1, pp. 1–19 (1999)
11. Coffield, F., Moseley, D., Hall, E., Ecclestone, K.: Should we be using learning styles? What research has to say to practice. Technical report, The Learning and Skills Research Centre (2014)
12. Galanes, G.J., Adams, K.K.L.: Effective Group Discussion: Theory and Practice. McGraw-Hill, Boston (2013)
13. Garcia, P., Balmaceda, J.M., Schiaffino, S., Amandi, A.: Automatic detection of team roles in computer supported collaborative work. IEEE Latin Am. Trans. **11**(4), 1066–1074 (2013)
14. Gogoulou, A., Gouli, E., Boas, G., Liakou, E., Grigoriadou, M.: Forming Homogeneous, Heterogeneous and Mixed Groups of Learners, pp. 33–40 (2007)
15. Graf, S., Bekele, R.: Forming heterogeneous groups for intelligent collaborative learning systems with ant colony optimization. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 217–226. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_22](https://doi.org/10.1007/11774303_22)
16. JCR: Journal Citation Reports (2017)
17. Johnson, D.W., Johnson, R.T.: Assessing Students in Groups: Promoting Group Responsibility and Individual Accountability. Corwin Press, Thousand Oaks (2004)
18. Kitchenham, B., et al.: The impact of limited search procedures for systematic literature reviews: a participant-observer case study. In: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 336–345 (2009)
19. Kravitz, D.A., Martin, B.: Ringelmann rediscovered: the original article (1986)
20. Lescano, G., Costaguta, R., Amandi, A.: Genetic algorithm for automatic group formation considering student’s learning styles. In: 8th Euro American Conference on Telematics and Information Systems, EATIS 2016, Cartagena, Colombia, 28–29 April 2016, pp. 1–8. IEEE (2016)
21. Li, C.T., Shan, M.K.: Composing activity groups in social networks. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012, pp. 2375–2378. ACM, New York (2012)
22. Liu, S., Joy, M., Griffiths, N.: An exploratory study on group formation based on learning styles. In: Proceedings of the IEEE 13th International Conference on Advanced Learning Technologies, pp. 95–99, July 2013
23. Lykourentzou, I., Antoniou, A., Naudet, Y.: Matching or crashing? Personality-based team formation in crowdsourcing environments (2015)
24. Lykourentzou, I., Antoniou, A., Naudet, Y., Dow, S.P.: Personality matters: balancing for personality types leads to better outcomes for crowd teams. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016, pp. 260–273. ACM, New York (2016)
25. Manske, S., Hoppe, H.U.: Managing knowledge diversity: towards automatic semantic group formation. In: Chang, M., Chen, N., Huang, R., Kinshuk, Sampson, D.G., Vasiu, R. (eds.) 17th IEEE International Conference on Advanced Learning Technologies, ICALT 2017, Timisoara, Romania, 3–7 July 2017, pp. 330–332. IEEE Computer Society (2017)

26. Maria-Iuliana, D., Constanta-Nicoleta, B., Alexandru, B.: Platform for creating collaborative e-learning communities based on automated composition of learning groups. In: Proceedings of the 3rd Eastern European Regional Conference on the Engineering of Computer Based Systems, pp. 103–112, August 2013
27. Odo, C.R., Masthoff, J.F.M., Beacham, N.A., Alhathli, M.A.E.: Affective State for Learning Activities Selection, May 2018
28. Pujari, A.K.: Data Mining Techniques. Universities Press (2001)
29. Roy, S.B., Lakshmanan, L.V.S., Liu, R.: From group recommendations to group formation (2015)
30. Sadeghi, H., Kardan, A.A.: Toward effective group formation in computer-supported collaborative learning. *Interact. Learn. Environ.* **24**(3), 382–395 (2016)
31. Sancho-Asensio, A., Sol, X., Montero, J., Navarro, J., Canaleta, X., Vernet, D.: Support tool for the formation of working groups in collaborative learning environments. In: Proceedings of the 9th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6, June 2014
32. Silva, R.L.S., Neiva, F.W.: Systematic literature review in computer science - a practical guide (2016)
33. So, A., Agbayewa, J.O.: Effect of homogenous and heterogeneous ability grouping class teaching on student's interest, attitude and achievement in integrated science. *Int. J. Psychol. Couns.* **3**(3), 48–54 (2011)
34. Srba, I., Bielikova, M.: Dynamic group formation as an approach to collaborative learning support. *IEEE Trans. Learn. Technol.* **8**(2), 173–186 (2015)
35. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2009)
36. Vygotskii, L.S.L.S.: Thought and Language. M.I.T. Press, Massachusetts Institute of Technology (1962)
37. Wang, D.Y., Lin, S.S.J., Sun, C.T.: DIANA: a computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. *Comput. Hum. Behav.* **23**, 1997–2010 (2007)
38. Wichmann, A., Hecking, T., Elson, M., Christmann, N., Herrmann, T., Hoppe, H.U.: Group formation for small-group learning: are heterogeneous groups more productive? In: Proceedings of the 12th International Symposium on Open Collaboration, OpenSym 2016, pp. 14:1–14:4. ACM, New York (2016)
39. Zamani, M.: Cooperative learning: homogeneous and heterogeneous grouping of Iranian EFL learners in a writing context. *Cogent Educ.* **3**(1), 1149959 (2016)



# AI Meets Austen: Towards Human-Robot Discussions of Literary Metaphor

Natalie Parde<sup>1</sup>(✉) and Rodney D. Nielsen<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Illinois at Chicago, Chicago, USA  
parde@uic.edu

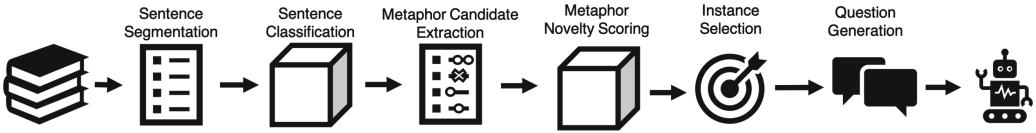
<sup>2</sup> Department of Computer Science and Engineering, University of North Texas,  
Denton, USA  
rodney.nielsen@unt.edu

**Abstract.** Artificial intelligence is revolutionizing formal education, fueled by innovations in learning assessment, content generation, and instructional delivery. Informal, lifelong learning settings have been the subject of less attention. We provide a proof-of-concept for an embodied book discussion companion, designed to stimulate conversations with readers about particularly creative metaphors in fiction literature. We collect ratings from 26 participants, each of whom discuss Jane Austen’s *Pride and Prejudice* with the robot across one or more sessions, and find that participants rate their interactions highly. This suggests that companion robots could be an interesting entryway for the promotion of lifelong learning and cognitive exercise in future applications.

## 1 Introduction

Robotic companions have been examined in many educational settings, acting as learning partners [14], intelligent tutors [5, 8, 12, 26], teachable agents [9, 18, 27], and feedback providers [1]. A common goal among most robots filling these roles to date has been the furtherance of specific learning objectives. They have been underutilized in informal learning settings, which may call for robots to tackle fuzzier objectives for which open-ended conversation is a better avenue of interaction. Reading is a cognitively rewarding way to engage in informal lifelong learning [2, 4, 15, 16, 22, 23], but the potential for companion robots to play a role in motivating lifelong reading behaviors has remained untapped. We set out to fill that void by developing a proof-of-concept embodied conversational companion capable of engaging readers in discussions about books.

We select creative metaphor (a particularly cognitively demanding form of rhetoric [11]) as our literary focus, and demonstrate that an automatic metaphor novelty scoring approach can be harnessed to identify interesting metaphors in literature. We design a conversational dialogue system that makes use of questions generated about those metaphors, and implement it in a companion robot. This is the first approach, either computational or otherwise, to employ metaphor as an impetus for lifelong cognitive exercise, and the first embodied



**Fig. 1.** Processing content from raw text to a series of questions.

conversational system created chiefly to promote such exercise. We empirically confirm that users of the completed prototype rate it as likeable and engaging, and maintain this sentiment over multiple sessions. These contributions form an essential proof-of-concept for an embodied lifelong learning companion.

## 2 Related Work

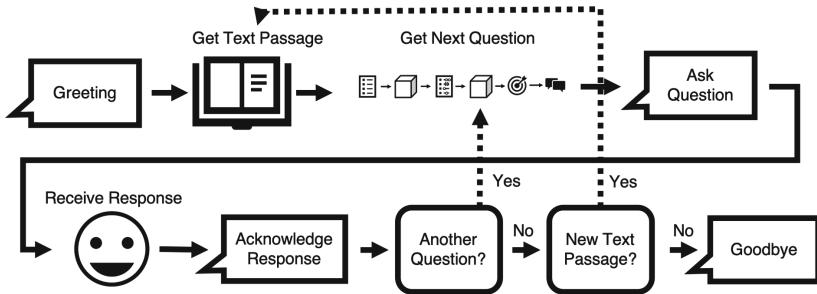
Educational scenarios to which social robots have been deployed have been primarily formal settings with child learners [1, 5, 8, 9, 14, 18, 27]. Our focus is on a different setting: informal, conversational lifelong learning. *Lifelong learning* is the process of acquiring knowledge and/or exercising cognitive faculties across the human lifespan, outside of traditional academic contexts. Research involving social robots in lifelong learning scenarios has been scarce, with most work deploying social robots to adult populations focusing on psychological or physical healthcare needs instead [7, 10, 24, 29]. However, Tapus et al. [28] designed a human-robot music guessing game to stimulate cognition in older adults suffering from dementia, and Deublein et al. [6] created a social robot to scaffold motivation in adult second language learners. Schodde et al. [26] also explored second language learning in adults, although their robot's behaviors were originally designed with children in mind. A common theme across these systems is the absence of open-ended conversation: all cases utilize buttons and multiple-choice answers as their input. The inability to converse naturally limits a robot's potential to engage in cognitively meaningful interactions, particularly when dealing with more subjective topics like literature or metaphor interpretation.

Although virtual avatars could implement the same methods as robots in most of these cases, they may fall short of achieving the same goals. Research has demonstrated that physically embodied robots elicit longer conversations and more positive perceptions than computer agents [25], and are better able to influence people than virtual avatars or videos of the same robots [13]. In accordance with these findings, we implement our system using a physically embodied robot to maximize its anticipated utility.

## 3 System Design

Our system converts raw text to questions about the *novel metaphors*<sup>1</sup> within it using the pipeline in Fig. 1. It embeds the pipeline into the dialogue system

<sup>1</sup> Creative or unexpected metaphors, e.g., “She frowned like a thunderstorm.”



**Fig. 2.** Dialogue system flow.

in Fig. 2. Our pipeline begins by segmenting text into sentences, and classifying each sentence as likely to contain a novel metaphor or not using a neural network model that considers sentence-level context<sup>2</sup> and psycholinguistic features [21]. We train the model using sentences from an existing dataset for which word pairs were labeled with metaphor novelty scores [20]. To repurpose the dataset for sentence-level classification, we label each sentence with a binary value depending on whether any word pairs within it exceed a threshold novelty score.

The system extracts all syntactically-related pairs of *content words* (nouns, verbs, adjectives, and adverbs) from sentences classified as likely to contain a novel metaphor, and predicts a metaphor novelty score for each word pair using our scoring approach defined in prior work [21]. We train our scoring model on a combination of datasets: (1) the only publicly available metaphor novelty dataset, consisting of continuous metaphor novelty scores for 18,439 word pairs from multiple domains [20], and (2) a smaller dataset of 2100 word pairs extracted from Project Gutenberg ([www.gutenberg.org](http://www.gutenberg.org)) books, for which we crowdsourced ratings along the same continuous scale.

Finally, the system selects a word pair based on chronological order, predicted novelty score, similarity to word pairs for which questions have already been generated, similarity of the word pair's source sentence to those for which questions have already been generated, and estimated completion time. It generates a question for the selected word pair using the template-based *Questioning the Author* (QtA) framework [19]. QtA is a questioning technique that prompts readers to consider the author's underlying motivations in crafting prose [3]. We previously showed that automatically-generated QtA questions are cognitively deep and comparable to those generated by humans about the same topics [19].

## 4 Usability Evaluation

Twenty-six participants each discussed Jane Austen's *Pride and Prejudice*<sup>3</sup> with our learning companion robot for 1–3 separate, approximately 30-minute sessions. The system was implemented on a NAO robot named Grace, and we encoded contextual gestures and life-like swaying motions to facilitate natural

<sup>2</sup> Words are represented using Word2Vec embeddings trained on Google News [17].

<sup>3</sup> *Pride and Prejudice* is the most-downloaded book on Project Gutenberg.

**Table 1.** Mode, median, 95% confidence interval, and  $p$  for each statement, for each session. For sessions 1, 2, and 3,  $n = 26$ ,  $n = 18$ , and  $n = 7$ , respectively.

Statement	Mode			Median			95% C.I.			$p$		
	1	2	3	1	2	3	1	2	3	1	2	3
S1: I found Grace easy to understand	4	4	5	4.0	4.0	4.0	3.9 ± .3	3.9 ± .5	3.9 ± .8	.00	.00	.11
S2: I knew what I could say or do at each point of the dialogue	3	3	4	3.5	3.5	4.0	3.5 ± .4	3.6 ± .4	4.1 ± .5	.04	.02	.00
S3: The system worked the way I expected	4	3	5	4.0	3.5	5.0	3.7 ± .3	3.7 ± .5	4.6 ± .4	.00	.02	.00
S4: I would like to use this system regularly	3	4	3	4.0	3.5	3.0	3.6 ± .4	3.3 ± .5	3.4 ± .7	.00	.33	.29
S5: I like interacting with Grace	5	4	5	4.5	4.0	4.0	4.3 ± .3	3.8 ± .5	3.9 ± .8	.00	.01	.11
S6: Grace seems smart	3	3	3	3.0	3.0	4.0	3.5 ± .4	3.2 ± .5	3.7 ± .8	.01	.39	.14
S7: Grace's dialogue seems natural	2	4	4	3.0	4.0	4.0	3.3 ± .4	3.3 ± .5	3.9 ± .7	.23	.25	.08
S8: Grace asked interesting questions about the text we were discussing	3	3	4	4.0	3.5	4.0	3.5 ± .5	3.6 ± .5	4.3 ± .5	.04	.03	.00
S9: It made sense for Grace to ask the questions we discussed	4	4	5	4.0	4.0	5.0	3.8 ± .3	3.7 ± .5	4.6 ± .4	.00	.01	.00

interactions. We employed a Wizard-of-Oz speech recognition technique, but all other aspects of the system functioned autonomously. Following interaction sessions, participants were asked to rate their agreement on a five-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree) for each of nine statements (provided in Table 1) covering different aspects of the interaction.

Twenty-six (12 M/14 F) participants completed one interaction session, 18 (9 M/9 F) additionally completed a second, and 7 (4 M/3 F) completed a third. The survey results, including the mode and median (ties broken by averaging) scores, the 95% confidence interval for each survey statement, and the  $p$  values resulting from a one sample  $t$ -test that compared the sample mean to an expected population mean of 3.0 (“Neither Agree Nor Disagree”) are shown in Table 1. All average scores expressed positive sentiment, and most differences between the average and the Likert scale midpoint (3.0) were statistically significant. The results establish that adults are receptive to an embodied lifelong learning companion, persistently rating it as both likeable and engaging. No sharp reductions in scores were observed over repeated sessions, which may suggest that the interactions are engaging enough to appeal to users for regular use.

## 5 Conclusions

In this work, we design and implement an embodied lifelong learning companion that engages users cognitively via human-robot book discussions. We conduct a

usability evaluation of our prototype, and find that users rate the learning companion as likeable and engaging across multiple sessions. Future work will focus on personalization and conversation quality, driving the system closer to our goal of automatically facilitating the types of conversations one might encounter during a cognitively stimulating book discussion with a human companion.

**Acknowledgements.** This material was based upon work supported by a National Science Foundation Graduate Research Fellowship under Grant 1144248, and the National Science Foundation under Grant 1262860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

1. Ahmed, I., Lubold, N., Walker, E.: ROBIN: using a programmable robot to provide feedback and encouragement on programming tasks. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 9–13. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_2](https://doi.org/10.1007/978-3-319-93846-2_2)
2. Barnes, D.E., Tager, I.B., Satariano, W.A., Yaffe, K.: The relationship between literacy and cognition in well-educated elders. *J. Gerontol. Ser. A* **59**(4), M390 (2004). <https://doi.org/10.1093/gerona/59.4.M390>, <https://doi.org/10.1093/gerona/59.4.M390>
3. Beck, I.L., McKeown, M.G.: Improving Comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach. Scholastic, New York (2006). <https://books.google.com/books?id=gcw0AAAACAAJ>
4. Berns, G.S., Blaine, K., Prietula, M.J., Pye, B.E.: Short-and long-term effects of a novel on connectivity in the brain. *Brain Connectivity* **3**(6), 590–600 (2013)
5. Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H., Barendregt, W., Nabais, F., Bull, S.: Towards empathic virtual and robotic tutors. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 733–736. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_100](https://doi.org/10.1007/978-3-642-39112-5_100)
6. Deublein, A., Pfeifer, A., Merbach, K., Bruckner, K., Mengelkamp, C., Lugrin, B.: Scaffolding of motivation in learning using a social robot. *Comput. Educ.* **125**, 182–190 (2018). <https://doi.org/10.1016/j.compedu.2018.06.015>, <http://www.sciencedirect.com/science/article/pii/S0360131518301581>
7. El Kamali, M., Angelini, L., Caon, M., Andreoni, G., Khaled, O.A., Mugellini, E.: Towards the nestore e-coach: a tangible and embodied conversational agent for older adults. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, UbiComp 2018, pp. 1656–1663. ACM, New York (2018). <https://doi.org/10.1145/3267305.3274188>, <http://doi.acm.org/10.1145/3267305.3274188>
8. Gordon, G., Breazeal, C.: Bayesian active learning-based robot tutor for children’s word-reading skills (2015). <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9280>
9. Hood, D., Lemaignan, S., Dillenbourg, P.: When children teach a robot to write: an autonomous teachable humanoid which uses simulated handwriting. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2015, pp. 83–90. ACM, New York (2015). <https://doi.org/10.1145/2696454.2696479>, <http://doi.acm.org/10.1145/2696454.2696479>

10. Kidd, C.D., Breazeal, C.: Robots at home: understanding long-term human-robot interaction. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3230–3235, September 2008. <https://doi.org/10.1109/IROS.2008.4651113>
11. Lai, V.T., Curran, T., Menn, L.: Comprehending conventional and novel metaphors: an ERP study. *Brain Res.* **1284**, 145–155 (2009)
12. Leyzberg, D., Spaulding, S., Toneva, M., Scassellati, B.: The physical presence of a robot tutor increases cognitive learning gains. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 34 (2012)
13. Li, J.: The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int. J. Hum. Comput. Stud.* **77**, 23–37 (2015). <https://doi.org/10.1016/j.ijhcs.2015.01.001>, <http://www.sciencedirect.com/science/article/pii/S107158191500004X>
14. Lu, Y., Chen, C., Chen, P., Chen, X., Zhuang, Z.: Smart learning partner: an interactive robot for education. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 447–451. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_84](https://doi.org/10.1007/978-3-319-93846-2_84)
15. Mar, R.A., Oatley, K., Hirsh, J., dela Paz, J., Peterson, J.B.: Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *J. Res. Pers.* **40**(5), 694–712 (2006)
16. Mar, R.A., Oatley, K., Peterson, J.B.: Exploring the link between reading fiction and empathy: ruling out individual differences and examining outcomes. *Communications* **34**(4), 407–428 (2009)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, USA, pp. 3111–3119 (2013)
18. Muldner, K., Lozano, C., Girotto, V., Burleson, W., Walker, E.: Designing a tangible learning environment with a teachable agent. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 299–308. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_31](https://doi.org/10.1007/978-3-642-39112-5_31)
19. Parde, N., Nielsen, R.: Automatically generating questions about novel metaphors in literature. In: Proceedings of the 11th International Conference on Natural Language Generation, pp. 264–273 (2018)
20. Parde, N., Nielsen, R.D.: A corpus of metaphor novelty scores for syntactically-related word pairs. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, pp. 1535–1540 (2018). <http://www.lrec-conf.org/proceedings/lrec2018/pdf/242.pdf>
21. Parde, N., Nielsen, R.D.: Exploring the terrain of metaphor novelty: a regression-based approach for automatically scoring metaphors. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, pp. 5336–5373 (2018)
22. Payne, B.R., Gao, X., Noh, S.R., Anderson, C.J., Stine-Morrow, E.A.: The effects of print exposure on sentence processing and memory in older adults: evidence for efficiency and reserve. *Aging Neuropsychol. Cogn.* **19**(1–2), 122–149 (2012)
23. Payne, B.R., Grison, S., Gao, X., Christianson, K., Morrow, D.G., Stine-Morrow, E.A.: Aging and individual differences in binding during sentence understanding: evidence from temporary and global syntactic attachment ambiguities. *Cognition* **130**(2), 157–173 (2014). <https://doi.org/10.1016/j.cognition.2013.10.005>, <http://www.sciencedirect.com/science/article/pii/S0010027713002072>

24. Piatt, J., et al.: Companionship with a robot? therapists' perspectives on socially assistive robots as therapeutic interventions in community mental health for older adults. *Am. J. Recreation Ther.* **15**(4), 29–39 (2017)
25. Powers, A., Kiesler, S., Fussell, S., Fussell, S., Torrey, C.: Comparing a computer agent with a humanoid robot. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, HRI 2007, pp. 145–152. ACM, New York (2007). <https://doi.org/10.1145/1228716.1228736>, <http://doi.acm.org/10.1145/1228716.1228736>
26. Schodde, T., Bergmann, K., Kopp, S.: Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2007, pp. 128–136. ACM, New York (2017). <https://doi.org/10.1145/2909824.3020222>, <http://doi.acm.org/10.1145/2909824.3020222>
27. Tanaka, F., Matsuzoe, S.: Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* **1**(1), 78–95 (2012). <https://doi.org/10.5898/JHRI.1.1.Tanaka>
28. Tapus, A., Tapus, C., Mataric, M.: The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In: 2009 IEEE International Conference on Rehabilitation Robotics, pp. 924–929, June 2009. <https://doi.org/10.1109/ICORR.2009.5209501>
29. Winkle, K., Caleb-Solly, P., Turton, A., Bremner, P.: Social robots for engagement in rehabilitative therapies: design implications from a study with therapists. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, pp. 289–297. ACM, New York (2018). <https://doi.org/10.1145/3171221.3171273>, <http://doi.acm.org/10.1145/3171221.3171273>



# Discovery of Study Patterns that Impacts Students' Discussion Performance in Forum Assignments

Bruno Elias Penteado<sup>1</sup>(✉) Seiji Isotani<sup>1</sup> Paula Maria Pereira Paiva<sup>2</sup> Marina Morettin-Zupelari<sup>2</sup> and Deborah Viviane Ferrari<sup>2</sup>

<sup>1</sup> Institute of Mathematical and Computer Sciences, University of São Paulo,  
São Carlos, SP 13566-900, Brazil  
[brunopenteado@usp.br](mailto:brunopenteado@usp.br)

<sup>2</sup> Speech Language Pathology and Audiology Department - Bauru Dental School,  
University of São Paulo, Bauru, SP 17012-901, Brazil

**Abstract.** Student-centered courses rely on the active participation of the students in forum assignments. In this work, we investigate a course where the forum assignment discusses a clinical case among professional students ( $N = 94$ ). We propose a method to discover navigation patterns related to performance grades, using behavioral actions in an LMS platform. We selected a set of significant course actions and built per-user sequences along the course module. Then, we applied the GSP algorithm to identify ordered patterns from this navigational data. The identified patterns were then used as features for a linear regression model, to predict the assignments' performance, graded manually by the teachers, and controlling for factors that may influence it. Results show some rules correlated to the students' performances. These results can be used to better inform course designers on how to improve the courseware and instructors on how to better guide their students.

**Keywords:** Study strategies · Active learning · Learning analytics

## 1 Introduction

The social constructivist theory considers that learning takes places in a social context, through relationships with other students before it is internalized [1], positing that social interactions are a crucial requirement for the development of higher order cognitive skills. The most common way to achieve this in online education is by the means of discussion forums. Previous research demonstrated the positive effect of discussion forums in student achievement [2, 3]. In this work, we sought to understand the students' trajectories, through the analysis of sequences of web pages visited before the interaction with the discussion forums, by finding behavioral patterns when the students navigated on the LMS and to explore which ones led to better grades.

Many different approaches have been used to model and understand study trajectories. Most usually, stochastic models for time series are adopted, such as Markovian models [4, 5], to model the probability of transition among different actions. However, other techniques like process mining [6, 7], sequential data mining [8], trajectory analysis [9], and interaction variable [10] were used to model different learning behaviors in different phases of computer mediated courses. Given the complexity of considering different options for the study of sequential patterns and the lack of consensus of the most appropriate methods, Van Laer [11] presented a methodological framework for the application of sequence analysis. This work proposes a method based on the analysis of sequences of actions, less computationally complex than estimations of Markovian models, yet interpretable by the impact of individual strategies, being adaptable for other LMS environments.

## 2 Methodology

The data was gathered from a specialization distance course in Audiology – *Auditory rehabilitation in children*, developed by the Speech-Language Pathology and Audiology Department (University of São Paulo in Bauru); Samaritano Association and Brazilian Ministry of Health. In total, 94 students (95,7% women, avg. age: 36.3, sd: 7.7), working in public hearing healthcare clinics across Brazilian territory, were considered. All communication and interactions between students and staff were asynchronous. The pedagogical model adopted sought to build collaborative activities so that students could share their experiences, reflect upon their practice and exchange of experiences, having the forum as a central tool. Each module of the course had an instructional sequence suggested for all the students; still, navigation within a module, as well as within its sections, was open and restricted only by the due date. This form of navigation offered students more autonomy and empowerment to choose their own learning path, according to their needs, prior knowledge, skills, and experiences, among others. In this work, we only used data from one module regarding auditory assessment in pediatric population. In the forum assignment, the participants were asked to critically discuss the pediatric protocol for behavioral auditory assessment employed in their clinics, including proposals for improvements, when necessary. Students could read their peers' posts only after submitting their first responses, which could not consist of phatic expressions only (e.g.: 'hello', 'thanks'). Instructors monitored the Forum and provided feedback, giving opportunity for students to correct and expand their responses, when necessary. The instructors also manually graded the assignment, based on the quality and completeness of the posts, using a rubric, with grading criteria and performance levels within those criteria. The dataset was composed of the log records registered by Moodle, from which only some actions were considered relevant, and encoded, as shown in Table 1. For each user, we created a sequence concatenating multiple instances of these codes throughout the module with an average length of 67 actions (sd: 43). The focus of this study is on the user discussions, operationalized by the posts

made in the module's forum. So, we segmented all these sequences in smaller subsequences with a *PostFOR* in its end. As this approach resulted in big subsequences with low support, we opted to limit to only the 10 last actions before *PostFOR*, as this number can be considered a good proxy for the student's study strategy.

**Table 1.** LMS actions encoded for this study.

Code	Action
Post FOR	The student posted a message in the Forum
Post PD	Posted a question on the Standby Support (PD in Portuguese) Forum (a specific tool for timely doubts - content or technical-wise)
View CONT	Visualized content material of the module (videos or books in Moodle)
View CTXT	Visualized the contextualization section of the problem to be discussed
View FAQ	Visualized the FAQ section, having questions about the content of the course, based on a previous edition of the course
View FOR	Visualized a discussion thread in the Forum. The main thread was created by the professor and involved the clinical case to be discussed
View MAP	Student visualized some supporting material - links or bibliographical references given by the professor or colleagues
View PD	Visualized some message posted in the Standby Support
View WIKI	Visualized some page of the wiki tool, to edit it or view it. The wiki was also used for the discussion of some procedures

The Generalized Sequential Pattern (GSP) algorithm was applied to these subsequences. GSP is a state-of-the-art algorithm for sequence mining, used in diverse educational problems (e.g. [12]). In this study, we set a support of 40%. From all the patterns identified, we selected only those which ended with an action of type *PostFOR*, with a total of 226 patterns, with lengths varying from 1 to 8 actions. The patterns were used to create a matrix  $M = |U|x|P| + 3$ , having  $u$  rows equal to the number of users and  $p$  denoting the number of patterns extracted in the previous step. The value  $M_{u,p} \in \{0, 1\}$  denotes if the pattern  $p_i$  is present for the user  $u_i$ . Besides the target variable (the grade, ranging from 0 to 10), two additional columns were used: the student's age and experience with the subject matter. The experience data were collected through a self-report instrument applied at the beginning of the course, in which the students could check multiple competencies of a list of professional backgrounds

and one of them was related to the subject of this module. These two variables were used as control variables since, in a previous study, they were the most significantly related to describing the variation of browsing patterns [9].

### 3 Results

As the number of variables (patterns) was bigger than the number of instances (users), we performed variable selection to find the best subset of variables to describe the data variance and avoid overfitting. We used a regularization technique - lasso regression from R package *glmnet* and it resulted in a model with 56 patterns and the two control variables: age and experience in the subject matter ( $R^2$ : 0.6806). Table 2 presents some of the patterns selected for the final model. Due to space limitation, we only list the top 3 patterns with highest and lowest coefficients and their respective representations. The advantage of using lasso regression was both for variable selection and shrinkage of the coefficients.

**Table 2.** Highest absolute coefficient values from the model.

Pattern ID	Coeffs	Representation
Pattern #42	0.610	<i>ViewCTXT &gt; ViewCONT &gt; PostFOR</i>
Pattern #98	0.275	<i>ViewCONT &gt; ViewCONT &gt; PostFOR &gt; PostFOR</i>
Pattern #29	0.226	<i>ViewFAQ &gt; ViewFOR &gt; PostFOR</i>
Pattern #9	-0.020	<i>ViewCTXT &gt; PostFOR</i>
Pattern #35	-0.022	<i>ViewCTXT &gt; ViewFOR &gt; PostFOR</i>
Pattern #46	-0.087	<i>ViewWIKI &gt; ViewWIKI &gt; PostFOR &gt; PostFOR</i>

### 4 Discussion

The resulting patterns suggest good face validity. For example, patterns #42 and #98 show some parts of the main sequence of the course, which was thoroughly curated by pedagogic experts, validating it. Pattern #98 suggests that the student improved his response based on the feedback of the instructor, regardless of what other strategy he might have used. The students who presented both of these sequences, in order, showed almost 1 point in the final score - having all other variables held constant. The pattern #29 is unexpected and needs more investigation, since the FAQ section was not a crucial step for the module assignment. As for the lowest coefficients, more investigation is also needed. Even with slight decreases, the use of the wiki or the context sections were not supposed to have detrimental effect. The method presented in this study can help course designers, which find this information useful for validating or correcting the flow

of the course, making easier the occurrence of some patterns, given their improvement in the grades. It can also help instructors to identify what sort of behaviors should be fostered with their students. The presence and monitoring of instructors in active learning environments like the one studied here are fundamental since they help in scaffolding students' knowledge. A limitation of this work is the limited sample scope, as only one module of the course was considered. Also, the maximum length for the subsequences was arbitrarily set; however, as the final model showed, the final subsequences were all shorter than that. As future work, we intend to expand this same method to other modules and validated it with other courses like this, where the discussion forums play a central role in the pedagogical strategy. Also, we seek to better explore the patterns generated by the method and create new information by combining them or discovering new information.

**Acknowledgements.** The ‘Auditory Rehabilitation in Children’ course was funded by the Brazilian Ministry of Health - Support Program for Institutional Development of the National Health System (Proadi/SUS - Grant 25000.024953/2015-89). The authors also thanks CNPq (Grant 307887/2017-0), CAPES and FAPESP (Grant15/24507-2) for the funding support.

## References

1. Vygotsky, L.S.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge (1978)
2. Jacob, S.M.: Mathematical achievement and critical thinking skills in asynchronous discussion forums. *Procedia Soc. Behav. Sci.* **31**, 800–804 (2012). <https://doi.org/10.1016/j.sbspro.2011.12.144>
3. Koole, S., Vervaeke, S., Cosyn, J., De Bruyn, H.: Exploring the relation between online case-based discussions and learning outcomes in dental education. *J. Dental Educ.* **78**(11), 1552–1557 (2014)
4. Köck, M., Paramythios, A.: Towards adaptive learning support on the basis of behavioural patterns in learning activity sequences. *J. Intell. Networking Collaborative Syst.* 100–107 (2010). <https://doi.org/10.1109/INCOS.2010.76>
5. Jeong, H., Biswas, G.: Mining student behavior models in learning-by-teaching environments. In: International Conference on Educational Data Mining, pp. 127–136 (2008)
6. Maldonado, J.J., Palta, R., Vázquez, J., Bermeo, J.L., Perez-Sanagustín, M., Munoz-Gama, J.: Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles. In: 42nd Latin American Computing Conference, pp. 1–12 (2016). <https://doi.org/10.1109/CLEI.2016.7833356>
7. Davis, D., Chen, G., Hauff, C., Houben G.J.: Gauging MOOC learners' adherence to the designed learning path. In: International Conference on Educational Data Mining, pp. 54–61 (2016)
8. Fournier-Viger, P., Faghihi, U., Nkambou, R., Nguifo, E.M.: CMRules: mining sequential rules common to several sequences. *J. Knowl. Based Syst.* **25**(51), 63–76 (2012). <https://doi.org/10.1016/j.knosys.2011.07.005>
9. Penteado, B.E., Isotani, S., Paiva, P.M., Morettin-Zupelari, M., Ferrari, D.V.: Detecting behavioral trajectories in continued education online courses. In: 19th IEEE International Conference on Advanced Learning Technologies (2019)

10. Venant, R., Sharma, K., Vidal, P., Dillenbourg, P., Broisin, J.: Using sequential pattern mining to explore learners' behaviors and evaluate their correlation with performance in inquiry-based learning. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 286–299. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_21](https://doi.org/10.1007/978-3-319-66610-5_21)
11. Van Laer, S.: Supporting learners in control: investigating self-regulated learning in blended learning environments. Ph.D Thesis. KU Leuven, Belgium (2018). <https://lirias.kuleuven.be/2169216?limo=0>
12. Romero, C., Ventura, S., Delgado, J.A., De Bra, P.: Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In: Duval, E., Klammer, R., Wolpers, M. (eds.) EC-TEL 2007. LNCS, vol. 4753, pp. 292–306. Springer, Heidelberg (2007). [https://doi.org/10.1007/978-3-540-75195-3\\_21](https://doi.org/10.1007/978-3-540-75195-3_21)



# Automatic Construction of a Phonics Curriculum for Reading Education Using the Transformer Neural Network

Cassandra Potier Watkins<sup>1</sup> , Olivier Dehaene<sup>2</sup>,  
and Stanislas Dehaene<sup>1,3</sup>

<sup>1</sup> INSERM, UMR992, CEA, Neurospin Center,  
University Paris Saclay, Gif-sur-Yvette, France  
`cassandra.potier-watkins@cea.fr`

<sup>2</sup> Owkin France, 75 rue de Turbigo, 75003 Paris, France

<sup>3</sup> College de France, 11 place Marcelin Berthelot, 75231 Paris Cedex 05, France

**Abstract.** Key to effective phonics instruction is the teaching of grapheme-phoneme (GP) correspondences in a systematic progression that starts with the most frequent and consistent pronunciation rules. However, discovering the relevant rules is not an easy task and usually requires subjective analysis by a native speaker and/or expert linguist. We describe GPA4.0, a submodule to the Transformer neural network model that automatizes the task of grapheme-to-phoneme (g2p) transcription and alignment. The network is trained with four different languages of decreasing orthographic transparency (Spanish < Portuguese < French < English). Our results show that the Transformer model improves on the current state-of-the-art in g2p transcription and that the attention mechanism allows for the alignment of graphemes to their corresponding phonemes. From the g2p aligned words, our software provides an optimally ordered phonics progression based on frequency and consistency in the target language, as well as an ordered list of words that teachers can use. This work exemplifies a practical way that neural networks can be used to develop educational materials for research and teachers. Submodules and phonics output are available at, <https://github.com/OlivierDehaene/GPA4.0>.

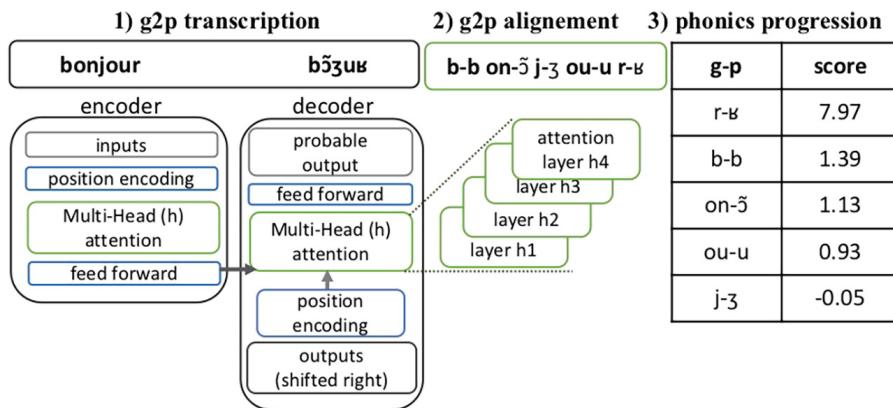
**Keywords:** Phonics instruction · g2p · Attention

## 1 Introduction

Early phonics introduction is endorsed as the foundation of successful reading instruction in both education research (meta-analysis by the National Reading Panel [1, 2]) and cognitive neuroscience [3, 4]. However, phonics instruction is not universally used. One factor for its relative disaffection could be that knowing what grapheme-phoneme (GP) correspondences to teach, and in what order to teach them, can be a difficult task, given that letter-sound relationships do not all have a one-to-one relationship. Take for example Spanish, a highly *transparent* language, meaning that a given letter is nearly always pronounced the same. In stark contrast is English, which can have many different sounds for a single grapheme (e.g. the ‘a’ in ‘cat’, ‘mate’,

‘what’ or ‘about’). Cross-language research demonstrates that orthographic transparency influences the time and difficulty children have in learning to read [5–9].

Orthographic transparency is also a conundrum in neural network text-to-speech applications that rely on grapheme-to-phoneme (g2p) transcription. G2p refers to converting words to their phonemes. The current state-of-the art applies long short-term memory (LSTM) networks and recurrent neural networks using sequence-to-sequence (seq2seq) modeling combined with an attention-mechanism [10]. More recently, the Transformer model has brought notable improvements in neural machine language transcription and language parsing [11]. These tasks are fairly analogous to g2p transcription (both depend on long range dependencies and contextual influences). Improvement made by the Transformer model is in part due to parallel position encoding that curtails the need for recurrence and a self-attention field that enables the concatenation of information between sequences, regardless of their distance. The goal of the current project, GPA4.0 (Fig. 1), is to test for g2p transcription improvements, for the first time to our knowledge, using the Transformer model. With this achieved, we take advantage of the Transformer’s attention mechanism to align grapheme input to phoneme output, thus permitting the construction of a phonics progression based on the frequency and consistency of all found GP correspondences for any alphabetic language word list.



**Fig. 1.** GPA4.0 steps to constructing a phonics progression. (1) g2p transcription is done using the Transformer neural network. (2) g2p alignment uses attention weights to align the ‘grapheme inputs’ to their ‘phoneme outputs’. (3) a phonics progression is built by according each g2p alignment an aggregated z-score based on frequency and consistency in the word corpus.

## 2 Experiment

We tested the Transformer model for improved g2p transcription compared to the current-state-of-the-art results [10, 12] on the CMUDict database [13] while also comparing, for the first time to our knowledge, the results of five different languages of varying orthographic transparency: Spanish < Portuguese < French < English. Training was done using one 1080TI NVIDIA GPU on the base models for a total of 10,000 steps. We use Tensor2Tensor (T2T) [14] an open-source system for training deep learning models in TensorFlow [15]. G2p alignment in our model is made possible

using the attention weights of the Transformer model. G2p alignment accuracy was analyzed in French, the only language for which we had a reference for comparison. Table 1 describes the word lists used and provides the minor adjustments made to accommodate the small amount of training data. Training was conducted on 80% of the data. The model’s performance was tested on the remaining 20% of data.

To generate a language’s phonics progression, we extract all the GP correspondences in the list of g2p aligned words. For each GP correspondence found, we measure its frequency, g2p consistency and phoneme-to-grapheme consistency. The GP correspondences are then sorted by an aggregate weight of the prementioned measures’ z-scores (we apply weights of 0.7, 0.25 and 0.05 respectively, but these can be adjusted in the code). The weights are designed to 1) give priority to the most frequent GP correspondences when a pair is particularly consistent and less frequent but highly consistent correspondences.

**Table 1.** Language wordlists used and adjustments made to the Transformer architecture

Language	Number of words used for training	Number of words used for testing	Number of hidden layers	3
Spanish [16]	10,400	2,600	Hidden size, number of neurons per layer	256
Portuguese [17]	31,200	7,800	Filter size	512
French [18]	8,000	2,000	h, number of attention heads	4
English [19]	8,000	2,000	Attention dropout rate	0.2
English [13]	95,069	23,767	Dropout rate	0.3

### 3 Results

#### 3.1 g2p Transcription

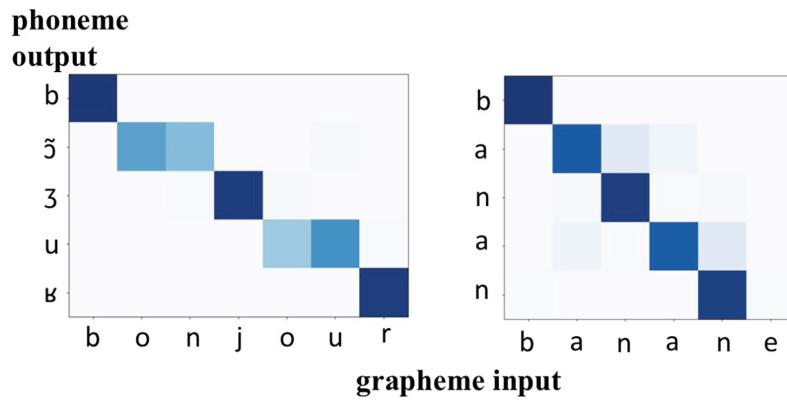
The standard measures of word error rate (WER) and phoneme error rate (PER) are reported in Table 2. WER is the total number of output errors in which there is at least one phoneme error/total number of words. PER is the Levenshtein distance [20] (the minimum number of single-character edits needed to change one word to the other) of the predicted phoneme sequence to the reference from the original database/the number of phonemes in the reference. Language WER and PER scores reflect, as expected, decreasing orthographic transparency. We report a slight gain over Toshniwal and Livescu’s best prior score on the CMUdict database.

**Table 2.** Word error rate (**WER**) and Phoneme error rate (**PER**) in four languages of decreasing orthographic transparency

	Spanish<	Portuguese<	French<	English	CMUDict
WER	0.38%	2.77%	3.18%	15.04%	20.87%
PER	0.07%	0.55%	0.89%	4.50%	4.59%
Previous best results using the CMUDict database:				WER = 21.69%	
				PER = 5.04%	

### 3.2 g2p Alignment

Figure 2 provides an example of encoder-decoder attention (taken from layer-4 multi-head attention in the decoder, see Fig. 1). As the network reads the word “bonjour” or “banane”, it attends to distant information required to know if a vowel followed by the letter ‘n’ will make a single nasal sound (e.g. ‘on’) or two distinct phonemes (e.g. a + n). GPA4.0 aligns graphemes to phonemes based on the attention carrying the most weight. G2p alignment error rate was assessed for French using the sequence error rate, a correct or incorrect score for each word and the g2p alignment error rate (Levenshtein distance [20]). We report scores of 27.76% and 10.20% respectively. The relatively high sequence error rate compared to the low g2p alignment score is due to the difficulty in parsing silent letters not coded in the phonology of the trained wordlist. 56% of words in the list contain silent letters.



**Fig. 2.** Encoder-decoder attention in g2p transcription

## 4 Conclusion

Our results demonstrate improved g2p transcription by the Transformer model. Our submodule, GPA4.0, takes a novel approach to developing applicable phonics tools for the classroom by taking advantage of neural network performance in g2p transcription and, in particular, the attention field for g2p alignment. This work highlights the difficulties for neural networks to learn the GP correspondences in decreasingly transparent languages. The phonics progressions for the four languages analyzed and their ordered wordlists are freely available. These datafiles can be used as a ‘paper’

support to guide reading instruction, or as stimuli for game-based reading applications (e.g. the GraphoGame software [3, 21]). We hope that the GPA4.0 submodule will be taken up as a tool for researchers and educators to generate their own phonics lessons with 100% decodable reading materials. GPA4.0 combines cognitive science and neural network technology for evidence-based reading education. Phonics progressions and word lists for the four different languages analyzed in this paper, as well as the GPA4.0 submodule code, can be downloaded at <https://github.com/OlivierDehaene/GPA4.0>.

## Bibliography

1. Cunningham, J.W.: The national reading panel report. *Reading Res. Q.* **36**, 326–335 (2001). <https://doi.org/10.1598/RRQ.36.3.5>
2. Castles, A., Rastle, K., Nation, K.: Ending the reading wars: reading acquisition from novice to expert. *Psychol. Sci. Public Interest* **19**, 5–51 (2018). <https://doi.org/10.1177/1529100618772271>
3. Brem, S., et al.: Brain sensitivity to print emerges when children learn letter–speech sound correspondences. *Proc. Nat. Acad. Sci.* **107**, 7939–7944 (2010). <https://doi.org/10.1073/pnas.0904402107>
4. Dehaene, S., et al.: How learning to read changes the cortical networks for vision and language. *Science* 1194140 (2010). <https://doi.org/10.1126/science.1194140>
5. Seymour, P.H.K., Aro, M., Erskine, J.M.: Foundation literacy acquisition in European orthographies. *Br. J. Psychol.* **94**, 143–174 (2003). <https://doi.org/10.1348/000712603321661859>
6. Goswami, U., Gombert, J.E., de Barrera, L.F.: Children’s orthographic representations and linguistic transparency: nonsense word reading in English, French, and Spanish. *Appl. Psycholinguistics* **19**, 19–52 (1998). <https://doi.org/10.1017/S0142716400010560>
7. Landerl, K.: Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *Eur. J. Psychol. Educ.* **15**, 239 (2000). <https://doi.org/10.1007/BF03173177>
8. Serrano, F., et al.: Variations in reading and spelling acquisition in Portuguese, French and Spanish: a cross-linguistic comparison. *J. Portuguese Linguist.* **10**, 183–204 (2011). <https://doi.org/10.5334/jpl.106>
9. Ziegler, J.C., et al.: Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. *Psychol. Sci.* **21**, 551–559 (2010). <https://doi.org/10.1177/0956797610363406>
10. Toshniwal, S., Livescu, K.: Jointly learning to align and convert graphemes to phonemes with neural attention models. [arXiv:1610.06540](https://arxiv.org/abs/1610.06540) [cs] (2016)
11. Vaswani, A., et al.: Attention is all you need. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) [cs] (2017)
12. Rao, K., Peng, F., Sak, H., Beaufays, F.: Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In: Presented at the April (2015). <https://doi.org/10.1109/ICASSP.2015.7178767>
13. Weid, R.L.: The CMU pronouncing dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
14. Vaswani, A., et al.: Tensor2Tensor for neural machine translation. [arXiv:1803.07416](https://arxiv.org/abs/1803.07416) [cs, stat] (2018)
15. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous distributed systems, vol. 19. Software available from [tensorflow.org](https://tensorflow.org) (2015)

16. Corral, S., Ferrero, M., Goikoetxea, E.: LEXIN: a lexical database from Spanish kindergarten and first-grade readers. *Behav. Res. Methods* **41**, 1009–1017 (2009). <https://doi.org/10.3758/BRM.41.4.1009>
17. Derived Corpora and Counts. <https://childe.talkbank.org/derived/>
18. Lété, B., Sprenger-Charolles, L., Colé, P.: MANULEX: a grade-level lexical database from French elementary school readers. *Behav. Res. Methods Instrum. Comput.* **36**, 156–166 (2004)
19. Masterson, J., Stuart, M., Dixon, M., Lovejoy, S.: Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *Br. J. Psychol.* **101**, 221–242 (2010). <https://doi.org/10.1348/000712608X371744>
20. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **10**, 707–710 (1966)
21. Richardson, U., Lyytinen, H.: The GraphoGame method: the theoretical and methodological background of the technology-enhanced learning environment for learning to read. *Hum. Technol.* **10** (2014). <https://doi.org/10.17011/ht.urn.201405281859>



# An Annotation Protocol for Collecting User-Generated Counter-Arguments Using Crowdsourcing

Paul Reisert<sup>1,3(✉)</sup>, Gisela Vallejo<sup>2</sup>, Naoya Inoue<sup>1,3</sup>, Iryna Gurevych<sup>2</sup>,  
and Kentaro Inui<sup>1,3</sup>

<sup>1</sup> RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan  
[paul.reisert@riken.jp](mailto:paul.reisert@riken.jp)

<sup>2</sup> Ubiquitous Knowledge Processing Lab (UKP), Department of Computer Science,  
Technische Universität Darmstadt, Darmstadt, Germany  
[{vallejo,gurevych}@ukp.informatik.tu-darmstadt.de](mailto:{vallejo,gurevych}@ukp.informatik.tu-darmstadt.de)

<sup>3</sup> Tohoku University, Sendai, Japan  
[{naoya-i,inui}@ecei.tohoku.ac.jp](mailto:{naoya-i,inui}@ecei.tohoku.ac.jp)

**Abstract.** Constructive feedback is important for improving critical thinking skills. However, little work has been done to automatically generate such feedback for an argument. In this work, we experiment with an annotation protocol for collecting user-generated counter-arguments via crowdsourcing. We conduct two parallel crowdsourcing experiments, where workers are instructed to produce (i) a counter-argument, and (ii) a counter-argument after identifying a fallacy. Our analysis indicates that we can collect counter-arguments that are useful as constructive feedback, especially when workers are first asked to identify a fallacy type.

**Keywords:** Critical thinking · Counter-argument · Fallacy ·  
Crowdsourcing · Annotation study · Constructive feedback

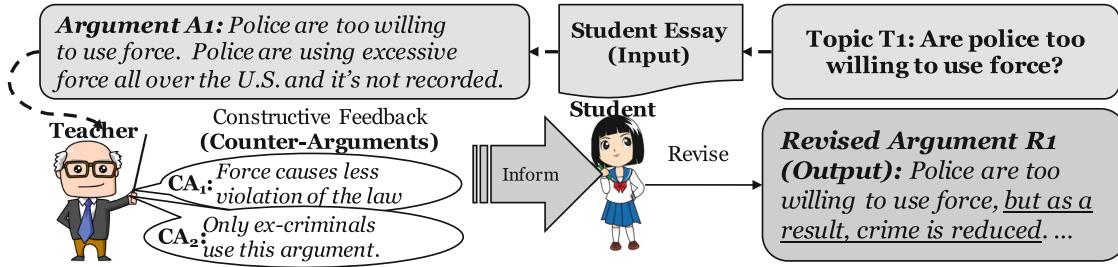
## 1 Introduction

Automatic essay scoring is the task of automatically evaluating a wide-range of essay criteria in a pedagogical context, such as organization [10], self-directed learning [7], thesis clarity [11] and author stance [12]. Several works have also integrated argumentative features [2, 8, 13] for evaluation. Applications such as Grammarly<sup>1</sup> and eRater<sup>2</sup> have received wide attention for automatically assess the contents of an essay.

An example of the usefulness of constructive feedback is shown in Fig. 1. In response to the *topic*,  $T_1$ , the argument  $A_1$  extracted from a student's essay. In response to  $A_1$ , a teacher would provide constructive feedback to the student for

<sup>1</sup> <https://www.grammarly.com/>.

<sup>2</sup> <https://www.ets.org/erater>.



**Fig. 1.** Example of argument revision via constructive feedback.

improving their argument (e.g.,  $CA_1$  &  $CA_2$ ). Afterwards, a student could revise their argument to produce a stronger one (i.e.,  $R_1$ ) and improve their critical thinking skills for future essays.

We aim to create a method for improving automatic constructive feedback generation, which can help reduce time for graders and allow writers to instantly learn their mistakes. Towards this goal, fallacy detection and counter-arguments have been shown to be useful. Habernal et al. [3] created a game which allowed users to identify fallacies. In the pedagogical context, several studies have identified common fallacies in student essays [1, 6, 9]. For counter-arguments, Wachsmuth et al. [14] created a task for retrieving the best counter-argument for a given argument, and Hua and Wang [5] generated counter-arguments by extracting external evidence. However, it still remains an open issue as to how to create a corpus useful for modeling constructive feedback.

In this work, we conduct two parallel crowdsourcing experiments in order to determine if a large-scale, high-quality corpus of user-generated counter-arguments required for modeling constructive feedback can be created. We instruct non-expert workers (i) to produce counter-arguments simply given an argument, and (ii) to produce a counter-argument after identifying a specified fallacy type. We then conduct an analysis on the collected counter-arguments for determining their usefulness. Our results suggest that workers can produce useful counter-arguments, especially when instructed to identify a fallacy type.

## 2 Collecting User-Generated Counter-Arguments

### 2.1 Data and Crowdsourcing

We conduct our experiments on top of the Argument Reasoning Comprehension (ARC) corpus [4]. ARC contains 2,477 context-independent, micro-level (i.e., single claim and premise) arguments with 172 diverse topics, making the corpus ideal for modeling constructive feedback.

We use the crowdsourcing platform Figure Eight<sup>3</sup> for quickly collecting counter-arguments. We assume that a large-scale corpus of counter-arguments can be produced by non-expert crowdworkers with appropriate guidelines.

<sup>3</sup> <http://www.figure-eight.com>.

**Table 1.** CAG-F distribution and inner-annotator agreement between annotators  $\langle \mathbf{A}, \mathbf{B} \rangle$ .

Fallacy type	Yes	No	Unsure	Cohen's $\kappa$
Appeal to common practice	$\langle 13, 17 \rangle$	$\langle 5, 1 \rangle$	$\langle 2, 2 \rangle$	0.44
Begging the question	$\langle 14, 18 \rangle$	$\langle 6, 2 \rangle$	$\langle 0, 0 \rangle$	0.41
Hasty generalization	$\langle 15, 15 \rangle$	$\langle 4, 5 \rangle$	$\langle 1, 0 \rangle$	0.68
Questionable cause	$\langle 15, 14 \rangle$	$\langle 4, 4 \rangle$	$\langle 1, 2 \rangle$	0.46
Red herring	$\langle 15, 17 \rangle$	$\langle 4, 2 \rangle$	$\langle 1, 1 \rangle$	0.49
Agreed instances	64	10	0	

**Counter-Argument Generation Without Fallacy Identification (CAG).** We first conduct trial experiments on Figure Eight for calibrating appropriate interface, guidelines, and settings. Per given *topic*, the worker is shown a *claim* and *premise* and instructed to produce a sentence-long counter-argument that attacks one or both of them. We use the following settings for CAG: 10 second *minimum time per instance*, level 3 annotators (i.e., high-quality), and \$0.10 per answer.

**Counter-Argument Generation with Fallacy Identification (CAG-F).** We conduct a parallel experiment in which crowdworkers were asked if a pre-specified fallacy type exists in the original argument. We randomly select 5 fallacy types and their examples from SoftSchools<sup>4</sup>: *appeal to common practice*, *begging the question*, *hasty generalization*, *questionable cause*, and *red herring*. We create separate crowdsourcing jobs for each fallacy type. Workers are instructed to answer if the fallacy type exists, and if so, they are asked to produce a counter-argument. We use the same settings as CAG. However, annotators are not required to write a counter-argument if they select *no* or *unsure*, so we award each answer with \$0.05 and offer workers a bonus if they produce a *good* counter-argument.

## 2.2 Annotation Statistics

For CAG, we collect 100 user-generated counter-arguments for 100 arguments. The time to complete the experiment was roughly 2.5 h. For each of the 5 jobs in CAG-F, we employed 5 crowdworkers per argument (100 arguments total). The average time to complete each experiment was roughly 1.3 h.

## 3 Analysis and Discussion

We conduct a qualitative analysis using two annotators specializing in the field of argumentation. One annotator created the crowdsourcing guidelines and conducted the experiments. We asked both annotators to judge the quality of CAG

---

<sup>4</sup> <http://www.softschools.com/examples/fallacies/>.

**Table 2.** Examples of CAG and CAG-F counter-arguments agreed as *yes*.

Claim	Premise	CAG	CAG-F
Unpaid internship exploit college students	Interns are replacing employees	<i>unpaid internship offer students chance of getting experience and therefore do not exploit them</i>	<i>its too hasty to say that all Interns are replacing employees</i> ( <b>Hasty Generalization</b> )
Home schoolers deserve a tax break	Home schooled children should get the same state financial backing given to public school attendees [...]	<i>most of the time they may not get equal education facilities of public attendees</i>	<i>no tax relief is needed because there are no real costs for such learning.</i> ( <b>Begging the Question</b> )

and CAG-F counter-arguments by the following: *Is the counter-argument attacking the claim, premise, or both?*, and *Using the counter-argument, could you make the original argument better?*. If one answer was *no*, the counter-argument was labeled as *no*. For CAG, we have both annotators answer the above questions for the 100 counter-arguments. For CAG-F, for each of the 5 fallacy types, we randomly select 20 arguments with a unique topic to the fallacy type, where some arguments are shared across different fallacy types.

### 3.1 Results

Table 1 shows the distribution of answers and the inner-annotator agreement for CAG-F, and Table 2 shows examples from both stages. For CAG, the Cohen’s kappa<sup>5</sup> ( $\kappa$ ) between both annotators is 0.29, which is slightly lower than CAG-F (0.37). In total, 74 (64 *yes* and 10 *no*) instances were agreed upon, indicating a slight improvement (20%) over CAG.

**Disagreements.** For CAG, we observed all but one instance of the 21 instances labeled as *no* by one annotator (**B**) were labeled as *no* by the other (**A**). When observing the 20 remaining instances labeled as *no* by **A**, we found that most were labeled as a *simple contradiction, unrelated, or incomprehensible/ungrammatical*. We believe this attributes to the fact that **A** created the guidelines and experiments and was more critical of the quality. For CAG-F, we observed that **A** labeled *no* 3 times when **B** said *yes*. We discovered that the reasons are *agreeing stance, irrelevant, and untrue* (e.g., “*Cyclists have nothing to do with bike lanes*”). **B** said *no* 11 times when **A** answered *yes* with the following reasons: *non-counter-argument, untrue, and unclear*.

## 4 Conclusion

Towards automatically generating constructive feedback, in this work, we experimented with constructing an annotation protocol for collecting user-generated

---

<sup>5</sup> We calculate the Cohen’s kappa after filtering out *unsure* instances.

counter-arguments via crowdsourcing. We conducted two parallel crowdsourcing experiments where, given an argument, workers were instructed to (i) produce a counter-argument, and (ii) first identify a fallacy type and then produce a counter-argument. Our results indicate that we can collect counter-arguments useful as constructive feedback in both settings, especially when workers were instructed to first identify a fallacy in the original argument.

## References

1. El Khoiri, N., Widiati, U.: Logical fallacies in Indonesian EFL learners' argumentative writing: students' perspectives. *Dinamika Ilmu* **17**(1), 71–81 (2017)
2. Ghosh, D., Khanam, A., Han, Y., Muresan, S.: Coarse-grained argumentation features for scoring persuasive essays. In: Proceedings of the 54th Annual Meeting of ACL (Volume 2: Short Papers), pp. 549–554 (2016)
3. Habernal, I., Pauli, P., Gurevych, I.: Adapting serious game for fallacious argumentation to German: pitfalls, insights, and best practices. In: Proceedings of the Eleventh International Conference on LREC, pp. 3329–3335 (2018)
4. Habernal, I., Wachsmuth, H., Gurevych, I., Stein, B.: The argument reasoning comprehension task: identification and reconstruction of implicit warrants. In: Proceedings of the 2018 Conference of NAACL: HLT, Volume 1 (Long Papers), pp. 1930–1940. Association for Computational Linguistics (2018)
5. Hua, X., Wang, L.: Neural argument generation augmented with externally retrieved evidence. In: Proceedings of the 56th Annual Meeting of ACL (Volume 1: Long Papers), pp. 219–230 (2018)
6. Indah, R.N., Kusuma, A.W.: Fallacies in English department students' claims: a rhetorical analysis of critical thinking. *Jurnal Pendidikan Humaniora* **3**(4), 295–304 (2015)
7. Lucas, C., Gibson, A., Buckingham Shum, S.: Utilization of a novel online reflective learning tool for immediate formative feedback to assist pharmacy students' reflective writing skills. *Am. J. Pharm. Educ.* (2018)
8. Nguyen, H.V., Litman, D.J.: Argument mining for improving the automated scoring of persuasive essays. In: The Thirty-Second AAAI Conference on Artificial Intelligence, pp. 5892–5899 (2018)
9. Oktavia, W., Yasin, A., et al.: An analysis of students' argumentative elements and fallacies in students' discussion essays. *Engl. Lang. Teach.* **2**(3) (2014)
10. Persing, I., Davis, A., Ng, V.: Modeling organization in student essays. In: Proceedings of the 2010 Conference on EMNLP, pp. 229–239. Association for Computational Linguistics (2010)
11. Persing, I., Ng, V.: Modeling thesis clarity in student essays. In: Proceedings of the 51st Annual Meeting of ACL (Volume 1: Long Papers), vol. 1, pp. 260–269 (2013)
12. Persing, I., Ng, V.: Modeling stance in student essays. In: Proceedings of the 54th Annual Meeting of ACL (Volume 1: Long Papers), vol. 1, pp. 2174–2184 (2016)
13. Wachsmuth, H., Al-Khatib, K., Stein, B.: Using argument mining to assess the argumentation quality of essays. In: Proceedings of the 26th International Conference on COLING, pp. 1680–1692 (2016)
14. Wachsmuth, H., Syed, S., Stein, B.: Retrieval of the best counterargument without prior topic knowledge. In: Proceedings of the 56th Annual Meeting of ACL (Volume 1: Long Papers), vol. 1, pp. 241–251 (2018)



# Towards an Automatic Q&A Generation for Online Courses - A Pipeline Based Approach

Sylvio Rüdian<sup>1,2(✉)</sup> and Niels Pinkwart<sup>1,2</sup>

<sup>1</sup> Humboldt-Universität zu Berlin, Berlin, Germany

ruediasy@informatik.hu-berlin.de

<sup>2</sup> Weizenbaum Institute for the Networked Society, Berlin, Germany

**Abstract.** Personalization of online courses is one of the challenges of the 21st century. Although different methodologies for personalization in educational contexts are already existing, there is a bottleneck: personalization by context is always limited to existing learning material; creation of those is a time-consuming task. In this paper we introduce a pipeline to generate questions and valid answers based on educational texts, limited to factual questions for given sentences. We combined NLP technologies with an efficient methodology that is normally used in bioinformatics and adjusted it to generate Q&A-pairs. Instructors can suggest corrections in natural language. Our system generates questions and corresponding answers based on sentences of which 70% make sense.

**Keywords:** Personalization · Online courses · Q&A-generation · POS · NLP

## 1 Introduction

Personalization of online courses is a current trend designed for optimizing learning by adapting it to individuals [1]. One class of personalization of learning environments focuses on content-based recommendations to support learners' needs. Typically, educational recommender systems personalize online courses based on already existing learning material. For instructors, the creation of this material (much more than what would be needed for a non-adaptive course) can be a highly time-consuming task. If we do not have the resources to create learning material for all learners' individual needs, the provision of personalized learning via content specific to individuals will necessarily be limited. On the other hand, the Internet contains a large amount of high quality resources, which could be used as a basis to create online courses. Such a creation requires different steps that range from the selection or creation of appropriate learning materials up to their evaluation. During conversations with instructors that use Learning Management Systems like Moodle we identified that the creation of interactive media and quizzes is the most time consuming task, whereas finding existing textual material online can be done faster as there are lots of descriptive texts available on the Internet. Many online courses use multiple choice questions as an essential question type. A subset of these questions is used to remember and recall facts and

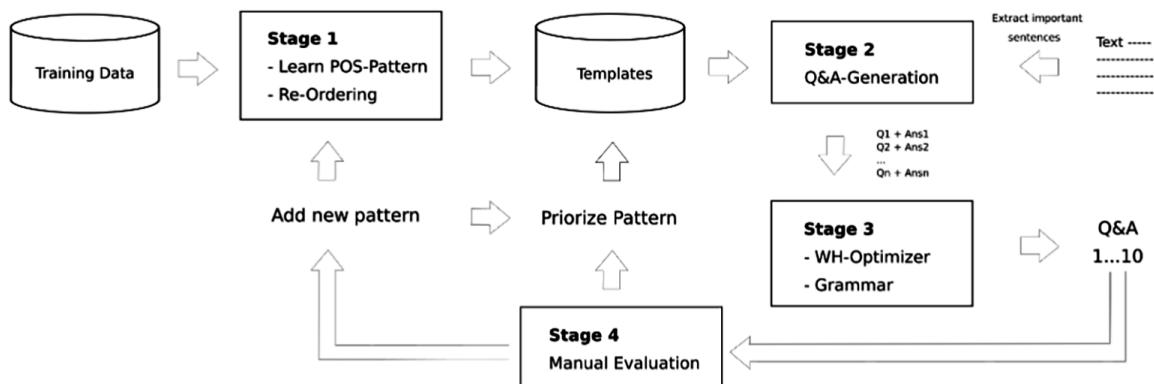
basic concepts [2]. We show that most of these questions can be generated automatically to help instructors.

## 2 Related Work

Heilman et al. [3] introduced an approach of question generation via overgenerating, transformations and ranking. They used sentences as inputs and generated multiple questions following a rule-based approach. Therefore, they manually improved rules and expressions, applied manual conditions and defined 12 features to rank generated questions with a supervised approach. 43.3% of generated questions were acceptable. In 2010, Heilman et al. [4] simplified sentences first and used manually defined rules to transform sentences into questions. The acceptance rate of generated questions reached 52%. Le et al. [5, 6] created an educational question generator which is dependent on existing templates and databases that contain structured knowledge. These approaches require human cognitive skills to create templates.

Zhao et al. [7] investigated an approach that makes use of automatically generated templates for creating questions. Templates contain placeholders that can be replaced by entities, but they do not generalize well. Rodrigues et al. [8] introduced a framework to generate questions based on different levels of linguistic information. They used triples as training data, with each triple consisting of a question, an answer, and a text snippet that answers the question. Their final model consists of 23 semantic patterns that can be used to generate Q&A pairs, but no studies on the generation quality were published.

We introduce a pipeline to generate factual questions and corresponding answers for online courses based on texts and learned triples [8]. We show how instructors can use the system for adding improvements by using natural language without the need for expert knowledge in creating templates or rules. The approach of Heilman et al. [4] achieved an acceptance rate of 52%, which is not good enough for generating learning materials in practical settings. Our aim is to improve this level and at the same time offer a methodology that allows instructors to participate and improve results.



**Fig. 1.** Pipeline of Q&A-generation

### 3 Methodology

Our pipeline for Q&A generation consists of four steps (Fig. 1). We used the Stanford Question Answering Dataset (SQuAD 1.1 [9]), extracted 10,000 questions and created triples like in [8]. We created a mapping of sentences and answers. Each sentence's words were mapped to a new position of corresponding questions. This reordering of sequences can be used to generate a Q&A pair based on a sentence. The same is done by mapping sentences and their corresponding answers. We trained our model by learning the structure of all sentences. We replaced all words and symbols with the Part-of-Speech representation (POS). Compared to most Q&A generation approaches, this approach does not require manual rule/template creation. We generalized our training data to avoid splitting related words like “New York” and noun phrases by integrating some rules (e.g.  $[Determiner(DT), X] \rightarrow X$  or  $[X, X] \rightarrow X$ ) to combine related words. This kind of an entity recognizer reduced templates by an average of 26.3% (Standard Deviation: 9.5%).

To generate questions and answers from a text, we have to extract sentences that contain the most important information to ask for. For this task, we can use existing text summarizers. Now we use each extracted main sentence separately. Our general idea is to find an already known POS sequence that is most similar to the new sentence's POS sequence. This sequence can be used to apply the learned mapping with the new sentence, which is enriched by its POS tags and generalized as we did before. We used the End-Gap Free Alignment approach that is often used in bioinformatics to find non exact matches of DNA or proteins [10]. Instead of nucleotides (A, G, T, C) we used POS tags. The approach is useful; it avoids splitting the sequence into several sub pieces. In addition, the algorithm calculates a score that represents the quality of the match. Due to a limited complexity of  $O(n * m)$  [10], the matching score can be calculated fast. If the score is calculated using any known POS sequence, the resulting list can be sorted by score to obtain the most fitting ones. Now we can use each sequence and apply the learned mapping to generate Q&A pairs. As language is complex, our approach works very well for short sentences, because there are plenty of learned POS patterns which are similar to new short sentences. The approach is helpful to find sub phrases as well. Yet, as sentences become longer, richer POS patterns are required, which are hard to learn from training corpora. To overcome this problem, we propose an approach to extend learning data and to improve generated Q&A pairs. Instructors can validate them by choosing the best of ten generated Q&A pairs. If there is a sentence whose POS tags look like the manually chosen one, the previously best chosen Q&A pair receives a better rank due to its prioritization. If none of the questions is good enough to be used, the user can add a question and corresponding answer manually. Resulting Q&A pairs can thus become better with increasing usage. For instructors, this type of interaction is much more acceptable than creating templates or handcrafted rules, because they are able to improve the system by using natural language.

## 4 Results

We used 100 sentences from [11], where sentences with general knowledge were given that include facts we wanted to ask for. 6 participants had the task to label the meaningfulness of every generated Q&A pair (binary choice: meaningful or not). From 100 Q&A pairs, 70% were rated to be meaningful on average (standard deviation SD: 5.3), with an inter-rater-reliability of 0.84. For 30% of all questions, participants gave some proposals to improve the corresponding Q&A pair's grammar. Additionally, we split sentences by length to distinguish between long and short sentences. As assumed, shorter sentences (up to 10 words) perform better (86.7% meaningful questions, SD 0.06) than longer sentences (55.8% meaningful, SD 0.07).

Our acceptance rate (70%) was higher than those reported by Heilman et al. (52%) by using SQuAD to train the model. The acceptance rate for longer sentences is also better than in the previous approaches. The correction of Q&A pairs requires less time than creating new ones, so finally we believe that our approach can support instructors during the process of creating online courses. Our evaluation does not contain any manual training, results might become better, the more the system will be used.

## 5 Conclusion and Further Research

Most online courses use multiple choice questions for self-assessment and final exams. Their creation is expensive and still requires human effort. We introduced a pipeline to generate questions and corresponding answers based on textual inputs. We limited this study to factual questions at sentence level [2]. The question generator becomes more valuable if other question types of Blooms Taxonomy [2] can be generated as well. Thus the study should be extended to use more than one sentence, which can be solved by applying the End-Gap Free Alignment, too. As we used pre-defined sentences to generate Q&A pairs, we cannot evaluate whether the generated Q&A pairs are acceptable from an educational perspective. This evaluation can be done with generated questions, where the generator extracts sentences of texts that are used to generate Q&A pairs, which is the scope of further research. To create MC questions, we need the context of texts to provide wrong answers as well. Therefore, structured domain knowledge from existing databases or the WordNet is required [12], which is not addressed in this paper. Instructors have the possibility to use the Q&A generator and can improve the system without knowing the processes behind the system. We were able to achieve an acceptance rate of 70% for generated Q&A pairs without having used the opportunity to get improvements by instructors. Although created Q&A pairs might not be perfect, even semi-automatically produced pairs that require manual correction can be more time-efficient than creating Q&A pairs manually [13]. This result shows that our Q&A generator can be used to assist instructors in creating online courses.

**Acknowledgments.** This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII116 (Weizenbaum-Institute). The responsibility for the content of this publication remains with the authors.

## References

1. National Academy of Sciences: Advance personalized learning NAE Grand Challenges for Engineering, Updated 2017, pp. 45–47 (2008)
2. Bloom, B.: Bloom’s Taxonomy of Educational Objectives, Vol. 1: Cognitive Domain, New York, McKay (1965)
3. Heilman, M., Smith, N.A.: Question generation via overgenerating transformations and ranking. Pittsburgh, Language Technologies Institute (2009)
4. Heilman, M., Smith, N.A.: Good question! Statistical ranking for question generation. In: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, HLT 2010, pp. 609–617 (2010)
5. Le, N.-T., Shabas, A., McLaren, P.: QUESGEN: a framework for automatic question generation using semantic web and lexical databases. In: Spector, J.M., et al. (eds.) Frontiers of Cyberlearning. LNET, pp. 69–89. Springer, Singapore (2018). [https://doi.org/10.1007/978-981-13-0650-1\\_4](https://doi.org/10.1007/978-981-13-0650-1_4)
6. Le, N.-T., Shabas, A., Pinkwart, N.: A question generation framework for teachers. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 182–186. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_33](https://doi.org/10.1007/978-3-319-93846-2_33)
7. Zhao, S., Wang, H., Li, C., Liu, T., Guan, Y.: Automatically generating questions from queries for community-based question answering. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, AFNLP, pp. 929–937 (2011)
8. Rodrigues, H.P., Coheur, L., Nyberg, E.: QGASP: a framework for question generation based on different levels of linguistic information. In: Proceedings of the 9th International Natural Language Generation conference, Edinburgh, UK, pp. 242–243. Association for Computational Linguistic (2016)
9. Rajpurkar, P., Jian Zhang, K.L., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Conference on Empirical Methods in Natural Language Processing (2016)
10. Altschul, S.F., Erickson, B.W.: Optimal sequence alignment using affine gap costs. Bull. Math. Biol. **48**, 603–616 (1986)
11. LM Digital Media: General Knowledge Quiz for kids to Intrigue their Senses, 11 October 2017. [kidsworldfun.com](http://kidsworldfun.com). Accessed 01 Feb 2018
12. Fellbaum, C.: WordNet. In: Chapelle, C. (ed.) The Encyclopedia of Applied Linguistics. Blackwell Publishing Ltd., Chichester (2012)
13. Karamanis, N., Ha, L.A., Mitkov, R.: Generating multiple-choice test items from medical text: a pilot study. In: Proceedings of the Fourth International Natural Language Generation Conference, Sydney, pp. 111–113. Association for Computational Linguistics (2006)



# Semantic Matching of Open Texts to Pre-scripted Answers in Dialogue-Based Learning

Ştefan Ruşeti<sup>1</sup>, Raja Lala<sup>2</sup>, Gabriel Guțu-Robu<sup>1</sup>, Mihai Dascălu<sup>1(✉)</sup>,  
Johan Jeuring<sup>2</sup>, and Marcell van Geest<sup>2</sup>

<sup>1</sup> Computer Science Department, University Politehnica of Bucharest,  
Bucharest, Romania

{stefan.ruseti, gabriel.gutu, mihai.dascalu}@cs.pub.ro  
<sup>2</sup> Computer Science Department, Utrecht University, Utrecht, The Netherlands  
{r.lala, j.t.jeuring}@uu.nl, marcell@marcell.nl

**Abstract.** Gamification is frequently employed in learning environments to enhance learner interactions and engagement. However, most games use pre-scripted dialogues and interactions with players, which limit their immersion and cognition. Our aim is to develop a semantic matching tool that enables users to introduce open text answers which are automatically associated with the most similar pre-scripted answer. A structured scenario written in Dutch was developed by experts for this communication experiment as a sequence of possible interactions within the environment. Semantic similarity scores computed with the SpaCy library were combined with string kernels, WordNet-based distances, and used as features in a neural network. Our experiments show that string kernels are the most predictive feature for determining the most probable pre-scripted answer, whereas neural networks obtain similar performance by combining multiple semantic similarity measures.

**Keywords:** Answer matching · Semantic similarity · Natural Language Processing · Neural network

## 1 Introduction

Serious games incorporated in various learning environments are usually aimed at stimulating users' creativity, as well as their engagement. However, most games frequently use pre-scripted interactions that require the specific selection of one option from a list of predefined potential candidates or actions; in return, this approach limits players' immersion and cognition. Our aim is to address this limitation by enabling learners to type free input answers, that are afterwards mapped onto existing alternatives defined within the game.

This study explores different Natural Language Processing (NLP) techniques for matching free-text student responses to pre-scripted answers in a Dutch serious game. The game is based on a communication scenario in which a player converses with a virtual character throughout a simulation. The entire scenario is scripted by an expert as a sequence of potential interactions and questions that form a decision tree with

branches corresponding to pre-scripted answers [1]. Instead of selecting a reply from a list of predefined answers given the sequence of questions, users are now encouraged to write their responses, thus providing them with freedom in writing their own responses; in return, these are mapped to the pre-scripted scenario answers.

Nevertheless, we must emphasize from the beginning a limitation of the matching process, namely that both players' and pre-scripted answers are short [2], which in return limits the performance of some NLP methods. Our aim is to explore different semantic relatedness methods and potential manners in which they can be effectively combined in order to best match responses and augment the existing rule-based system incorporated in most serious games.

The problem tackled in this paper is similar to an answer selection task in question answering, if we consider the candidate replies from our scenario as the possible answers. Several datasets exist for English that cover different versions of this problem, like SQuAD [3], MCTest [4], or InsuranceQA [5]. However, these datasets are significantly larger and the complex deep learning models that obtain the best results on these tasks cannot be applied in our case. Thus, we focused mostly on unsupervised methods.

## 2 Method

Our dataset was gathered in guided sessions with students who played our serious game and provided free-text inputs throughout their gameplay. In addition, players were given the list of pre-scripted answers after providing their text inputs, as well as a “no match” option when their answer was unrelated to any pre-scripted alternative and were asked to select the option which was closest to their answer. The user inputs contained 52.34 characters/9.84 words on average, while pre-scripted answers were similar in size, but still short having limited contextual information: 59.33 characters and 10.44 words on average. Two experts annotated each student's answer by matching themselves all responses to the closest corresponding pre-scripted answers from a semantic point of view. There were 1,143 evaluations overall, out of which 974 cases were kept based on a majority agreement criterion (i.e., two or more people agree out of the initial players and the two experts). These items were used in the experiments that follow. We ran a two-way random effects model of ICC and Cronbach's alpha which denoted acceptable agreement (Cronbach's alpha of .777) and a high average ICC measure of .742.

The following splitting procedure was used. We considered the two most-answered questions and the two least-answered questions to be outliers and put them in the training set. We were left with 20 questions which were ordered by the number of matching items. We assigned consecutive groups of five questions randomly to training (3), testing (1), and validation (1), thus resulting in a dataset with 12 training, 4 testing and 4 validation questions, each set having a significant number of matching items.

We considered several semantic models in order to maximize the matching process. First, SpaCy (<https://spacy.io>) is an advanced NLP framework written in Python, which contains a very fast syntactic parser designed for production usage. It incorporates pre-trained Dutch semantic models for part-of-speech tagging and dependency parsing. SpaCy computes similarity scores based on the cosine similarity of average word

vectors of two texts. Second, WordNet is a lexicalized ontology whose Dutch version, the Open Dutch WordNet, contains more than 115,000 *synsets* (i.e., sets of similar words) and corresponding relationships [6]. Semantic distances between words available for Dutch include path length [7], the Leacock-Chodorow and the Wu-Palmer methods [8].

Third, string kernels compute a similarity between two texts by counting common character  $n$ -grams, without the need for any language-specific tools. This method performs well when comparing texts without the need for a large training set [9]. Different scores can be computed by varying the size of the  $n$ -grams or by changing the way the sum is computed. The most common types of string kernels are *presence*, *intersection* and *spectrum* [10], each representing different ways of computing character  $n$ -grams overlap. When evaluated as a single method, we computed the average of the three types of string kernels for  $n$ -grams ranging in size between 3 to 7 characters.

Given the scores computed with each individual method described above, one possible way of improving the performance of the system is to compute an aggregate score. We implemented a neural network (NN) with one hidden layer that computes the best combination of scores. Several experiments were performed on the validation set to select the most relevant features and hyper-parameters of the network. The network receives as input in the training phase two pairs containing a candidate answer and a given answer, one being a positive match, the other negative (either it matches another candidate or doesn't match anything). The network computes a score for each pair and learns to separate them as much as possible. While considering string kernels as features for the neural network, we computed each of the three types with different values for  $n$ -gram sizes, namely: 2–3, 4–5, 6–7, and 8–9.

### 3 Results

Given the matches annotated by experts, we evaluate the performance of each method based on the following three types of accuracy: (a) accuracy when a pre-scripted answer is matched (1-match) – 147 out of 224 input texts; (b) accuracy for not matching any pre-scripted answer (no match) – 77 out of 224 input texts; and (c) global accuracy. Results on the test data are presented in Table 1. The neural network combination was trained on both the training and validation partitions after selecting the best configuration on the validation dataset. The threshold used to determine if there is a match was selected based on the validation data.

The neural network combination obtained the highest overall score, but with only a small improvement compared to the String Kernels method (only one more correct example), which seems to be the best method for this task, by far. One possible explanation for the success of the String Kernels is its ability to detect common keywords (in different forms) in the two texts, while not being influenced by the other words in the sentence. All the other methods take into account all the words in the text by using an average over individual word pairs. String Kernels also have the advantage of working at character-level, thus being more suitable to cases when users provide short answers.

**Table 1.** Accuracies for the semantic methods applied on the test data.

Method	1-match	No match	Global accuracy
SpaCy	(38/147) 26%	(77/77) <b>100%</b>	(115/224) 51%
WN path length	(25/147) 17%	(74/77) 96%	(99/224) 44%
WN Leacock Chodorow	(19/147) 13%	(74/77) 96%	(93/224) 42%
WN Wu-Palmer	(19/147) 13%	(76/77) 99%	(95/224) 42%
String kernels	(72/147) <b>49%</b>	(64/77) 83%	(136/224) 61%
Average of SpaCy and string kernels	(33/147) 22%	(77/77) <b>100%</b>	(110/224) 49%
Neural network	(72/147) <b>49%</b>	(65/77) 84%	(137/224) <b>61%</b>

In general, direct subword matching is advantageous in cases where the testing domain has a different lexical distribution to the background data used to develop a matching model (e.g., word embedding data). It is likely that partitioning the data set based on scenario questions may have had an effect on the nature of responses between data partitions, especially in terms of word overlap. Moreover, the validation set is more skewed towards *No match* cases, which in turn may bias methods tuned on this set. With these factors in mind, it appears that methods that don't rely on prior knowledge perform better on the test set. For a general-purpose matching method, string kernels proved to be the best option between the selected methods. However, there is clearly still quite a lot of room for improvement.

## 4 Conclusions

This paper describes the research of text-matching methods for mapping open text input to predefined scripted dialogue response options. We implemented a number of domain-independent text-matching methods including WordNet semantic distances and string kernels, as well as corpora dependent methods (i.e., spacy models). We evaluated these alongside a neural network which integrates our text matching scorers. Overall, the NN combination method achieved the best performance on our test data. However, its performance was quite close to string kernels. Given the additional overhead required to train and run the NN, it appears that string kernels are the best option for integrating a generic, domain independent text-matcher into a serious game.

Iterative improvement of the dialogue design using text analysis methods appears to be a promising way to help ensure open text inputs are dealt with appropriately. For example, shaping the dialogue to encourage user responses to be more specific to the topic discussed will likely make matching easier and semantic models more useful.

We also expect that incorporating more dialogue context into text matching methods may be beneficial when enough consistent user data is available. In this case, we could make more use of the sequence-to-sequence methods that drive many conversational AI chatbots, without sacrificing control of the dialogue structure.

**Acknowledgments.** This activity has received funding from the European Institute of Innovation and Technology (EIT). This body of the European Union receives support from the European Union's Horizon 2020 research and innovation programme. This research was also partially supported by the 644187 EC H2020 RAGE project.

## References

1. Jeuring, J., et al.: Communicate! — a serious game for communication skills —. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 513–517. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24258-3\\_49](https://doi.org/10.1007/978-3-319-24258-3_49)
2. Holtgraves, T., Han, T.L.: A procedure for studying online conversational processing using a chat bot. *Behav. Res. Methods* **39**(1), 156–163 (2007)
3. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint [arXiv:1606.05250](https://arxiv.org/abs/1606.05250) (2016)
4. Richardson, M., Burges, C.J.C., Renshaw, E.: MCTest: a challenge dataset for the open-domain machine comprehension of text. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, Washington, USA, pp. 193–203. ACL (2013)
5. Feng, M., Xiang, B., Glass, M.R., Wang, L., Zhou, B.: Applying deep learning to answer selection: a study and an open task. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 813–820. IEEE (2015)
6. Postma, M., van Miltenburg, E., Segers, R., Schoen, A., Vossen, P.: Open Dutch WordNet. In: Global WordNet Conference, p. 300, January 2016
7. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: 21st International Conference on AAAI, Boston, Massachusetts, vol. 1, pp. 775–780. AAAI Press (2006)
8. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. *Comput. Linguist.* **32**(1), 13–47 (2006)
9. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. *J. Mach. Learn. Res.* **2**, 419–444 (2002)
10. Ionescu, R.T., Popescu, M., Cahill, A.: Can characters reveal your native language? A language-independent approach to native language identification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1363–1373 (2014)



# Developing Game-Based Models of Cooperation, Persistence and Problem Solving from Collaborative Gameplay

Maria Ofelia Z. San Pedro<sup>1(✉)</sup>, Ruitao Liu<sup>1</sup>,  
and Tamera L. McKinniss<sup>2</sup>

<sup>1</sup> ACT, Inc., Iowa City, USA

{sweet.san.pedro, ruitao.liu}@act.org

<sup>2</sup> Iowa City, USA

**Abstract.** Collaborative-problem solving (CPS) is an important 21st century skill and it continues to be a complex skill to model and assess. We approached this challenge by first looking at the individual level primary cognitive and social aspects of CPS. This paper demonstrates ongoing work of designing and developing game-based models of three CPS components: cooperation, problem-solving, and persistence. A study was conducted collecting data from the game-play of 11 groups (three middle school students in each group) tasked with solving challenges in Physics Playground. We employed evidence-centered design principles to develop behavioral indicators of cooperation, problem-solving and persistence. These were used to code each student's behavior during three hours of video-recorded gameplay. For each CPS component, we applied hierarchical clustering on this video-coded data and qualitatively evaluated two generated clusters of students across groups. For cooperation, there was more communication with other students in working towards a solution for one group. For problem-solving, one group had more instances of talking about possible solutions. For persistence, one group had more attempts in a challenge and was more on-task. Implications of results, limitations and future work were discussed.

**Keywords:** Cooperation · Persistence · Problem-solving · Evidence-centered design · Collaborative problem solving · Video coding · Hierarchical clustering

## 1 Introduction

Collaborative problem solving (CPS) is a problem-solving strategy where a number of participants (2 or more) have shared experiences related to a given problem state and goal, and involves cognitive and social skills to achieve a solution [1–3]. However, CPS continues to be a complex skill to model and assess. Current work that assesses collaboration often involve simulations, games and team-based activities [4, 5] which

---

T. L. McKinniss—Independent Scholar.

engage students as well as provide opportunities for evidence that can be used to measure skills that are hard to address with more familiar item types. In this paper, we present preliminary qualitative findings from ongoing work to design and develop models for three CPS components: cooperation, persistence and problem-solving, using a collaborative game activity as context. Limitations and future work are provided.

## 2 Methods

We used an evidence-centered design (ECD) approach [6] in defining the constructs that comprised the CPS student model (cooperation, problem-solving, persistence). We designed the tasks, student roles, and instructions to elicit evidence of each construct, and facilitate the measurement of these constructs. We defined indicators using in-game and out-of-game observed behaviors. The study involved three hours of collaborative gameplay from 11 teams of 3 middle students randomly assigned based on availability. Students were instructed to work together to solve problems when playing the educational game Physics Playground (formerly known as Newton's Playground in [7]). Each team consisted of 7th or 8th grade students, with a mix of male and female students for some teams. The goal in each challenge was to move the green ball to hit the red balloon by drawing simple machine agents (ramp, lever, pendulum, and springboard). Each member of a team was assigned a specific role (e.g., player, questioner, and recorder) in each challenge, rotating roles for each new challenge. Teams were instructed to verbally discuss ways to solve the challenges and what happened as a result of their actions. Gameplay sessions for each team were video recorded using two camera views (front focused on faces, rear capturing gestures) and a computer screen capture. Together with the knowledge of the game features and the CPS constructs themselves, evidence rules were designed to define behavioral indicators for each construct. Using the resulting rubrics for cooperation, problem-solving and persistence, the behaviors for each student were coded [8, 9] by trained raters (inter-rater reliability of above 0.70). A mark was noted for each occurrence of a behavior and different behavioral indicators were linked to each construct. We applied agglomerative clustering based on Ward's algorithm [10] to extract student clusters (across all 33 students) for cooperation, problem-solving and persistence. The clustering process used the sum indicator occurrence for each student across eight identified games (played by all groups). Below is a list of the final indicators used to code behavior for each construct from the video data, as well as the resulting dendrogram of student clusters for each construct.

## 3 Results

The resulting behavioral indicators of cooperation, problem-solving and persistence (Table 1) were iteratively engineered and mapped by researchers based on existing frameworks of each construct [3, 11] and behaviors commonly observed during group gameplay by the students. The behavioral indicators across all three constructs can be categorized by students discussing solutions, action steps taken by students to solve the

challenge, student behaviors after solving a challenge, and student interactions with one another. Using this rubric, expert raters observed the videos and coded instances when these indicators were observed for each student in each challenge they attempted to solve.

**Table 1.** ECD-based behavioral indicators of cooperation, problem-solving and persistence

#	Behavioral indicators	Cooperation	Problem-solving	Persistence
1	Talks about the challenge situation			X
2	Generally talks about possible solutions			X
3–6	Talks about using a ramp/lever/pendulum/springboard to solve the Challenge			X
7–9	Talks about the weight/mass/height/length of an object needed			X
10	Builds on ideas/provides ideas for improving attempted solutions	X		X
11	Provides reasons to support implementing a potential solution/action			X
12	Asks questions about why a solution should be tried or what took place			
13	Talks about results after implementing the solution			
14	Provides information to the Recorder to help complete the Challenge Log	X		
15	Makes an initial attempt after discussion	X		X
16	Tries again after discussion		X	
17	Tries again without discussion		X	
18	Completes the Challenge Log			X
19	Player asks if others want to try taking action in the game		X	
20	Asks to take action in the game before Player asks for help	X		
21–23	Brings up leaving a Challenge before solving it/trying for a Gold after receiving a Silver/trying for a trophy on a different agent		X	
24	Visibly not focused on the game activities and assigned role		X	
25	Initiates off-topic conversation or other distractions during the game		X	
26	Joins off-topic conversations during the game		X	

(continued)

**Table 1.** (*continued*)

#	Behavioral indicators	Cooperation	Problem-solving	Persistence
27	Does not respond when spoken to by others	X		
28	Willing to compromise during disagreements	X		
29	Tries to confirm understanding of what others said by paraphrasing	X		
30	Compliments or makes encouraging comments to another team member	X		
31	Makes fun of, criticizes, or is rude to others	X		
32	Interrupts or talks over other students	X		

After conducting the clustering process, two student clusters emerged and were qualitatively evaluated for cooperation, problem-solving and persistence (looking at the average instances per behavioral indicator and inspection of its values per cluster). From the resulting dendrogram for cooperation (dendrogram figures are not included in this paper due to page limits), one student cluster (13 students) exhibited more instances of communicating by building and improving ideas from others (Indicator 4 in Table 1), talking about solutions (indicator 2) and discussing before attempting a challenge (indicators 7, 8) compare to the other student cluster (20 students). For problem-solving, one student cluster (10 students) emerged to have more discussions in coming up with a solution for the challenges using physics terms (indicators 2 to 8) and providing reasons for a solution or action (indicator 11) than the other student cluster (23 students). Lastly, for persistence, one small student cluster emerged (2 students) that exhibited far higher instances of attempting a challenge (with or without discussion, indicators 16, 17) and far lower instances of engaging in off-topic conversation (indicators 25, 26).

## 4 Discussion and Future Work

We present in this paper the creation of game-based behavioral indicators for the CPS components cooperation, problem-solving and persistence, in the context of collaborative gameplay. An evidence-centered design approach was used. This study is part of ongoing work to develop valid measurement models for each CPS construct and CPS itself. Findings presented in this paper include mapping of these behavioral indicators to each construct and preliminary analysis of video-coded data using these indicators. This qualitative analysis generated two student clusters for each component and showed how these designed indicators were able to distinguish student groups based on construct definitions. For example in cooperation, indicators related to communication were more evident in one group than the other. In persistence, indicators related to

attempts were more evident in one group than the other. And in problem-solving, discussing physics-related concepts in solving the challenge was more evident in one group. Knowing such prevalent indicators per construct may be useful in potentially designing AI-driven pedagogical agents that can evaluate CPS competencies and provide scaffolds in team-based learning activities. Although a limitation in this study included a relatively low number of students, there were numerous identified indicators for each construct. The next phase in this work includes (1) creating Item Response Theory (IRT) models to measure each component and compare observations to self-reported construct measures, and (2) including game-log information in the creation of the CPS models (i.e., multimodal data analysis).

**Acknowledgments.** We are grateful to Prof. Valerie Shute and Dr. Lubin Wang from Florida State University for their support in this study.

## References

1. Griffin, P., Care, E. (eds.): *Assessment and Teaching of 21st Century Skills: Methods and Approach*. EAIA. Springer, Dordrecht (2015). <https://doi.org/10.1007/978-94-017-9395-7>
2. Organization for Economic Co-operation and Development (OECD): *PISA 2015 collaborative problem solving frameworks* (2013). <http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm>
3. Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P.: A framework for teachable collaborative problem solving skills. In: Griffin, P., Care, E. (eds.) *Assessment and Teaching of 21st Century Skills*. EAIA, pp. 37–56. Springer, Dordrecht (2015). [https://doi.org/10.1007/978-94-017-9395-7\\_2](https://doi.org/10.1007/978-94-017-9395-7_2)
4. Liu, L., Hao, J., von Davier, A.A., Kyllonen, P., Zapata-Rivera, J.D.: A tough nut to crack: measuring collaborative problem solving. In: *Handbook of Research on Technology Tools for Real-World Skill Development*, pp. 344–359. IGI Global (2016)
5. Chang, C.J., et al.: An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations. *J. Comput. Assist. Learn.* **33**(6), 649–662 (2017)
6. Mislevy, R.J., Riconscente, M.M.: Evidence-centered assessment design: layers, concepts, and terminology. In: Downing, S., Haladyna, T. (eds.) *Handbook of Test Development*, pp. 61–90. Erlbaum, Mahwah (2006)
7. Shute, V.J., Ventura, M.: *Stealth Assessment: Measuring and Supporting Learning in Video Games*. MIT Press, Cambridge (2013)
8. Leighton, J.P.: Avoiding misconception, misuse, and missed opportunities: the collection of verbal reports in educational achievement testing. *Educ. Measur. Issues Pract.* **23**(4), 6–15 (2004)
9. Saldaña, J.: *The Coding Manual for Qualitative Researchers*. Sage, London (2015)
10. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. SSS. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-84858-7>
11. Camara, W., O'Connor, R., Mattern, K., Hanson, M.A.: *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success*. ACT Research Report Series 2015 (4). ACT, Inc. (2015)



# An Intelligent-Agent Facilitated Scaffold for Fostering Reflection in a Team-Based Project Course

Sreecharan Sankaranarayanan<sup>(✉)</sup>, Xu Wang, Cameron Dashti, Marshall An,  
Clarence Ngoh, Michael Hilton, Majd Sakr, and Carolyn Rosé

Carnegie Mellon University, Pittsburgh, USA

{sreechas,xuwang,cdashti1,haokanga,pcn,mhilton,msakr,cprose}@cs.cmu.edu  
<http://teel.cs.cmu.edu/>

**Abstract.** This paper reports on work adapting an industry standard team practice referred to as Mob Programming into a paradigm called Online Mob Programming (OMP) for the purpose of encouraging teams to reflect on concepts and share work in the midst of their project experience. We present a study situated within a series of three course projects in a large online course on Cloud Computing. In a  $3 \times 3$  Latin Square design, we compare students working alone and in two OMP configurations (with and without transactivity-maximization team formation designed to enhance reflection). The analysis reveals the extent to which grading on the produced software rewards teams where highly skilled individuals dominate the work. Further, compliance with the OMP paradigm is associated with greater evidence of group reflection on concepts and greater shared practice of programming.

**Keywords:** Online Mob Programming · Project-based learning · Computer-supported collaborative learning · Conversational agents

## 1 Introduction and Prior Work

Although team-project based courses are valued as opportunities to integrate and apply knowledge while refining skills learned in more basic courses, the anecdotal experience of many is that the inherent reward structure resulting from grades based on the quality of the end-project fosters a performance orientation, where the most capable students take on the lion's share of the work, doing tasks they already know how to do well, while undercutting the opportunity for other students to practice and for the group to reflect on underlying concepts. The contribution of this paper is a new intelligent-agent enabled paradigm for project based teamwork in computer science courses that is based on an adaptation of an industry standard team practice referred to as Mob Programming, with the goal of combating this tendency, and fostering more equal sharing of practice opportunities and group reflection on concepts.

Online Mob Programming is adapted from the industrial practice of Mob Programming, where a group of 3–6 participants rotate through four roles in order to afford participants the opportunity to experience the work from distinctly different vantage points [11, 12]. Each participant will experience several rounds of all the roles throughout a single mob programming session, getting an opportunity to contribute as well as observe different perspectives and approaches to solve problems in the same session. The roles include - **Mob**: A participant or group of participants who consider and deliberate between multiple alternative implementations ultimately informing the decision of the Navigator. **Navigator**: A single participant who solicits input from the Mob, decides on the next action and communicates that to the Driver to be implemented into code. **Driver**: A single participant who converts high-level instructions from the Navigator into code. **Facilitator**: A single participant who observes and intervenes when necessary, such as to indicate when roles are to switch and to keep the activity progressing. This role is taken up by an Intelligent Conversational Agent, which monitors group processes [6, 8] and supports the uptake and rotation of roles.

Prior work on in-person Mob Programming describes benefits including a structured process for utilizing distributed knowledge, a unanimous positive perception of the process from the knowledge sharing, learning, and developer satisfaction perspectives [5], and the ability to learn from more experienced developers [3]. Remote Mob Programming [7] has also been attempted.

OMP participants collaboratively code in the online AWS Cloud9 IDE which includes an editor, terminal, text-chat and file navigation all on one screen. The intelligent conversational agent based on the open-source Bazaar framework [1] is integrated into the text-chat and uses a combination of static scripts that structure the activity, and dynamic role assignment based on the number of students in the chat room at any given time. Additionally, the agent receives data from the chat and the code edits to determine instances of compliant behavior associated with each role and highlights them as examples that participants can emulate. The introduction of the agent opens up the possibility of more dynamic context-sensitive conversational support for students and their roles in the future.

## 2 Method

Because of the distribution of responsibilities to roles in an inter-dependent fashion, we can hypothesize that - **Hypothesis 1** - Teams that demonstrate increased compliance to the OMP paradigm will discuss project-relevant conceptual content more substantively, contribute work towards the group solution more evenly, and produce a group product with as high of quality as individuals or teams with lower compliance. In designing this current study, we build on prior work developing a team formation strategy that provides benefits associated with idea sharing [10] and reduced problems with distribution of labor and conflict [9]. This team assignment paradigm uses a measure of observed exchange of transactive discussion [2, 4] as an estimate of pairwise collaboration potential between students and then groups teams within a class in such a

way as to maximise the estimated pairwise collaboration potential within teams across the class as a whole [10]. Transactivity as a conversational construct can potentially interface well with the hypothesized benefits of OMP including more even distribution of work and more substantive discussions. Thus, **Hypothesis 2** - Groups formed transactively will demonstrate higher compliance with OMP practices that will then be associated with an intensification of the observed benefits of OMP compliance.

In order to test the two above hypotheses, we experimentally contrast the mob programming scaffold in randomly formed and transactively formed groups against individual programming in a 3 (Condition)  $\times$  3 (Programming Assignment) Latin Square between-subjects design embedded within a completely online, graduate Cloud Computing course offered to the students of Carnegie Mellon University and its campuses worldwide. Students first participate in a training activity with random groups. They are then assigned to one of three tracks within which they are assigned to individual, transactively formed, and random teams with 4–5 students per team such that no two students who have worked together in one group are together in another (including the training session). The conditions are then counterbalanced across tracks in order to prevent ordering effects. A total of 120 students took the course allowing 40 students to be assigned to each condition for each exercise. Each activity, including a training activity in the first week of the course, lasted 80 min with 10 min reserved for introductions and wrap-up and roles switching every 7 min. The role-switching was kept relatively frequent in order to promote observation of the problem from multiple perspectives.

### 3 Results and Discussion

Code contributions, chat logs, grades, post-assignment and post-course survey data was collected in all conditions to facilitate our analyses. In order to quantify whether students complied with the structure suggested by the OMP activity, we calculated a *Compliance Score* which was measured as the ratio of code contributions by the driver to the average code contributions by the other team members. Since the driver is expected to make all of the code contributions, a higher compliance score constitutes more compliance with the OMP structure. We further calculated the percentage of code contributions made by each group member which was used to compute an *Evenness Deviation* score, which measured the difference between this percentage and what percentage would be observed for the group if work was distributed evenly. Finally, in order to quantify the extent of activity-relevant *Conceptual Content* being discussed in the chat, we measured the vector similarity of the topic representation of the chat of a student with the primer corresponding to that activity using a latent semantic indexing model with the number of topics set to 5. A higher document similarity score meant that more of the conceptual content from the primer was discussed in the chat.

We begin by checking to ensure that students in all three conditions achieved equivalent grades on the assignments, and indeed, there was no significant difference. In order to test the extent to which the reward structure substantially does encourage an uneven distribution of labor, we computed an ANCOVA model with a three-way split on the Evenness Deviation variable (Top Quartile, Middle, Lower Quartile) as the independent variable, average grade prior to the experiment as Covariate, Assignment time point as a random variable, and Grade at time point as the Dependent variable. The upper quartile had deviation scores of higher than .33, and the lower quartile had deviation scores less than .08. The median deviation score was .2. In 3% of teams, a single member did all of the work. There was no significant effect of the deviation variable on grade, though the trend was in the expected direction both for the auto graded and manually graded portions of the assignment. Thus, students may falsely believe it is necessary to deviate from an even distribution of labor when in fact it does not help their grade.

We first tested for an association between Compliance scores with Conceptual Content scores and found the association to be highly significant ( $R = .26$ ,  $p < .005$ ) such that students in more compliant groups focused more on conceptual content in their chats. Then we tested for an association between Compliance scores and Evenness Deviation scores. In this model, there was no main effect of condition, and in Random teams, there was no effect of Compliance, but in Transactive teams, there was a marginal effect such that more compliant teams had a lower Evenness Deviation score  $F(1,149) = 1.93$ ,  $p = .055$ ). We also tested for an effect of compliance on grade and there was no significant or marginal effect of Compliance in either condition. Thus, we have correlational, though not causal, evidence to support the first hypothesis that Mob practices are associated with more conceptual focus and more even distribution of labor without harm to grade on assignment.

To test the second hypothesis we computed an ANOVA model with Condition as the independent variable and Assignment time point as a random variable. Compliance score was the dependent variable. Here we found a trend consistent with the hypothesis, but it was not significant. Thus, Hypothesis 2 is not supported, though above we indicated that Transactivity maximized teams showed a more even distribution of labor when they complied, whereas the Random assigned teams did not. It is possible that the OMP structure acts as its own scaffold for idea exchange, which might explain why the primary enhancement we observe of the Transactivity maximization condition is that students in that condition cooperated better in terms of division of labor.

Overall, this study provides correlational evidence in favor of OMP as a set of team practices that might serve to counter the performance focus of team behavior in project courses and instead encourage teamwork and reflection on project relevant concepts. One limitation of the current study is that all of the teams were trained and supported in OMP practices. Without manipulating whether OMP was encouraged, we lack causal evidence in favor of the value of OMP. Thus, a follow-up study is necessary to obtain such evidence.

This research was funded in part by the National Science Foundation grants ACI-1443068, IIS 1546393, and IIS 1822831 as well an AWS Educate Grant, Microsoft Azure Educator Grant Award and a Google Cloud Platform Grant.

## References

1. Adamson, D., Dyke, G., Jang, H., Rosé, C.P.: Towards an agile approach to adapting dynamic collaboration support to student needs. *Int. J. Artif. Intell. Educ.* **24**(1), 92–124 (2014)
2. Azmitia, M., Montgomery, R.: Friendship, transactive dialogues, and the development of scientific reasoning. *Soc. Dev.* **2**(3), 202–221 (1993)
3. Buchan, J., Pearl, M.: Leveraging the mob mentality: an experience report on mob programming. In: Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018, pp. 199–204. ACM (2018)
4. Gweon, G., Jain, M., McDonough, J., Raj, B., Rosé, C.P.: Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *Int. J. Comput.-Support. Collaborative Learn.* **8**(2), 245–265 (2013)
5. Kattan, H.M., Oliveira, F., Goldman, A., Yoder, J.W.: Mob programming: the state of the art and three case studies of open source software. In: Santos, V.A., Pinto, G.H.L., Serra Seca Neto, A.G. (eds.) WBMA 2017. CCIS, vol. 802, pp. 146–160. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-73673-0\\_12](https://doi.org/10.1007/978-3-319-73673-0_12)
6. Kumar, R., Rosé, C.P., Wang, Y.C., Joshi, M., Robinson, A.: Tutorial dialogue as adaptive collaborative learning support. *Front. Artif. Intell. Appl.* **158**, 383 (2007)
7. Malmgren, U.: Remote mob programming set up, January 2017. <https://natooktesting.wordpress.com/2017/01/30/remote-mob-programming-set-up/>
8. Rosé, C.P., Ferschke, O.: Technology support for discussion based learning: from computer supported collaborative learning to the future of massive open online courses. *Int. J. Artif. Intell. Educ.* **26**(2), 660–678 (2016)
9. Sankaranarayanan, S., Dashti, C., Bogart, C., Wang, X., Sakr, M., Rosé, C.P.: When optimal team formation is a choice - self-selection versus intelligent team formation strategies in a large online project-based course. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 518–531. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_38](https://doi.org/10.1007/978-3-319-93843-1_38)
10. Wen, M., Maki, K., Dow, S., Herbsleb, J.D., Rose, C.: Supporting virtual team formation through community-wide deliberation. In: Proceedings of the ACM on Human-Computer Interaction (CSCW), vol. 1 (2018)
11. Wilson, A.: Mob programming - what works, what doesn't. In: Lassenius, C., Dingsøyr, T., Paasivaara, M. (eds.) XP 2015. LNBP, vol. 212, pp. 319–325. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-18612-2\\_33](https://doi.org/10.1007/978-3-319-18612-2_33)
12. Zuill, W., Meadows, K.: Mob programming: a whole team approach. In: Agile 2014 Conference, Orlando, Florida (2016)



# I Wanna Talk Like You: Speaker Adaptation to Dialogue Style in L2 Practice Conversation

Arabella J. Sinclair<sup>1</sup>(✉), Rafael Ferreira<sup>1</sup>, Dragan Gašević<sup>1,2</sup>,  
Christopher G. Lucas<sup>1</sup>, and Adam Lopez<sup>1</sup>

<sup>1</sup> University of Edinburgh,

Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK

{s0934062,dragan.gasevic,alopez,cglucas2}@ac.ed.uk

<sup>2</sup> Monash University, 19 Ancora Imparo Way, Clayton, VIC 3800, Australia

dragan.gasevic@monash.edu

**Abstract.** We present a novel method to analyse speaker alignment in second language practice dialogue. Our method represents utterances as Dialogue Acts and applies Epistemic Network Analysis to their use. ENA makes convergence between speakers visible, and enables us to confirm hypotheses that both initial similarity and final convergence increase with student ability; and that Dialogue Act use changes with ability, and over the course of an interaction. Our results can inform personalised automatic tutoring tools as well as formative assessment and feedback.

**Keywords:** Epistemic Network Analysis · Scaffolding · Dialogue · Natural language processing · Alignment · Zone of proximal development

## 1 Introduction

One-to-one spontaneous dialogue practice is important for Second Language (L2) learning in both classrooms and online learning platforms. It has been shown to provide greater opportunities for L2 learning [1, 2, 7, 9, 15] since learners improve from practice and from observation of their interlocutors. We use Dialogue Acts (DAs) [8] to label utterance roles and analyse alignment at this level. DAs are used to infer discourse structure, and for automatic understanding of spontaneous dialogue [19]. DAs provide a high level, topic-agnostic representation. We use Epistemic Network Analysis (ENA) [16] to model speaker DA usage within L2 dialogues at different levels, quantifying dialogic contribution. We investigate the following research questions: RQ1: *What is the relationship between DA symmetry and student ability?* and RQ2: *How does DA usage change over the*

---

We acknowledge useful input from Ed Fincham, Pablo Leon, Nicolas Collignon, Kate McCurdy, Clara Vania, Naomi Saphra, Toms Bergmanis, Sameer Bansal, Ida Szubert, Maria Corkery, Federico Fancellu and our anonymous reviewers.

*course of a dialogue?* We hypothesise DA usage will be more similar as student ability improves, as speaker contributions become increasingly symmetric [4] and that speakers will converge within a single dialogue [18]. Alignment consists of interlocutor interaction adaptation, resulting in convergence, or in their sharing of the same concept space [3, 6, 12]. Typically, alignment is measured at either a *lexical* (use of the same words [17, 21] or phrases [5] as each other) or a *syntaxtic* (parts of speech patterns e.g. similar noun-phrase constructions, or similar adjuncts [13, 14]) level. Measurement methods range from count statistics [5] to linear regression on prime-to-target distance<sup>1</sup> [20] to using general linear mixed models to account for the random effects present in dialogue [14, 17].

Our work contributes to the literature on speaker adaptation within L2 dialogue, providing evidence to support our hypothesis that alignment can be seen at the level of DAs both with increasing ability level and across dialogues. We present a novel method for modelling dialogue contribution by combining the descriptive powers of ENA with DAs. This has implications for formative assessment in an instructional setting, and continuous feedback for tutors and students. Our work also has implications for (i) the design of learning analytic tools, (ii) informing tutoring strategy, and (iii) the design of automatic tutoring systems.

## 2 Data and Methods

The Barcelona English Language Corpus (BELC) [11] consists of 118 transcripts (of length 60–140 utterances) from English learner conversational practice. Tutors’ instructions were to elicit as much naturalistic conversation as possible, following a similar script. It is divided into four general levels of student ability, from beginner to intermediate. We use DA annotations [18], chosen from [19] for their relevance to the corpus. Table 1 shows DA labels and example dialogues from the highest and lowest level students, demonstrating differences in DA use.

Epistemic Network Analysis (ENA) [16] is a graph-based analysis method which captures relationships between different concepts (*codes*, in our case DAs) within an *analysis unit* (speakers at each level) in textual datasets. Codes are considered related if they appear in the same *stanza*: full dialogue (RQ1), or dialogue quarters divided by number of utterances (RQ2). Each utterance is represented as a vector of the presence (1) or absence (0) of each code. A co-occurrence matrix is derived for each dialogue from these code vectors. Dimensionality reduction is performed using Singular Value Decomposition (SVD) [10] representing the projection graph in a two dimensional space [svd1, svd2].

## 3 Results and Discussion

To answer RQ1, stanza is the *full dialogue*, the unit of analysis *speaker* and *ability level*. Figure 1(a) shows individual speakers’ networks at different abilities.

---

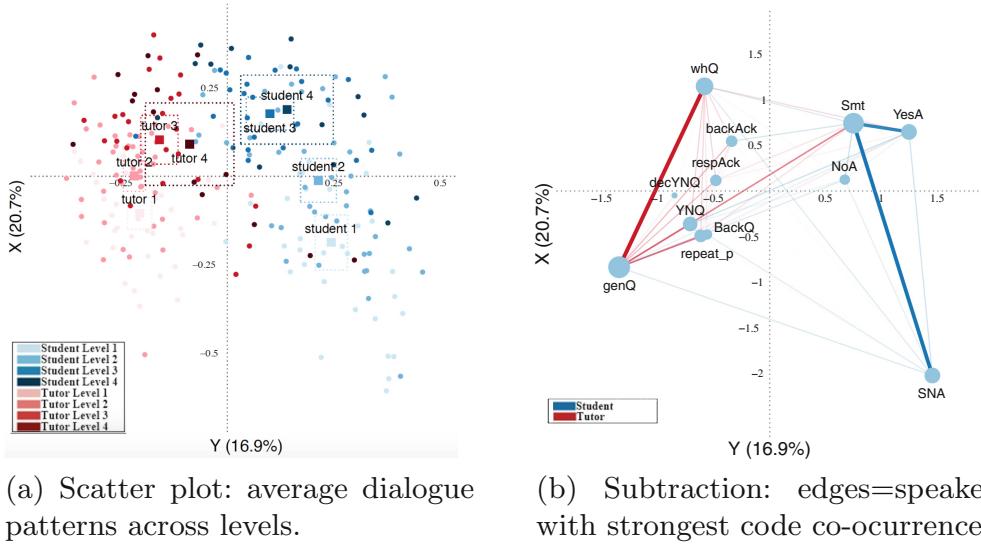
<sup>1</sup> The item being aligned to in this context is known as the *prime*, and the subsequent usage of this prime by the other speaker is known as the *target*, or sign of alignment.

**Table 1.** DA dialogue examples at Levels 1 (Highest) and 4 (Lowest) in BELC **P** = Participant **DA** = Dialogue Act **SPA** = Code-switching in Spanish

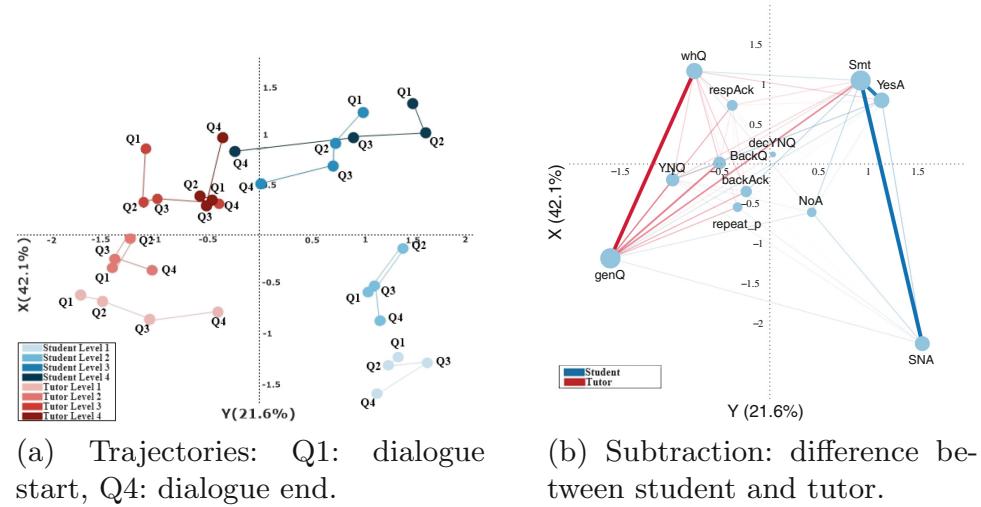
P	Level 1	DA	P	Level 4	DA	DA key
T	do you like the school?	YNQ	T	do you like this school?	YNQ	<b>YNQ:</b> <i>yes-no-question</i>
T	[- spa] m-entens?	SPA	S	yes	YesA	<b>RespAck:</b> <i>response</i>
S	0 [= says nothing]	SNA	T	yes?	RAck	<i>acknowledgement</i>
T	“do you like”?	YNQ	T	what are you planning to do next year?	WhQ	<b>decYNQ:</b> <i>declarative YNQ</i>
T	do you like the school?	YNQ	S	I would like to study zoology	Smt	<b>BackQ:</b> <i>backchannel-Q</i>
S	xxx	SNA	T	what time did you arrive here this morning?	WhQ	<b>whQ:</b> <i>wh-question</i>
T	no si t-agrada l-escola?	SPA	S	this morning?	GenQ	<b>GenQ:</b> <i>General-Other-Q</i>
T	do you like the school?	YNQ	T	yes	YesA	<b>YesA:</b> <i>yes answers</i>
S	yes	YesA	S	I ... I am here since eight o'clock	Smt	<b>NoA:</b> <i>no answers</i>
T	yes ok	RAck	T	uhhuh right quite early	Smt	<b>NA:</b> <i>non-understanding</i>
T	now what time do you begin in the morning?	WhQ	T	and when will you leave?	WhQ	<b>Smt:</b> <i>statement</i>
		SNA	S	I ... I finish my time-table in half-past-two	Smt	<b>repeat:</b> <i>repeat-phrase</i>
S	0 [= says nothing]	SPA				<b>backAck:</b> <i>backchannel-acknowledge</i>
T	[- spa] m-entens?	SPA				

Interlocutor means are closest at higher student ability levels. We see evidence of tutor movement within DA space (t-tests reveal significant difference between Tutor Level 1 (T1) and Tutor Level 4 (T4): ( $D = 1.26 p < 0.001$ ), which we interpret as tutors' adapting their strategy to learner ability. Students show *more* movement across ability level than tutors (t-tests reveal significant difference between S1 and S4: ( $D = 1.79 p < 0.001$ )), indicating that ability influences the sorts of DAs produced, with a more active role (*Wh-Questions (whQ)*, *Response-Acknowledgements (RespAck)* and *Statements (Smt)*) being taken by higher level students. Figure 1(b) shows students have more connections between *statements*, *signal-non-understanding* and *yes-answers* than tutors, who have more connections in general, specifically between *questions*, *back-channeling* and *repetition*.

To answer RQ2, stanzas are *dialogue quartiles* and unit of analysis *speaker* and *ability level*. Figure 2(a) shows trajectories over the four quartiles, points represent mean speaker position in the same DA space as Fig. 2(b). Figure 2(a)



**Fig. 1.** ENA projection graph: Students and tutors at Higher levels are closer to DAs such as *WH-questions*, and *Statements*; whereas at lower levels, DAs such as *general questions*, and *Signal-non-understanding* (*SNA*).



**Fig. 2.** ENA trajectory. t-tests show sig. diff. between Q1 & Q4 for each trajectory except L1 Students ( $x(D = 0.28 p = 0.26)$ ,  $y(D = 0.08 p = 0.74)$ ). Highest effect sizes: Tutor L1 ( $D = 1.49 p = 0.001$ ) and Student L4 ( $D = 1.46 p = 0.001$ ).

shows greater speaker DA similarity at Q4 than at Q1 for all levels, supporting our hypothesis of DA convergence over an interaction. This is most pronounced at high ability levels. High level students show most convergence (greater distance between Q1 and Q4). We can interpret this as indication that their ability allows them to align more, or that DA usage becomes more diverse with ability. Tutors show less movement, except to converge with L1 students. We interpret this as evidence of tutor strategy: converging when the student cannot, and adapting less when they are capable. Higher ability dialogues have been shown to become more symmetric [18], mirroring native speakers. Here, we are able to

see that this is the case for interlocutors' use of DAs. While evidence of alignment at a lexical level has been found in BELC [17], our work shows this at a more abstract level in terms of the conversational dynamics via DAs.

## 4 Contributions and Conclusions

We contribute a novel method for analysis of L2 dialogue, combining DA labels with ENA. Our findings support the hypothesis that L2 speakers in dialogue practice exhibit a degree of convergence, both as ability level increases and over the course of a dialogue. This better understanding of tutor adaptation can inform the design of tutoring dialogue systems. This method can be used by practitioners in learning analytics for the design of new tools across different dialogue modalities. The corpus used is not large or diverse enough for us to make generalisations about particular dialogue characteristics at certain levels thus we limit our interpretation to high-level adaptation phenomena. Next we plan to explore particular DA functions and difficulty in context. The shift of speaker DA position suggests different DA patterns are used to better suit student ability. We hypothesise certain DA sequences may be more indicative of learner support, and others of conversational symmetry.

## References

1. Bailey, K.M.: What my EFL students taught me. *PAC J.* **1**(1), 7–31 (2001)
2. Birjandi, P., Jazebi, S.: A comparative analysis of teachers' scaffolding practices (2014)
3. Branigan, H.P., Pickering, M.J., McLean, J.F., Cleland, A.A.: Syntactic alignment and participant role in dialogue. *Cognition* **104**(2), 163–197 (2007)
4. Costa, A., Pickering, M.J., Sorace, A.: Alignment in second language dialogue. *Lang. Cogn. Process.* **23**(4), 528–556 (2008)
5. Duplessis, G.D., Clavel, C., Landragin, F.: Automatic measures to characterise verbal alignment in human-agent interaction. In: 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 71–81 (2017)
6. Garrod, S., Pickering, M.J.: Alignment in dialogue. In: *The Oxford Handbook of Psycholinguistics*, pp. 443–451 (2007)
7. Hawkes, R.: Learning to talk and talking to learn: how spontaneous teacher-learner interaction in the secondary Foreign languages classroom provides greater opportunities for L2 learning. Ph.D. thesis, University of Cambridge (2012)
8. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard dialog act corpus. International Computer Science Institute, Berkeley, CA, Technical report (1997)
9. Lantolf, J.P.: Second language learning as a mediated process. *Lang. Teach.* **33**(2), 79–96 (2000)
10. Mandel, J.: Use of the singular value decomposition in regression analysis. *Am. Stat.* **36**(1), 15–24 (1982)
11. Muñoz, C.: Age and the rate of Foreign language learning. *Multilingual Matters*, vol. 19 (2006)
12. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* **27**(2), 169–190 (2004)

13. Reitter, D., Keller, F., Moore, J.D.: Computational modelling of structural priming in dialogue. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, NAACL-Short 2006, pp. 121–124. Association for Computational Linguistics, Stroudsburg (2006). <http://dl.acm.org/citation.cfm?id=1614049.1614080>
14. Reitter, D., Moore, J.D.: Alignment and task success in spoken dialogue. *J. Mem. Lang.* **76**, 29–46 (2014)
15. Samana, W.: Teacher's and students' scaffolding in an EFL classroom. *Acad. J. Interdiscip. Stud.* **2**(8), 338 (2013)
16. Shaffer, D.W., et al.: Epistemic network analysis: a prototype for 21st-century assessment of learning. *Int. J. Learn. Media* **1**(2), 33–53 (2009). <https://doi.org/10.1162/ijlm.2009.0013>
17. Sinclair, A., Lopez, A., Lucas, C., Gasevic, D.: Does ability affect alignment in second language tutorial dialogue? In: Proceedings of the 19th Annual SIGDIAL Meeting on Discourse and Dialogue, pp. 41–50 (2018)
18. Sinclair, A., Oberlander, J., Gasevic, D.: Finding the zone of proximal development: student-tutor second language dialogue interactions. In: Proceedings of SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue, pp. 107–115 (2017)
19. Stolcke, A., et al.: Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **26**(3), 339–373 (2000)
20. Ward, A., Litman, D.: Dialog convergence and learning. *Front. Artif. Intell. Appl.* **158**, 262 (2007)
21. Ward, A., Litman, D.: Measuring convergence and priming in tutorial dialog. University of Pittsburgh (2007)



# Understanding Students' Model Building Strategies Through Discourse Analysis

Caitlin Snyder<sup>1</sup>(✉), Nicole Hutchins<sup>1</sup>, Gautam Biswas<sup>1</sup>,  
and Shuchi Grover<sup>2</sup>

<sup>1</sup> Vanderbilt University, Nashville, TN 37235, USA

caitlin.r.snyder@vanderbilt.edu

<sup>2</sup> Looking Glass Ventures, Palo Alto, USA

**Abstract.** The benefits of computational model building in STEM domains are well documented yet the synergistic learning processes that lead to the effective learning gains are not fully understood. In this paper, we analyze the discussions between students working collaboratively to build computational models to solve physics problems. From this collaborative discourse, we identify strategies that impact their model building and learning processes.

**Keywords:** Computational modeling · Collaborative discourse

## 1 Introduction

Technology-enhanced environments can be productive vehicles for engaging students in computational model building and problem solving, a process shown to be effective for learning K-12 science concepts (e.g. [2, 6, 18]). This mutually supportive approach to STEM and CT learning has produced synergistic learning environments [3, 9, 13, 17] where students express domain concepts and laws in a computational form, and then interpret the behaviors generated by these computational constructs to refine their knowledge of the domain. The necessity to combine, represent, interpret, and analyze the two simultaneously in a mutually supportive way is what we call *synergistic learning*. While we have theorized the advantages of synergistic learning [13], and assessments have demonstrated students' learning gains attributed to these environments [1, 2], how they develop and apply these synergistic learning processes to their learning and modeling tasks are not fully understood.

For this research, students learn by building, simulating, testing, and refining their models in C2STEM [9]. We analyze collaborative discourse as students work in small groups to develop a shared understanding of a phenomena by jointly constructing models [11]. While working on their model building tasks on a shared screen, students have the opportunity to discuss, explain, argue about and evaluate their models [14]. In this paper we use students' collaborative problem-solving dialogues along with information on how they progress in their model building to identify students' STEM and CT learning processes, while also gaining some insight into their group dynamics. Specifically, we perform an exploratory analysis to identify dialogue characteristics and

model building moves that may be indicative of strategies they use in their computational modeling tasks.

## 2 Background

Computational modeling of scientific processes provides an effective framework for learning scientific concepts and practices through computational representations and simulation models, as well as CT practices like those evaluated in [17]. Reciprocally, the concepts and practices emphasized in CT are better contextualized and, therefore, easier to understand and learn when they are situated in domain specific model building, analysis, and problem-solving tasks [4, 13]. Such environments that facilitate synergistic learning have proven to be effective in increasing learning gains in the STEM and CT domains [1, 2, 9]. Our work extends these approaches using a block-based computational modeling environment, C2STEM, equipped with tools aimed at scaffolding the learning of STEM and CT. These tools include a domain-specific modeling language (DSML) with physics constructs to help students create dynamic (simulation) models in Physics and control-structure blocks to initialize needed variables (Green Flag) and to program the dynamic behavior changes of each object (the Simulation Step block), aimed at evaluating the step-by-step update of the model via animations and data tools.

We analyze collaborative student dialogue with a learning and social framework to better understand successful and unsuccessful learning processes building on related work [8, 10]. Dialogue is characterized by the domain (Physics or CT) of focus during knowledge construction. Discussion are further characterized by a combination of the ICAP framework [5] and the framework proposed by Weinberger and Fischer [16]. The ICAP framework designates four different modes of learning: Interactive, Constructive, Active and Passive. The Passive mode is characterized by a learner receiving information without visible response, whereas an Active learner responds by manipulating the learned knowledge. Constructive learners add one more step by manipulating the information to construct something new. Interactive learners discuss and construct knowledge with a fellow learner. We incorporate the five different social modes in argumentative knowledge construction from Weinberger & Fischer's framework with the ICAP learner modes to interpret the types of dialogues. The social modes are defined as conflict-oriented consensus building, integration-oriented consensus building, quick consensus building, elicitation, and externalization. The three consensus building modes occur when there is a discussion between learners. Elicitation can lead to a consensus building or a learner may answer their own question. Externalization is a primarily singular mode where one learner is vocalizing what they are doing while the other learner(s) in their group are quiet. We combine these two frameworks by mapping the learning modes to the social modes [15].

### 3 Methods

We conducted a study with 26 high school sophomore students using C2STEM. The students spent one day a week for 2 months completing a CT training module, 3 kinematics modules: 1D and 2D motion including gravity, and 1 force module. Our curriculum included three types of tasks: instructional, model building, and challenge [9]. We divided the students into 9 different groups, 8 groups had three students per group, and the ninth was a group of 2 students. Each group was instructed to work together on one computer screen to build their models. There was discussion across groups. These were not discouraged and are reported as part of our analysis.

Our data sources are OBSTM screen-capture videos that recorded the students' screens along with the webcam video and audio and model scores on submitted tasks. We focused our qualitative analysis on the 2D motion with constant velocity challenge task. In this module, students modeled a boat crossing a flowing river while stopping at two different islands along the way. Model scores were computed utilizing a pre-defined rubric divided into use of physics and CT constructs in order to evaluate proficiency in each domain separately.

### 4 Results

Using the collaborative dialogue framework described above for qualitative analysis, we identified 5 predominate problem-solving strategies. Table 1 provides transcript evidence to support our identification of problem-solving approaches. We saw increased performances by groups 2, 4 and 5 over time. Interestingly, Group 2 seemed

**Table 1.** Dialogue examples of strategies

Strategy	G	Example quotes
Hardcoding	1, 2	S1: "it goes 5 m/s, but to go 6 meters forward it would be 1.2 s. So we need to figure out, we know the distance we know the time we know the change in distance over the change in time now that will give us the velocity. So 15/1.2 [calculates it on paper]. 12.5."
Data tools	4, 5	S10: "So we find x y coordinates and find the slope and then go there and there [referencing the islands] S11: "So where is this. Wait how do we look at the variables" S9: "Display x and y position"
Debugging	5	S12: "just for testing purposes let's make this an if else and put stop simulation in the else so once it gets there it should stop moving"
Trial and error	6, 8, 9	S22: "just change the velocity to be lower" S23: "okay we will change that" S22: "just trial and error, make it -4"
Replication/Help	7	Other group: "Here's the thing, your x velocity should be 5" S19: "No I think that since the river is 2 you need to add more to it"

to have the strongest CT skills from the onset of the curriculum. This group showed gains in Physics (75% to 87.5%) and CT (75% to 90%) performance. Group 5 started with a high performance in Physics (100%) and maintained that with a 100% on the 2D motion challenge (there was a slight dip in score when they started with 2D motion). We hypothesize this indicates some prior knowledge in Physics. Group 5 did show increases in CT over time (from a 62.5% to 90%). Group 7's Physics and CT performances dropped (from 75% to 25 and 62.5% to 20%, respectively). Groups 6 and 9 scored lower in CT but maintained their performance in Physics. We conjecture this correlates with the common difficulty of translating Physics knowledge to a computational model [13].

## 5 Discussion

The only group to receive lower scores in Physics and CT on this task compared to previous tasks utilized a replication problem solving approach, aiming to copy a solution of another group. This was the only group to engage in *constructive externalization*, instead of working together or communicating as a team. This group began with a successful strategy (using the data tools) but were unable to interpret the physics calculations. We conjecture that confidence in knowledge application or abilities to translate gained Physics knowledge to a more challenging computational model may have caused this decline as the group elected to seek help elsewhere.

The two groups that showed constant Physics scores but decreasing CT scores utilized trial and error or replicating code strategies. These strategies avoid switching between physics and CT. In fact, the groups who did trial and error primarily focused on the computational model and did not attempt to utilize physics concepts like the kinematics equations to solve the problem. Alternately, Group 8 utilized a combination of trial and error and the data tools. The combination of one unsuccessful strategy, trial and error, with a successful strategy, data tool use, seems to have resulted in neither a loss nor a gain in knowledge construction in both Physics and CT.

Finally, the groups whose model scores remained constant or increased in physics and CT, based on model scores utilized hardcoding, data tools, and debugging strategies. These strategies show switching of focus between physics and CT understanding. The hard coding of values into the computational model requires some physics knowledge. Utilizing the data tools, students identified initial positions values for use in their physics equations. Debugging strategies required students to interpret their model behaviors using physics constructs and to identify errors in their models. All of these strategies can be considered synergistic learning processes.

## 6 Conclusion

A systematic approach that combines quantitative and qualitative analysis of collaborative, computational model building strategies provides useful information into how students problem solve to learn Physics and CT simultaneously. Through careful evaluation of instances in which both Physics and CT knowledge are needed to build a

computational model, successful strategies can be characterized by the use of synergistic processes. For example, when a group implements a combination of strategies such as debugging (a CT strategy [7]) and data evaluation (a Physics strategy [12] and a CT process [17]), this results in increased scores in both domains. Unsuccessful strategies do not exploit synergy between the domains, and lead to drop in performance. Combination of good and bad strategies produce mixed results.

**Acknowledgements.** This research is supported by NSF grant #1640199.

## References

1. Basu, S., Biswas, G., Kinnebrew, J.S.: Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Model. User-Adap. Interact.* **27**(1), 5–53 (2017)
2. Basu, S., Biswas, G., Kinnebrew, J.S.: Using multiple representations to simultaneously learn computational thinking and middle school science. In: Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, pp. 3705–3711 (2016)
3. Basu, S., Dickes, A., Kinnebrew, J.S., Sengupta, P., Biswas, G.: CTSiM: a computational thinking environment for learning science through simulation and modeling. In: CSEDU, pp. 369–378 (2013)
4. Brennan, K., Resnick, M.: New frameworks for studying and assessing the development of computational thinking. In: presented at the American Education Researcher Association, Vancouver, Canada (2012)
5. Chi, M.T., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **49**(4), 219–243 (2014)
6. DiSessa, A.A.: *Changing Minds: Computers, Learning, and Literacy*. Mit Press, Cambridge (2001)
7. Grover, S., Pea, R.: Computational thinking: a competency whose time has come. In: Sentance, S., Carsten, S., Barendsen, E. (eds.) *Computer Science Education: Perspectives on Teaching and Learning*, pp. 19–38. Bloomsbury, New York (2018)
8. Grover, S., Hutchins, N., Biswas, G., Snyder, C., Emara, M.: Examining synergistic learning of physics and computational thinking through collaborative problem solving in computational modeling. In: AERA, Toronto, CA (2019)
9. Hutchins, N., Biswas, G., Maroti, M., Ledezci, A., Broll, B.: A design-based approach to a classroom-centered OELE. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 155–159. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_28](https://doi.org/10.1007/978-3-319-93846-2_28)
10. Hutchins, N.: Studying synergistic learning of physics and computational thinking in a learning by modeling environment. In: 26th International Conference on Computers in Education (ICCE), Manila, Philippines (2018)
11. Larkin, S.: Collaborative group work and individual development of metacognition in the early years. *Res. Sci. Educ.* **36**(1–2), 7–27 (2006)
12. NGSS Lead States: *Next Generation Science Standards: For States, By States*. The National Academies Press, Washington, DC (2013)
13. Sengupta, P., Kinnebrew, J.S., Basu, S., Biswas, G., Clark, D.: Integrating computational thinking with K-12 science education using agent-based computation: a theoretical framework. *Educ. Inf. Technol.* **18**(2), 351–380 (2013)

14. Sins, P.H., Savelsbergh, E.R., van Joolingen, W.R.: The difficult process of scientific modelling: an analysis of novices' reasoning during computer-based modelling. *Int. J. Sci. Educ.* **27**(14), 1695–1721 (2005)
15. Snyder, C., Hutchins, M., Biswas, G., Emara, M., Grover, S., Conlin, L.: Analyzing students' synergistic learning processes in physics and ct by collaborative discourse analysis. In: To be presented at CSCL, Lyon, France (2019)
16. Weinberger, A., Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Comput. Educ.* **46**(1), 71–95 (2006)
17. Weintrop, D., et al.: Defining computational thinking for mathematics and science classrooms. *J. Sci. Educ. Technol.* **25**(1), 127–147 (2016)
18. Wilensky, U., Reisman, K.: Thinking like a wolf, a sheep, or a firefly: learning biology through constructing and testing computational theories—an embodied modeling approach. *Cognit. Instr.* **24**(2), 171–209 (2006)



# Exploring Teachable Humans and Teachable Agents: Human Strategies Versus Agent Policies and the Basis of Expertise

John Stamper<sup>(✉)</sup> and Steven Moore

Carnegie Mellon University, Pittsburgh, PA 15213, USA  
jstamper@cs.cmu.edu, stevenjamesmoore@gmail.com

**Abstract.** In this research, we explore how expertise is shown in both humans and AI agents. Human experts follow sets of strategies to complete domain specific tasks while AI agents follow a policy. We compare machine generated policies to human strategies in two game domains, using these examples we show how human strategies can be seen in agents. We believe this work can help lead to a better understanding of human strategies and expertise, while also leading to improved human-centered machine learning approaches. Finally, we hypothesize how a continuous improvement system of humans teaching agents who then teach humans could be created in future intelligent tutoring systems.

**Keywords:** Policies · Agents · Strategies · Expertise

## 1 Introduction

In this research, we explore how AI agent policies might be used to teach humans. In complex tasks humans generate strategies which can be applied in many different situations. Combinations of strategies that lead to optimal outcomes can lead to expertise in a domain, although there is still no consensus among researchers as to what makes a person an expert and how expertise is defined. We explore the interactions of policies and strategies, looking at how both relate to expertise. Our long term goal is to see how humans can help teach agents and agents can help teach humans in a continuous loop. A start to this goal is a comparison of agent policies, generated with different techniques on several complex game domains, with strategies generated from human players.

## 2 Background and Domains

Expertise has been the subject at the crossroads of Psychology and Computer Science for some time. *The Nature of Expertise* [14] explored a wide variety of domains from human typing to sports to ill-defined domains. A key insight from this work is that in the early development of AI systems, expertise was tightly related to the concept of encoding human strategies into machines, such as early work involving chess players and intelligent tutors [4]. As work continued, the Psychology field moved into

architectures of cognition defined by ACT-R [1] and Soar [18] as examples. Computer Science moved towards agents and policy creation focusing early on reinforcement learning [29] and now advanced techniques built on deep learning [19].

The question of what exactly defines someone as an expert is still an open question and has a lot to do with the particular domain that is being studied. In chess, Chase and Simon posited that it takes 10,000 h of study to become an expert [4]. That number has also been suggested as the rough number of hours to become an expert musician [11] and is a general theory of expertise [10], although largely due to Simon's chess work. In the case of learning systems, we often define mastery using some form of knowledge tracing. These systems often set “mastery” as a probabilistic value that a learner knows a particular skill. The value of mastery varies on skills and domains, but a value of 90% or 95% are assumed to have achieved mastery [8]. Understanding skills that are used to solve problems has also been explored in many domains [16, 25]. Tasks to elicit knowledge from experts, such as cognitive task analysis (CTA) have been used by cognitive scientists to better understand the strategies that experts use, but may not explicitly recognize [6].

AI has been used now for decades to create agents that mimic human behavior. These agents are generally driven by a policy created by some form of machine learning, such as Q-learning [29]. The policy tells the AI agent what to do given a certain set of conditions. This is most often defined as a state-action graph that suggests the best possible next action for an agent assigned to a given state. In education, agents driven by policies have long been a foundational part of intelligent tutors and adaptive learning. Work has been done in modeling learning as a policy generated to predict what a student knows and what the next best instructional lesson is for a particular student [24]. Other research has been done using reinforcement learning (RL) with a focus on what pedagogical action would be best to use for a student when multiple actions are available [5]. Most closely associated with the research we are doing is work on the automatic generation of hints and feedback [23, 27]. This work uses state graphs and RL to identify the best path for solving problems. Then generates a just in time hint or provides feedback that can lead the student down a better path for learning.

We focus on two complex game domains: **connect four** (C4) and **Space Invaders** (SI). Both are well known games and chosen because of their simplicity of play and known human strategies for winning. They also have multiple agent implementations that we can exploit, which are explained in detail below.

The objective of C4 is to align four game pieces of the same color in a row, either diagonally, horizontally, or vertically. There are three possible states for each of the forty-two available game spaces. The board spaces can be occupied by the turn players piece, the opponent's piece, or it can be empty. This means there are  $3^{42}$  ( $\geq 10^{20}$ ) moves possible on the game board of seven columns by six rows, ranging from zero to forty-two pieces on it. Using binary decision diagrams, it has been shown there are exactly 4,531,985,219,092 legal board configurations [9]. Additionally, C4 is a solved zero-sum game, of moderate complexity, where the outcome of the game can correctly be predicted from any state [31]. There are many variants of C4 agents [12, 13]. Recent agents that solve the game using temporal difference learning, achieved a win percentage close to perfect, but require several millions of self-play games for training, thus being far off human performance [2]. Another study found that using 1,565,000

games for training data, their agent could reach an 80% success rate, but it required between 2–4 million games to produce what would be considered a strong-playing one [30]. The most successful agents of C4 make use of the MiniMax algorithm, which consists of heuristic evaluation function that is akin to these human strategies. It is often cited as the standard to compare different agent implementations against, as MiniMax can win virtually every time, depending on its search depth, with no training data required [30]. This is powerful since C4 is a zero sum game, and the heuristic function has the agent follow a set of optimal human-like strategies. The evaluation function can be summarized by five strategy points: (1) If there is a winning move, take it (2) If the opponent has a winning move, prevent it (3) Take the center square over edges and corners (4) Take corner squares over edges (5) Take edges if nothing else is available.

Just implementing a simple human strategy can have a profound effect on the size of the agents search space and number of game plays needed to generate an expert agent. For example, one basic strategy is when given the opportunity to go first, a player should always take the center position on the board, and if going second the player should take this position if available. From a simple computation we can see that this prunes 6 of the 7 high level branches in the initial graph leading to tremendously less possible game states in the expert player.

Space Invaders was a classic arcade game and one of the games available in the Atari Grand Challenge dataset (AGC) [17] based on the classic Atari 2600 home console game system. In dataset-1 of the AGC, there are 445 human game plays of SI. SI also represents a potentially easier game to follow in the Atari game space because the game dynamics remove some of the available moves. While Atari games allow for the use of four directional movements (left, right, up, down) plus a button, SI only allows the player to move left or right and use the button to shoot. This limits the complexity of this game compared to some others.

There are a number of human strategies that we have discovered from discussions with an expert of the game. This expert was able to achieve scores greater than 98% of all human players as reported by the Atari Grand Challenge site. The human strategies include (1) because only one shot can be on the screen at a time shooting lower invaders leads to faster shooting, (2) shooting entire columns from the left side first give additional time because of the right to left movement of the invaders, and (3) when the invaders reach the left side and begin moving right shoot the bottom row and move to shooting the rightmost column. These strategies keep the invaders largely in a square formation. It is disadvantageous to split the invaders into two squares, because that requires additional movement to get a shot off.

Using the Arcade Learning Environment (ALE) [3], agents have been created and trained to play Space Invaders. A summary of the scores of three agent based on different algorithms in a replication study of a number of previously built agents in the ALE framework [21] claimed agents did exceed human capabilities at times, although they did not average a score that was higher than the top 5% of human players presented in the AGC dataset. When we looked at data from the DQN agent, which was driven by a deep learning algorithm, visualizing the RAM states based on a t-SNE embedding [22] shows that many of the clusters do show evidence of human strategies, such as keeping the invaders in a single square formation. Watching replays of expert

agent players also shows expert human strategies, but more work is needed to delve into the actual policy to find clear evidence of a particular strategy.

### 3 Discussion and Conclusions

Heuristic driven policies, by means of a given evaluation function, are widely used to solve games such as chess, C4, Othello and Go [7]. The evaluation functions in these agents use information about the game. Much of this is directly related to a strategy that a human player would follow, as addressed previously with C4. These strategies represent expertise in a human player and are clearly identifiable in agent play. In the development of agents, it is the human encoding the strategy into the AI using their knowledge of the game. The majority of game-playing agents, however, make use of deep neural nets to develop their policies, which makes them black-box and often difficult to interpret by a human. Recent work has looked at making policies developed this way programmatically interpretable, but much work remains for humans to be able to clearly articulate what many of these agents have learned from their training [32].

It is debatable if these deep reinforcement learning agents make use of explicit strategies as they execute their given policies. A recent approach uses saliency maps to highlight key decision regions for agents playing Atari 2600 games, and found that their SI agent learned a sophisticated aiming strategy [15]. Another way to make policies less black-box, is to break the policy down into smaller subtasks that are comprised of a few actions that feed back into the overall policy [20]. These techniques of breaking down policies into smaller interpretable strategies and visually representing the mechanisms of an agent's policy are steps toward having humans learn strategies from agents, without directly encoding any into the agent itself.

Some previous work looks to use human seeding of policies in educational domains [26]. Another such study found that training on human data; they could achieve comparable scores to state-of-the-art reinforcement learning techniques and even beat the scores using just the top 50% of their collected data for more complicated games [17]. Combining a method that not only trains agents on expert human data, but also encodes their strategies into a form of an evaluation function has the potential to yield successful agents that require less computational time, while performing at greater levels than comparable agents.

We can identify human strategies in the policies generated by agent through post hoc human inspection. In the future, we will explore how to automate the process of identifying strategies within the agent policies similar to previous work on less complex educational domains [28]. This will require progress on explainable AI to extract human readable information from increasingly black-box policies. We plan to explore a number of additional domains where data and agents are available for study.

## References

1. Anderson, J.R., Matessa, M., Lebiere, C.: ACT-R: a theory of higher level cognition and its relation to visual attention. *Hum.-Comput. Interact.* **12**(4), 439–462 (1997)
2. Bagheri, S., Thill, M., Koch, P., Konen, W.: Online adaptable learning rates for the game Connect-4. *IEEE Trans. Comput. Intell. AI Games* **8**(1), 33–42 (2016)
3. Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M.: The arcade learning environment: an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253–279 (2013)
4. Chase, W.G., Simon, H.A.: Perception in chess. *Cognit. Psychol.* **4**(1), 55–81 (1973)
5. Chi, M., VanLehn, K., Litman, D., Jordan, P.: An evaluation of pedagogical tutorial tactics for a natural language tutoring system: a reinforcement learning approach. *Int. J. Artif. Intell. Educ.* **21**(1–2), 83–113 (2011)
6. Clark, R.E., Estes, F.: Cognitive task analysis for training. *Int. J. Educ. Res.* **25**(5), 403–417 (1996)
7. Clune, J.: Heuristic evaluation functions for general game playing. In: AAAI, vol. 7, pp. 1134–1139, July 2007
8. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Interact.* **4**(4), 253–278 (1994)
9. Edelkamp, S., Kissmann, P.: Symbolic classification of general two-player games. In: Dengel, A.R., Berns, K., Breuel, T.M., Bomarius, F., Roth-Berghofer, T.R. (eds.) KI 2008. LNCS (LNAI), vol. 5243, pp. 185–192. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-85845-4\\_23](https://doi.org/10.1007/978-3-540-85845-4_23)
10. Ericsson, K.A., Smith, J. (eds.): *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge (1991)
11. Ericsson, K.A., Prietula, M.J., Cokely, E.T.: The making of an expert. *Harvard Bus. Rev.* **85** (7/8), 114 (2007)
12. Faußer, S., Schwenker, F.: Neural approximation of monte carlo policy evaluation deployed in connect four. In: Prevost, L., Marinai, S., Schwenker, F. (eds.) ANNPR 2008. LNCS (LNAI), vol. 5064, pp. 90–100. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69939-2\\_9](https://doi.org/10.1007/978-3-540-69939-2_9)
13. Ghory, I.: Reinforcement learning in board games. Department of Computer Science, University of Bristol, Technical report 105 (2004)
14. Glaser, R., Chi, M.T., Farr, M.J. (eds.): *The Nature of Expertise*. Lawrence Erlbaum Associates, Hillsdale (1988)
15. Greydanus, S., Koul, A., Dodge, J., Fern, A.: Visualizing and understanding atari agents. arXiv preprint [arXiv:1711.00138](https://arxiv.org/abs/1711.00138) (2017)
16. Koedinger, K.R., Stamper, J.C., Leber, B., Skogsholm, A.: LearnLab’s datashop: a data repository and analytics tool set for cognitive science. *Top. Cognit. Sci.* **3**(5), 668–669 (2013)
17. Kurin, V., Nowozin, S., Hofmann, K., Beyer, L., Leibe, B.: The atari grand challenge dataset. arXiv preprint [arXiv:1705.10998](https://arxiv.org/abs/1705.10998) (2017)
18. Laird, J.E., Newell, A., Rosenbloom, P.S.: SOAR: an architecture for general intelligence. *Artif. Intell.* **33**(1), 1–64 (1987)
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
20. Lyu, D., Yang, F., Liu, B., Gustafson, S.: SDRL: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. arXiv preprint [arXiv:1811.00090](https://arxiv.org/abs/1811.00090) (2018)

21. Machado, M.C., Bellemare, M.G., Talvitie, E., Veness, J., Hausknecht, M., Bowling, M.: Revisiting the arcade learning environment: evaluation protocols and open problems for general agents. arXiv preprint [arXiv:1709.06009](https://arxiv.org/abs/1709.06009) (2017)
22. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518** (7540), 529 (2015)
23. Moore, S., Stamper, J.: Decision support for an adversarial game environment using automatic hint generation. In: International Conference on Intelligent Tutoring Systems (ITS 2019). Springer, Heidelberg (2019, to Appear)
24. Rafferty, A.N., Brunskill, E., Griffiths, T.L., Shafto, P.: Faster teaching by POMDP planning. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 280–287. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21869-9\\_37](https://doi.org/10.1007/978-3-642-21869-9_37)
25. Stamper, J., et al.: PSLC DataShop: a data analysis service for the learning science community. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6095, p. 455. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13437-1\\_112](https://doi.org/10.1007/978-3-642-13437-1_112)
26. Stamper, J., Barnes, T., Croy, M.: Enhancing the automatic generation of hints with expert seeding. *Int. J. AI Educ.* **21**(1–2), 153–167 (2011)
27. Stamper, J., Barnes, T.: Unsupervised MDP value selection for automating ITS capabilities. In: Educational Data Mining 2009, pp. 180–188 (2009)
28. Stamper, J.C., Barnes, T., Croy, M.: Extracting student models for intelligent tutoring systems. In: Proceedings of the National Conference on Artificial Intelligence, vol. 22, no. 2, p. 1900. AAAI Press, Menlo Park, CA. MIT Press, Cambridge (1999, 2007)
29. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
30. Thill, M.: Temporal difference learning methods with automatic step-size adaptation for strategic board games: Connect-4 and Dots-and-Boxes. Doctoral dissertation, Master thesis, Cologne University of Applied Sciences, June 2015
31. Tromp, J.: Solving Connect-4 on medium board sizes. *ICGA J.* **31**(2), 110–112 (2008)
32. Verma, A., Murali, V., Singh, R., Kohli, P., Chaudhuri, S.: Programmatically Interpretable Reinforcement Learning. arXiv preprint [arXiv:1804.02477](https://arxiv.org/abs/1804.02477) (2018)



# Learning from Videos Showing a Dialog Fosters More Positive Affect Than Learning from a Monolog

Samantha Stranc and Kasia Muldner<sup>(✉)</sup>

Carleton University, Ottawa, Canada  
{samantha.stranc, kasia.muldnner}@carleton.ca

**Abstract.** How much students learn from instructional videos is influenced by the type of video. Prior work has shown that when students are given a video showing a *dialog* between a tutor and a tutee, they learn more than if the video shows a *monolog* delivered by a tutor. To date, however, there does not exist work investigating how each type of video impacts student affect. To fill this gap, we apply sentiment analysis to transcripts of students learning in each context. We show that learning from videos with dialog fosters more positive affect for university-level students, but not for middle-school students.

**Keywords:** Sentiment analysis · Instructional videos · Dialog · Monolog

## 1 Introduction

In addition to contexts like classrooms and tutoring systems, another common way to learn is from instructional videos. Most videos include an auditory description of the topic being presented (e.g., the voice of a teacher going over a math problem), but some also show the instructor, who presents the materials through a monolog-style exposition. While showing an instructor in the video does not always improve learning [6, 11, 18], students find the presence of an instructor motivating, be that individual a human [11, 18] or a pedagogical agent [6]. Thus, there are benefits of including the instructor in instructional videos. An alternative approach to showing only the instructor is to include an instructor and a student going over the materials together, like they would in a one-on-one tutoring session or office hour, i.e., to show a *dialog* between two individuals rather than a *monolog* presented by one individual. Although the dialog approach is less common, there are indications that students learn better from videos that include a dialog over ones with only a monolog [3, 4, 12, 14]. There are a number of proposals for why this is the case [3, 10, 13]. One is that refutation of misconceptions helps learning and a dialog includes misconceptions expressed by the tutee and refuted by the tutor. Another is that overhearing questions about the material helps students learn, and again, a dialog includes questions posed by the tutee to the tutor and/or by the tutor to the tutee.

Thus, as highlighted above, there are cognitive benefits to learning from dialog, and thus some tutoring systems have incorporated this style of instruction [5]. However, it is an open question as to how the presentation format of instructional videos (monolog, dialog) impacts student affect. Addressing this question is important because affect

impacts student behaviors and outcomes [1, 8, 16, 17]. For instance, boredom is associated with gaming and poor learning [2], while confusion and uncertainty can be beneficial but only if students are aware of being in those states [9].

To fill this gap, in the present work we rely on a Natural Language Processing (NLP) tool, namely SEANCE [7], to automatically extract sentiment information from student discussions recorded as they watched an instructional video (dialog or monolog). Briefly, sentiment information pertains to human emotions, attitudes, and polarity. To date, NLP-based sentiment tools have been used to measure confusion in student discussions in MOOCs [20], markers of expertise in student learning [19], and sentiment in student evaluations of teaching [15], to name a few examples. To date, however, these tools have not yet been applied to examine potential differences in sentiment when students are given a dialog- vs. monolog-style instructional video.

The data for our analysis comes from a previous study [12] but to date its sentiment content has not been analyzed. In that study, two students worked together on a worksheet asking them questions about the process of diffusion. Students were told to talk and come to agreement with their partner before writing solutions down, as well as to take turns writing, and all conversations were recorded and transcribed; these transcripts provide the data for the current analysis. Each pair was given an instructional video in which a tutor went over the same worksheet (monolog condition), or a tutor worked with a tutee to complete the worksheet (dialog condition). Two populations were tested: middle school ( $N = 32$ , 16 dyads per condition) and university ( $N = 40$ , 20 dyads per condition). The main finding was that learning was higher from dialog than monolog [12], but this effect only held for the university population.

## 2 Results

As indicated above, we used the SEANCE sentiment tool [7] to analyze the transcripts of student discussions as they worked on the diffusion worksheet. SEANCE incorporates a number of dictionaries, producing over 200 core indices and 20 component indices as its output. Each sentiment index is assigned a numeric value by SEANCE representing how present that sentiment is in the text (zero representing lack of that sentiment). We selected indices that (i) were present in the current corpora (i.e., values were not zero) and (ii) were relevant for the present analysis – these are shown in Table 1. While we were especially interested in sentiment-related information, we also included indices related to cognition (see Table 1). Most of the indices are self-explanatory (see [7] for details) but the *hu\_liu\_neg* index warrants an explanation: it reports on the proportion of negative to positive sentiment in a given corpora, with higher numbers indicating more negative sentiment. Thus, this index is opposite to and includes a broader scope than the *joy component* index that was also included. Our analysis was guided by the following research questions:

- (1) Does type of instructional video (monolog vs. dialog) impact student sentiment?
- (2) Is there a relationship between sentiment and student learning?

Because we had data from two populations, we included that as a variable in our analysis to see if and how it influenced the results.

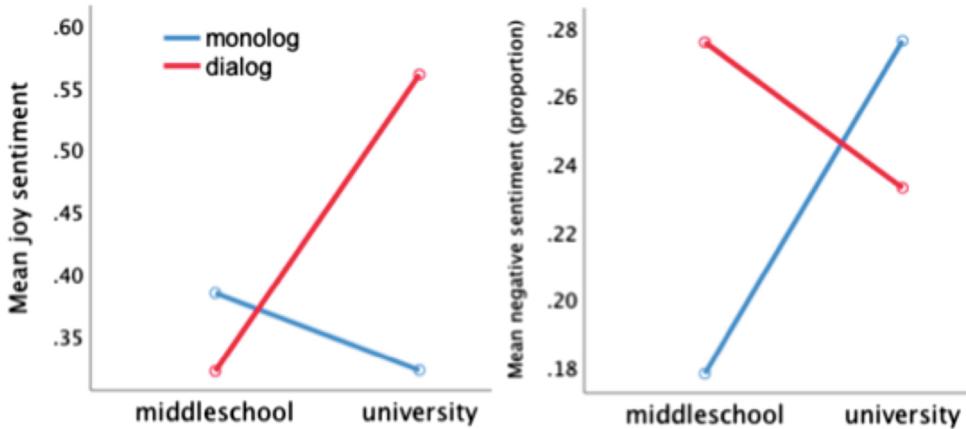
**Table 1.** Descriptive statistics for each condition and population (*M* and *SD*)

	<i>middle school</i>		<i>university</i>	
	dialog	monolog	dialog	monolog
<b>Category + SEANCE Indices</b>				
<b>positive / negative affect</b>				
<i>joy component</i>	.32 (.24)	.38 (.3)	.56 (.4)	.32 (.23)
<i>hu_liu_neg</i>	.28 (.19)	.18 (.2)	.23 (.2)	.28 (.14)
<b>arousal and dominance</b>				
<i>arousal</i>	4.1 (1.6)	4.4 (1.2)	4.8 (.3)	4.8 (.3)
<i>valence</i>	5.4 (2.2)	5.7 (1.5)	6.0 (.6)	6.1 (.5)
<i>dominance</i>	4.6 (1.8)	4.9 (1.3)	5.3 (.4)	5.3 (.3)
<b>cognition</b>				
<i>aptitude</i>	.18 (.04)	.19 (.02)	.21 (.04)	.21 (.04)
<i>attention</i>	.17 (.04)	.16 (.04)	.22 (.04)	.21 (.04)
<b>evaluation of information</b>				
<i>certainty component</i>	.19 (.07)	.18 (.04)	.22 (.06)	.25 (.03)
<i>action component</i>	.74 (.18)	.80 (.09)	.68 (.12)	.65 (.11)

**Effect of Instructional Video** To address the first research question, we analyzed the data from SEANCE for the selected indices using a  $2 \times 2$  between subjects ANOVA, with *instructional video* (dialog, monolog) and *group* (middle school, university) as the independent variables and the target SEANCE index as the dependent variable. Here, we focus on the significant results, which pertain to affect.

For the *joy component* index, we found a significant interaction between *instructional video* and *group* (see Fig. 1, left),  $F(1, 67) = 4.1, p = .048, \eta_p^2 = .06$ , indicating that the effect of instructional video type on *joy* depended on the population. As shown in Fig. 1 (left), university students in the dialogue condition expressed more *joy* when discussing the content with their partner than in the monologue condition, but this effect was not present for the middle school students (there was little difference for *joy* between monolog and dialog for that population).

These results are mirrored in a marginally-significant interaction for the *hu\_liu\_neg* component between *instructional video* and *group*,  $F(1, 65) = 2.8, p = .10, \eta_p^2 = .04$ , shown in Fig. 1 (right). Here we are analyzing negative affect, which is why the pattern is reversed compared to the *joy* analysis. The university dialog condition expressed proportionally less negative sentiment than did the monolog condition. While compatible with the results for *joy*, the *hu\_liu\_neg* analysis is broader in scope than the *joy* analysis because it encompasses all negative states and has the advantage of reporting the *proportion* of sentiment expressed, eliminating effects of verbosity (if any). Note that the analysis corresponding to *hu\_liu\_pos* index, focusing on the proportion of positive affect, produced identical statistics as expected given the proportional nature of this index. The other sentiment indices were highly similar between the two instructional video conditions (see Table 1) and did not show signs of interactions between them and population.



**Fig. 1.** Interaction between instructional video type and population for joy (left) and proportion of negative sentiment (right)

**Relation Between Sentiment and Learning** Prior analysis of the present data showed that university students learned more when given dialog videos as compared to monolog but that learning gains between the dialog and monolog conditions were similar for middle school students [12]. We wanted to see if this pattern manifested when we analyzed sentiment. Overall, collapsing across conditions, there was a trend that joy correlated with learning ( $r = .22, p = .07$ ), indicating that more joy was associated with higher learning (other relationships were not significant). How did condition and population affect that relationship? While we did not have sufficient power, descriptively we found a pattern mirroring the learning results reported in [12]. Specifically, for middle-school students, the regression slopes characterizing the association between learning and joy were flat and almost identical for monologue vs. dialogue. For the university population, the slope for the monolog condition was also flat, but the slope for the dialog condition was positive, indicating that joy was positively associated with learning.

### 3 Discussion and Future Work

To date, research on the benefits of learning from dialog in instructional videos over monolog have been confined to cognitive factors. To address this gap, we used NLP methods and specifically sentiment analysis to shed light on how each type of context (dialog, monolog), influenced learning for two populations (university, middle school). The data came from a study investigating student learning in a setting where pairs of students worked on a worksheet while watching an instructional video. Our results mirror the *cognitive patterns* observed in the original analysis of the data, namely that students learned more when given dialog-type instructional videos over monolog videos – and now we know they also expressed more positive sentiment when learning. The fact that this was only the case for the university population mirrors the prior cognitive findings [12], since the original analysis found that dialog fostered more learning only for the university students (middle school students learned similar amounts regardless of type of video). In general, given the recent interest in comparing

cognitive benefits of learning instruction that involves a dialog vs. a monolog, work is also needed on the affective (sentiment) front.

## References

1. Arroyo, I., Woolf, B., Burleson, W., Muldner, K., Rai, D., Tai, M.: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *Int. J. Artif. Intell. Educ.* **24**(4), 387 (2014)
2. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: 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. *Int. J. Hum Comput Stud.* **68**(4), 223–241 (2010)
3. Chi, M.T., Kang, S., Yaghmourian, D.L.: Why students learn more from dialogue-than monologue-videos: analyses of peer interactions. *J. Learn. Sci.* **26**(1), 10–50 (2017)
4. Craig, S.D., Chi, M.T., VanLehn, K.: Improving classroom learning by collaboratively observing human tutoring videos while problem solving. *J. Educ. Psychol.* **101**(4), 779–789 (2009)
5. Craig, S.D., Gholson, B., Brittingham, J.K., Williams, J.L., Shubeck, K.T.: Promoting vicarious learning of physics using deep questions with explanations. *Comput. Educ.* **58**(4), 1042–1048 (2012)
6. Craig, S., Driscoll, D., Gholson, B.: Constructing knowledge from dialog in an intelligent tutoring system: interactive learning, vicarious learning, and pedagogical agents. *J. Educ. Multimed. Hypermedia* **13**(2), 163–183 (2004)
7. Crossley, S., Kyle, K., McNamara, D.: Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social order analysis. *Behav. Res. Methods* **49**(3), 803 (2017)
8. D'Mello, S., Graesser, A.: Autotutor and affective autotutor. *ACM Trans. Interact. Intell. Syst.* **2**(4), 1–39 (2012)
9. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. *Learn. Instr.* **29**, 153–170 (2014)
10. Gholson, B., Craig, S.D.: Promoting constructive activities that support vicarious learning during computer-based instruction. *Educ. Psychol. Rev.* **18**(2), 119–139 (2006)
11. Kizilcec, R.: Showing face in video instruction: effects on information retention, visual attention, and affect. In: SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2095–2102 (2014)
12. Muldner, K., Lam, R., Chi, M.T.H.: Comparing learning from observing and from human tutoring. *J. Educ. Psychol.* **106**(1), 69–85 (2014)
13. Muller, D.A., Bewes, J., Sharma, M.D., Reimann, P.: Saying the wrong thing: improving learning with multimedia by including misconceptions. *J. Comput. Assist. Learn.* **24**(2), 144–155 (2008)
14. Muller, D.A., Sharma, M.D., Eklund, J., Reimann, P.: Conceptual change through vicarious learning in an authentic physics setting. *Instr. Sci.* **35**(6), 519–533 (2007)
15. Newman, H., Joyner, D.: Sentiment analysis of student evaluations of teaching. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 246–250. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_45](https://doi.org/10.1007/978-3-319-93846-2_45)
16. Pekrun, R., Goetz, T., Titz, W., Perry, R.P.: Academic emotions in students' self- regulated learning and achievement: a program of qualitative and quantitative research. *Educ. Psychol.* **37**(2), 91–105 (2002)

17. Putwain, D.W., Becker, S., Symes, W., Pekrun, R.: Reciprocal relations between students' academic enjoyment, boredom, and achievement over time. *Learn. Instr.* **54**, 73–81 (2018)
18. Van Gog, T., Verveer, I., Verveer, L.: Learning from video modeling examples: effects of seeing the human model's face. *Comput. Educ.* **72**, 323–327 (2014)
19. Worsley, M., Blikstein, P.: What's an expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In: Educational Data Mining Conference, pp. 235–240 (2011)
20. Yang, D., Kraut, R., Rose, C.: Exploring the effect of student confusion in massive open online courses. *J. Educ. Data Min.* **8**(1), 52–83 (2016)



# Automated Feedback on the Structure of Hypothesis Tests

Sietske Tacoma<sup>1</sup>(✉) , Bastiaan Heeren<sup>1,2</sup>, Johan Jeuring<sup>1,2</sup> , and Paul Drijvers<sup>1</sup> 

<sup>1</sup> Utrecht University, Utrecht, The Netherlands

s.g.tacoma@uu.nl

<sup>2</sup> Open University of the Netherlands, Heerlen, The Netherlands

**Abstract.** Hypothesis testing is a challenging topic for many students in introductory university statistics courses. In this paper we explore how automated feedback in an Intelligent Tutoring System can foster students' ability to carry out hypothesis tests. Students in an experimental group ( $N = 163$ ) received elaborate feedback on the structure of the hypothesis testing procedure, while students in a control group ( $N = 151$ ) only received verification feedback. Immediate feedback effects were measured by comparing numbers of attempted tasks, complete solutions, and errors between the groups, while transfer of feedback effects was measured by student performance on follow-up tasks. Results show that students receiving elaborate feedback solved more tasks and made fewer errors than students receiving only verification feedback, which suggests that students benefited from the elaborate feedback.

**Keywords:** Domain reasoner · Hypothesis testing · Intelligent tutoring systems · Statistics education

## 1 Introduction

Hypothesis testing is widely used in scientific research, and is therefore covered in most introductory statistics courses in higher education [2]. The topic is challenging for many students, because it requires an ability to follow a complex line of reasoning involving several abstract concepts and uncertainty [4, 6]. Students struggle to understand the role and interdependence of the concepts, or, in other words, the structure of hypothesis tests [14]. Appropriate feedback might support students in comprehending this structure. It should not only address the content of a current step, but also its relation to earlier steps. An Intelligent Tutoring System (ITS) can provide such sophisticated feedback on the level of steps and can provide diagnostics of student errors [11]. Feedback on the step level is generally more effective than feedback on the level of complete solutions [16].

Although ITSs vary considerably in design, they generally contain an expert knowledge module, a student model module, a tutoring module, and a user interface module [11]. Of these four components, the expert knowledge module, also referred to as domain reasoner [7], is the most domain-dependent. Two important paradigms for constructing domain reasoners are model-tracing, in which the ITS checks that a student

follows the rules of a model solution [1], and constraint-based modeling, in which the ITS checks whether a student violates constraints [10]. There exist ITSs that support hypothesis testing based on either of these approaches [9]. We combined the two in a single ITS supporting hypothesis tests. The contribution of this paper is a thorough evaluation of the impact of the combined ITS's feedback, which especially addresses the structure of hypothesis tests, on students' problem-solving behavior. It is guided by the question: does automated intelligent feedback on the structure of hypothesis tests contribute to student proficiency in carrying out hypothesis tests?

## 2 Methods

The domain reasoner for hypothesis testing is based on the Ideas framework [8], with a model-tracing approach as starting point, adding constraint-based modeling to identify inconsistencies in solution structure. For a description of its design, see [13].

The study consisted of a randomized controlled experiment in the context of a compulsory statistics course for first-year psychology students at a Dutch university. Students enrolled in the course were divided randomly into an experimental group (310 students) and a control group (309 students). Consent for the study was given by 163 students in the experimental group and 151 students in the control group. Participants were between 17 and 31 years old ( $M = 19.3$ ,  $SD = 1.7$ ) and 77% were female.

In five weeks of the ten-week course students received online homework sets in the Freudenthal Institute's Digital Mathematics Environment (DME; see [3]). The three homework sets that concerned hypothesis testing each contained two tasks in which students were asked to construct hypothesis tests by selecting steps from a drop-down menu and to completing these steps. For an example, see [13].

Two versions of the homework sets were designed: an experimental version with feedback on steps in the hypothesis testing procedure by the domain reasoner, and a control version with verification feedback on the contents of single steps only. Consequently, in the experimental version correct solutions needed to include four essential steps, since otherwise constraints would be violated. In the control version correct solutions only needed to include a correct conclusion about the null hypothesis.

Data for this study consisted of logs of the students' actions on the online homework sets, including all attempts students made to find correct answers, and all feedback requests. After exporting the logs from the DME, logs from students who did not give consent were deleted and all other logs were anonymized.

Three measures were used to assess immediate effects of feedback condition on the students' ability to solve hypothesis testing tasks: the number of tasks in which students attempted to construct steps, the number of tasks that students solved, and the number of errors students made in hypothesis test structure. Since samples were large, independent samples  $t$ -tests were used for all comparisons between groups [5]. Besides  $t$ -tests to compare groups over all tasks simultaneously, graphical representations were used to assess the differences between groups over time.

As promising effects of feedback on student performance do not automatically guarantee transfer to new tasks [12], student performance on follow-up tasks was also evaluated. From the three homework sets follow-up tasks on hypothesis testing were

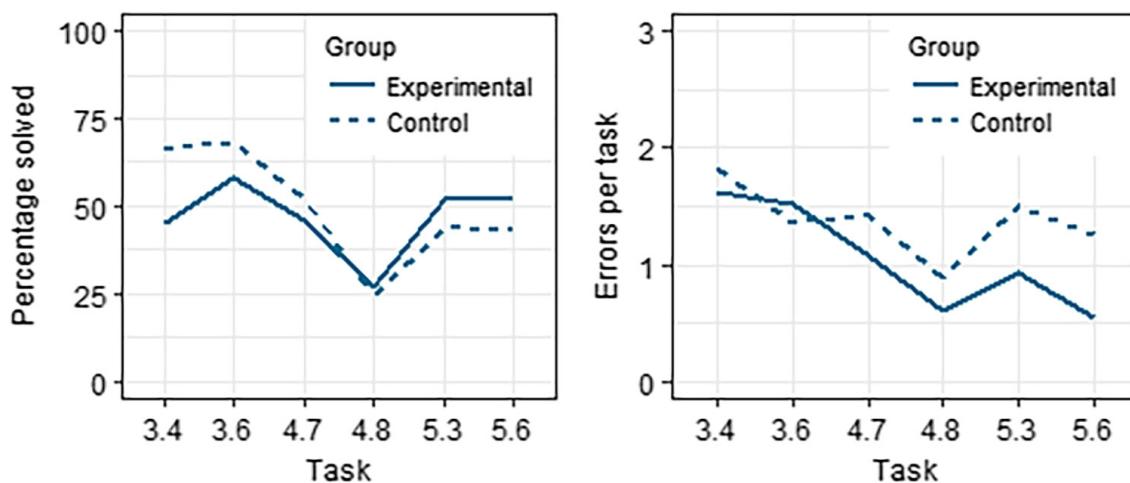
selected. For each student who received feedback on constructed steps at least once the ratio between number of selected tasks immediately answered correct and number of selected tasks attempted was calculated and ratios were compared between groups.

### 3 Results

In the hypothesis testing tasks students could choose to only fill in final answers, without constructing steps. Table 1 contains the mean number of tasks students worked on, the mean number of tasks in which they attempted to construct steps, and the mean number of complete solutions. In both groups, students attempted to construct steps for almost 80% of the tasks they worked on. The *t*-tests yielded no significant differences between groups. For the number of complete solutions, however, examining individual tasks did reveal different patterns. Figure 1 (left) displays the percentage of students who found complete solutions per task, as percentage of students who attempted to construct steps. For the first three tasks the control group outperformed the experimental group, while for the latter three tasks this was reversed.

**Table 1.** Mean number of tasks students worked on, constructed steps for and solved

	Experimental group (N = 163)	Control group (N = 151)	<i>t</i> (df = 312)	<i>p</i>
Tasks worked on	4.8 (1.5)	4.9 (1.5)	0.86	.391
Tasks tried constructing steps	3.8 (1.7)	3.9 (1.6)	0.62	.537
Tasks with complete solution	1.7 (1.8)	2.0 (1.7)	1.33	.184



**Fig. 1.** Percentage of students who correctly solved tasks according to group's assessment criteria (left) and mean number of errors in solution structure (right)

The final measure of immediate feedback effects was the number of errors students made in the structure of their hypothesis tests. The domain reasoner could diagnose 15

different errors in hypothesis test structure, such as a missing alternative hypothesis. On average, students in the experimental group made 1.12 ( $SD = 0.79$ ) different structure errors per solution, while students in the control group made 1.42 ( $SD = 0.86$ ) errors, which was significantly more,  $t(312) = 3.22$ ,  $p = .001$ , Cohen's  $d = .36$ . The graph in Fig. 1 (right) shows that in both groups the number of structure errors decreased over tasks, but this trend was stronger in the experimental group.

Regarding transfer to follow-up tasks, students in the experimental group ( $N = 158$ ) and the control group ( $N = 147$ ) were found to perform similarly: the mean ratio of correct answers was 0.72 ( $SD = 0.07$ ) in the experimental group and 0.71 ( $SD = 0.08$ ) in the control group. This implies that the domain reasoner feedback did not lead to better performance on follow-up tasks than verification feedback alone.

## 4 Conclusion and Discussion

We have evaluated the influence of ITS feedback addressing hypothesis test structure on student proficiency in carrying out hypothesis tests. The ITS feedback seemed to affect students' success in solving tasks completely; while students receiving ITS feedback performed worse than students receiving only verification feedback on the first three tasks, they outperformed the control group in the final three tasks, even with stricter assessment criteria. Additionally, students receiving ITS feedback made significantly fewer errors in hypothesis test structure than students receiving verification feedback only. This suggests that after familiarization, the ITS feedback effectively supported students in resolving their misunderstandings. This is in line with earlier findings that elaborate feedback is more effective than verification feedback [15]. Performance on follow-up tasks did not differ between groups, which implies that there was no automatic transfer from the positive results of the ITS feedback.

Such a lack of transfer has been found more often [12]. Here it could be caused by the design of the follow-up tasks, none of which specifically addressed the structure of hypothesis tests. From a research perspective, availability of tasks addressing the structure could have provided more insight in transfer of ITS feedback effects. From an educational perspective, availability of such tasks would have been valuable too, to avoid that students rely too much on the ITS feedback [12].

A second limitation of the study was that in this first large-scale implementation of the domain reasoner inevitably some unclarities became apparent. Nonetheless, even though sometimes receiving confusing feedback, students in general kept attempting the tasks and, as the results above show, did still benefit from the feedback.

Overall, this study has demonstrated that combining the model-tracing and constraint-based modeling paradigms can result in effective feedback on the structure of hypothesis tests. A challenging aspect of hypothesis testing that is not yet addressed by the ITS feedback is the role of uncertainty in the interpretation of the results from hypothesis tests [4]. Future research could focus on broadening the scope of the domain reasoner for hypothesis testing to include this reasoning with uncertainty.

**Acknowledgments.** We thank teachers Jeltje Wassenberg-Severijnen and Corine Geurts for their collaboration in designing teaching tasks and delivering the course. Furthermore, we thank Noeri Huisman, Martijn Fleuren, Peter Boon and Wim van Velthoven who helped with developing the domain reasoner.

## References

1. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. *J. Learn. Sci.* **4**(2), 167–207 (1995)
2. Carver, R., et al.: Guidelines for Assessment and Instruction in Statistics Education College Report 2016. American Statistical Association (2016). <http://www.amstat.org/education/gaise>
3. Drijvers, P., Boon, P., Doorman, M., Bokhove, C., Tacoma, S.: Digital design: RME principles for designing online tasks. In: Margolinas, C. (ed.) *Proceedings of ICMI Study 22 Task Design in Mathematics Education*, pp. 55–62. ICMI, Clermont-Ferrand (2013)
4. Falk, R., Greenbaum, C.W.: Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory Psychol.* **5**(1), 75–98 (1995)
5. Field, A.: *Discovering Statistics Using SPSS*, 3rd edn. Sage Publications, London (2009)
6. Garfield, J.B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., Zieffler, A.: Learning to reason about statistical inference. In: *Developing Students' Statistical Reasoning*, pp. 261–288. Springer, Dordrecht (2008). <https://doi.org/10.1007/978-1-4020-8383-9>
7. Goguadze, G.: ActiveMath - generation and reuse of interactive exercises using domain reasoners and automated tutorial strategies (Doctoral dissertation, Saarland University, Saarbrücken, Germany) (2011). <https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/26153/1/goguadzeDiss2011.pdf>
8. Heeren, B., Jeuring, J.: Feedback services for stepwise exercises. *Sci. Comput. Program.* **88**, 110–129 (2014). <https://doi.org/10.1016/j.scico.2014.02.021>
9. Kodaganallur, V., Weitz, R.R., Rosenthal, D.: A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *Int. J. Artif. Intell. Educ.* **15**(2), 117–144 (2005)
10. Mitrovic, A., Martin, B., Suraweera, P.: Intelligent tutors for all: the constraint-based approach. *IEEE Intell. Syst.* **4**, 38–45 (2007)
11. Nwana, H.S.: Intelligent tutoring systems: an overview. *Artif. Intell. Rev.* **4**(4), 251–277 (1990)
12. Shute, V.J.: Focus on formative feedback. *Rev. Educ. Res.* **78**(1), 153–189 (2008)
13. Tacoma, S., Heeren, B., Jeuring, J., Drijvers, P.: Automated feedback on the structure of hypothesis tests. In: Jankvist, U.T., van den Heuvel-Panhuizen, M., Veldhuis, M. (eds.) *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education*, pp. xx–yy. Freudenthal Group & Freudenthal Institute, Utrecht University and ERME, Utrecht, the Netherlands (2019)
14. Vallecillos, A.: Some empirical evidence on learning difficulties about testing hypotheses. In: *Bulletin of the International Statistical Institute: Proceedings of the Fifty-Second Session of the International Statistical Institute*, vol. 58, pp. 201–204 (1999)
15. Van der Kleij, F., Feskens, R., Eggen, T.: Effects of feedback in a computer-based learning environment on students' learning outcomes. A meta-analysis. *Rev. Educ. Res.* **85**(4), 475–511 (2015). <https://doi.org/10.3102/0034654314564881>
16. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**(4), 197–221 (2011). <https://doi.org/10.1080/00461520.2011.611369>



# Informing the Utility of Learning Interventions: Investigating Factors Related to Students' Academic Achievement in Classroom and Online Courses

Anna-Lena Theus<sup>(✉)</sup> and Kasia Muldner

Carleton University, Ottawa, Canada  
[{anna.theus, kasia.muldner}@carleton.ca](mailto:{anna.theus,kasia.muldner}@carleton.ca)

**Abstract.** To inform the utility of interventions delivered by adaptive educational technologies, we investigated the relationship between student grades and three target constructs, namely self-regulation, motivation, and self-theory of intelligence, in classroom and online settings. To do so, we collected data from a large sample of undergraduate university students ( $N = 1453$ ) enrolled in either a traditional face-to-face course or an online course and analyzed the data using hierarchical regression analysis. Prior research suggests that self-regulation, motivation, and self-theory of intelligence influence students' academic achievement. However, to date a hierarchical regression model including all three constructs has not been tested. Our results show that self-regulation and motivational constructs are positively associated with grades, but the self-theory of intelligence construct is not. Furthermore, we show that context does matter: the model for the classroom sample explained substantially more variance in grades as compared to the online model.

**Keywords:** Self-regulated learning · Motivation · Self-theory of intelligence · Academic achievement · Classroom courses · Online courses

## 1 Introduction and Related Work

There is established evidence that students' *self-regulation*, *motivation*, and *self-theory of intelligence* influence achievement in middle school, high school, and beyond. Students learn more when they self-regulate during learning (Garcia and Pintrich 1996; Pintrich and De Groot 1990; Pintrich et al. 1991), when they are motivated to learn (Garcia and Pintrich 1996), and when they believe that outcomes are controlled by effort rather than inherent ability (Blackwell et al. 2007; De Castella and Byrne 2015; Grant and Dweck 2003; Paunesku et al. 2015; Tempelaar et al. 2015). Given these findings, researchers in the artificial intelligence community have been investigating the utility of adaptive interventions integrated into learning platforms to promote beneficial behaviors related to these constructs (Arroyo et al. 2010; Karumbaiah et al. 2017; Hershkovitz and Nachmias 2008; Mudrick et al. 2018; Forbes-Riley and Litman 2011;

Rodrigo et al. 2008). To illustrate, Forbes-Riley and Litman (2011) manually labeled student utterances according to how motivated (engaged) students were, and correlated these labels with learning outcomes, showing that overall, lower engagement resulted in less learning. More recently, Karumbaiah et al. (2017) integrated messages into a tutoring system encouraging students to believe in the malleability of intelligence and showed that students who received more of these messages learned more.

While there is established evidence that the target constructs impact academic achievement *in general*, to date no work has investigated whether a given construct uniquely explains variance in grades over and above the other constructs. In the present paper, we present analysis on the relationship between target psychological constructs (i.e., self-regulation, motivation, and self-theory of intelligence) and students' grades through a hierarchical regression model that lets us analyze the relative contribution of each construct. To the best of our knowledge, our study is the first to analyze the relative contribution of each of the target constructs in both types of context (face-to-face vs. online), allowing us to understand the effects of the constructs in each setting and to help guide the process of designing and integrating interventions into instructional materials of large face-to-face and online university classes. In contrast to tutoring systems, these classes rely on a more basic form of educational technology, namely a Blackboard-style application that is used by instructors to post slides, instructional materials, or grades.

## 2 Method

We recruited a large sample of undergraduate university students ( $N = 1453$ ) enrolled in either a first year “Introduction to Psychology” traditional face-to-face classroom course ( $N = 707$ , referred to as the *classroom sample*) or a first year “Introduction to Psychology” for-credit online course ( $N = 746$ , referred to as the *online sample*). The participants completed an online personality traits survey made up of established questionnaires. We used all 15 scales of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al. 1991) to measure motivational and self-regulated learning (SRL) constructs, and the Implicit Theories of Intelligence Scale (Dweck 1999) to measure students’ self-theory of intelligence. Because self-efficacy is context specific (classroom vs. online), we included the Online Learning Self-Efficacy Scale (OLSES) (Zimmerman and Kulikowich 2016) for the online sample. As a measure of academic achievement, we used students’ final course grades.

The analysis relied on hierarchical regression analysis, a method that tests whether an independent variable or a set of independent variables accounts for a significant amount of variance in a dependent variable, over and beyond the variance accounted for by previously entered independent variables. The independent variables and their order of entry are specified prior to the analysis and based on previous work, with most relevant predictors entered into the model first (Cohen et al. 1983; Wampold and Freund 1987). To analyze the relative contribution of various sub-constructs (that were measured using varying scales), we used standardized regression coefficients.

### 3 Results

**Classroom Context.** To determine the unique contribution of the target constructs for explaining variance in student performance (i.e., grade in percentage) in the classroom context, we carried out a three-step hierarchical regression ( $N = 707$ ) with “grade” as the dependent variable and the constructs as the independent variables, entered in the following order (new categories of variables in a given step in italics):

**Step 1:** motivational constructs (these were six sub-constructs from the MSLQ, including intrinsic and extrinsic goal orientation, task value, control of learning beliefs, test anxiety, and self-efficacy)

**Step 2:** motivational constructs + *self-regulated learning (SRL) constructs* (the latter included nine sub-constructs from the MSLQ, including rehearsal, organization, elaboration, critical thinking, meta-cognition, time and study environment, effort regulation, peer learning and help seeking)

**Step 3:** motivational constructs + SRL constructs + *self-theory of intelligence* (one construct measuring belief in ability being fixed vs. malleable).

Model 1 was significant,  $F(6, 700) = 19.06, p < .001$ , and accounted for 13.3% of variance in grade (*adjusted R<sup>2</sup>* = 0.133). When we added the self-regulated learning (SRL) constructs in step 2, the model fit improved (*adjusted R<sup>2</sup>* = 0.173,  $p < .001$ ), and the overall model related to step 2 was significant,  $F(15, 691) = 10.84, p < .001$ . However, adding the self-theory of intelligence construct to the model (step 3) did not improve the fit of the model ( $p = .36$ ). Thus, self-theory of intelligence did not explain unique variance in course grades over and beyond the motivation and self-regulated learning constructs. The full model corresponding to step 3 was significant ( $p < .001$ ) and accounted for approximately 17.3% of the variance in course grades (*adjusted R<sup>2</sup>* = 0.173), with self-efficacy (a motivational sub-construct), effort regulation (a self-regulated learning sub-construct), and control of learning beliefs (a motivational sub-construct) being the strongest positive predictor variables for grades in classroom courses.

**Online Context.** To examine whether the context, online vs. face-to-face, influences the relationship between course grade and our target constructs, we repeated the analysis above but with the online sample ( $N = 746$ ). Thus, we ran a three-step hierarchical regression with “grade” as the dependent variable and the target constructs as independent variables entered in the same order as above:

**Step 1:** motivational constructs

**Step 2:** motivational constructs + *self-regulated learning (SRL) constructs*

**Step 3:** motivational constructs + SRL constructs + *self-theory of intelligence*.

Instead of using the self-efficacy sub-scale of the MSLQ, as done for the classroom sample analysis, we used the Online Learning Self-Efficacy Scale (OLSES) to measure students’ self-efficacy to take the context in which self-efficacy is measured into account. Mirroring the high-level pattern of results for the classroom sample, each of the models corresponding to the three steps were significant ( $p < .01$ ), and self-regulated learning constructs did significantly improve model fit when they were added

in step 2 ( $p < .01$ ). However, the model fit measured by adjusted  $R^2$  was quite modest for model 1 (*adjusted R<sup>2</sup>* = 0.02) and model 2 (*adjusted R<sup>2</sup>* = 0.04). Again, mirroring the results for the classroom context, adding self-theory of intelligence in step 3 did not improve model fit ( $p = .53$ ).

In the online context, while significant, the full model only accounted for 4% of variance in grade (*adjusted R<sup>2</sup>* = 0.04), a result we did not anticipate given that online classes arguably require more self-regulation than face-to-face classes.

## 4 Discussion

Our analysis using a three-step hierarchical regression model demonstrated that self-regulated learning accounts for unique variance in course grades over and beyond motivation both in classroom and online settings. These results are in line with prior work focusing on the classroom context (e.g., Bae 2014; Komarraju and Nadler 2013; Lynch 2006; Pintrich and De Groot 1990), although that work did not evaluate the relative contribution of each construct. In contrast, self-theory of intelligence did not significantly improve the model fit over and beyond the self-regulation and motivation constructs. Prior research did report a positive association between self-theory of intelligence and classroom grades, albeit with other statistical methods like path modeling (e.g., Blackwell et al. 2007; Chen and Pajares 2010; De Castella and Byrne 2015; Dweck and Master 2008; Gonida et al. 2006; Yeager et al. 2016). While it is true that we only entered the self-theory construct in step 3 of the model-building process, this does not appear to be the cause of difference between the present results and prior work, because the zero-order correlation between self-theory of intelligence and course grade in our data was not significant in either context ( $p > .1$ ), with very small corresponding  $r$  coefficients ( $r < .1$  in both classroom and online settings). This weak positive association between self-theory of intelligence and course grade we found is also reported in a recent meta-analysis. Specifically, Costa and Faria (2018) conducted a meta-analysis of 46 studies published between 2002 and 2017 to examine the relationship between self-theory of intelligence and students' academic achievement. The mean weighted effect size was quite small ( $r = 0.07$ ), similar to our classroom data. In our case, this weak relationship may be due to the domain, namely that we focused on psychology rather than the more challenging STEM topics used in some prior work (e.g., Blackwell et al. 2007), in which the impact of self-theory of intelligence may be more pronounced (Paunesku et al. 2015).

Our results have implications for the design of interventions embedded into course materials delivered through educational technologies, such as blackboards and tutoring systems, namely that they should focus on the motivation and self-regulation constructs, at least initially before moving on to the self-theory construct. This recommendation, however, comes with a key caveat: it only applies to the particular domain in our study and population, namely psychology classes with university students – it is an open question of how our results would transfer to, for instance, large university math or statistics classes. Our findings also highlight the importance of taking into account the context of student learning when making design recommendations for educational technologies. While the pattern of results between the classroom and online contexts

was the same in our work, the classroom model explained a modest amount of variance in grade, while the online model accounted only for a small amount of variance. This was somewhat unanticipated because we expected that self-regulated learning would be particularly important in an online setting, but results did not support that conjecture. We do not believe the instruments we used to be the cause, as they are established and validated through hundreds of studies. Nonetheless, our results show that there are variables not considered in our analysis that are contributing to grade (e.g., student behaviors *during* the class). These kinds of considerations will shape the next steps of our future work in designing interventions for large face-to-face and online classes.

## References

- Arroyo, I., Mehranian, H., Woolf, B.P.: Effort-based tutoring: an empirical approach to intelligent tutoring. In: Proceedings of Educational Data Mining Conference 2010, pp. 1–10 (2010)
- Bae, Y.: The relationships among motivation, self-regulated learning, and academic achievement. Doctoral dissertation, Texas A & M University (2014)
- Blackwell, L.S., Trzesniewski, K.H., Dweck, C.S.: Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. *Child Dev.* **78**(1), 246–263 (2007)
- Chen, J.A., Pajares, F.: Implicit theories of ability of Grade 6 science students: Relation to epistemological beliefs and academic motivation and achievement in science. *Contemp. Educ. Psychol.* **35**(1), 75–87 (2010)
- Cohen, J., Cohen, P., West, S.G., Aiken, L.S., et al.: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum, Hillsdale (1983)
- Costa, A., Faria, L.: Implicit theories of intelligence and academic achievement: a meta-analytic review. *Front. Psychol.* **9** (2018). <https://doi.org/10.3389/fpsyg.2018.00829>
- De Castella, K., Byrne, D.: My intelligence may be more malleable than yours: the revised implicit theories of intelligence (self-theory) scale is a better predictor of achievement, motivation, and student disengagement. *Eur. J. Psychol. Educ.* **30**(3), 245–267 (2015)
- Dweck, C.S.: Self-theories: Their Role in Motivation, Personality, and Development. Psychology Press, Philadelphia (1999)
- Dweck, C.S., Master, A.: Self-theories motivate self-regulated learning. In: Schunk, D.H., Zimmerman, B.J. (eds.) Motivation and Self-Regulated Learning: Theory, Research, and Applications, pp. 31–51. Taylor & Francis, New York (2008)
- Forbes-Riley, K., Litman, D.: When does disengagement correlate with learning in spoken dialog computer tutoring? In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNNAI), vol. 6738, pp. 81–89. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21869-9\\_13](https://doi.org/10.1007/978-3-642-21869-9_13)
- Garcia, T., Pintrich, P.R.: Assessing students' motivation and learning strategies in the classroom context: the motivated strategies for learning questionnaire. In: Birenbaum, M., Dochy, F.J.R.C. (eds.) Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge, vol. 42, pp. 319–339. Springer, Dordrecht (1996). [https://doi.org/10.1007/978-94-011-0657-3\\_12](https://doi.org/10.1007/978-94-011-0657-3_12)
- Gonida, E., Kiosseoglou, G., Leondari, A.: Implicit theories of intelligence, perceived academic competence, and school achievement: testing alternative models. *Am. J. Psychol.* **119**(2), 223–238 (2006)
- Grant, H., Dweck, C.S.: Clarifying achievement goals and their impact. *J. Pers. Soc. Psychol.* **85**(3), 541–553 (2003)

- Hershkovitz, A., Nachmias, R.: Developing a log-based motivation measuring tool. In: Proceedings of Educational Data Mining Conference 2008, pp. 226–233 (2008)
- Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B.P., Arroyo, I., Wixon, N.: Addressing student behavior and affect with empathy and growth mindset. In: Proceedings of the Educational Data Mining Conference 2017, pp. 96–103 (2017)
- Komarraju, M., Nadler, D.: Self-efficacy and academic achievement: why do implicit beliefs, goals, and effort regulation matter? *Learn. Individ. Differ.* **25**, 67–72 (2013)
- Lynch, D.J.: Motivational factors, learning strategies and resource management as predictors of course grades. *Coll. Stud. J.* **40**(2), 423–428 (2006)
- Mudrick, N.V., Sawyer, R., Price, M.J., Lester, J., Roberts, C., Azevedo, R.: Identifying how metacognitive judgments influence student performance during learning with MetaTutorIVH. In: Proceedings of the International Conference on Intelligent Tutoring Systems 2018, pp. 140–149 (2018)
- Paunesku, D., Walton, G.M., Romero, C., Smith, E.N., Yeager, D.S., Dweck, C.S.: Mind-set interventions are a scalable treatment for academic underachievement. *Psychol. Sci.* **26**(6), 784–793 (2015)
- Pintrich, P.R., De Groot, E.V.: Motivational and self-regulated learning components of classroom academic performance. *J. Educ. Psychol.* **82**(1), 33–40 (1990)
- Pintrich, P.R., Smith, D.A.F., Garcia, T., McKeachie, W.J.: A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). University of Michigan, Ann Arbor (1991)
- Rodrigo, M.M.T., et al.: The effects of motivational modeling on affect in an intelligent tutoring system. In: Proceedings of International Conference on Computers in Education 2008, pp. 64–72 (2008)
- Tempelaar, D.T., Rienties, B., Giesbers, B., Gijselaers, W.H.: The pivotal role of effort beliefs in mediating implicit theories of intelligence and achievement goals and academic motivations. *Soc. Psychol. Educ.* **18**(1), 101–120 (2015)
- Wampold, B.E., Freund, R.D.: Use of multiple regression in counseling psychology research: a flexible data-analytic strategy. *J. Couns. Psychol.* **34**(4), 372–382 (1987)
- Yeager, D.S., et al.: Using design thinking to improve psychological interventions: the case of the growth mindset during the transition to high school. *J. Educ. Psychol.* **108**(3), 374–391 (2016)
- Zimmerman, W.A., Kulikowich, J.M.: Online learning self-efficacy in students with and without online learning experience. *Am. J. Distance Educ.* **30**(3), 180–191 (2016)



# Auto-Sending Messages in an Intelligent Orchestration System: A Pilot Study

Kurt VanLehn<sup>1</sup>(✉), Salman Cheema<sup>2</sup>, Seokmin Kang<sup>1</sup>,  
and Jon Wetzel<sup>1</sup>

<sup>1</sup> Arizona State University, Tempe, AZ, USA

kurt.vanlehn@asu.edu

<sup>2</sup> Microsoft, Redmond, OR, USA

**Abstract.** FACT (Formative Assessment with Computational Technology) is an intelligent orchestration system. That is, because it helps the teacher manage the workflow of a complicated set of activities in the classroom, it is an orchestration system. Because it conducts tasks-specific and domain-specific analyses of the students' mathematical products and their group interactions, it is more intelligent than other orchestration systems. From analyzing videos of our iterative development trials, we realized that too many students needed help simultaneously, but the teacher could only visit one group at a time. Thus, we modified FACT to send a few messages to the students directly instead of sending all its advice to the teacher. This paper reports a successful pilot test of auto-sending.

**Keywords:** Classroom orchestration systems · Formative assessment · Intelligent tutoring system · Classroom evaluation

## 1 Introduction

Some lesson plans involve individual work, group work and whole-class discussions, and some also require that the teacher integrate workflows and ideas across all three planes of activity. “Classroom orchestration” refers to the planning and enacting of such integrated workflows [1]. A “classroom orchestration system” is intended to help the teacher with classroom orchestration [2–18].

Our system [19–23] was designed to increase the effectiveness of a particular set of mathematics lessons called the Classroom Challenges [24]. In their paper-based form, the Classroom Challenges (CCs) are known to be highly effective [25]. They exemplify teaching based on formative assessment [26], wherein teachers no longer give explanations and feedback, but instead keep students engaged in solving problems.

The CC students spend most of their time working in small groups on large posters, to which they add cards and handwriting. The posters can become extremely complicated. When teachers are circulating among the groups and they stop to visit a group, they often have only seconds to conduct a formative assessment of a complex poster.

We hypothesized that difficulties in formative assessment were preventing the CCs from being even more effective. Thus, the original goal of the FACT system was to

conduct and display formative assessments of posters in order to help visiting teachers. Thus, it was named Formative Assessment with Computational Technology (FACT).

FACT students edit an electronic document called a poster. Posters can have movable cards on them. Students can write or draw on the cards or the poster with a stylus, finger, mouse or keyboard. Students can also move the cards, pin them or resize them.

Students can edit their own individual poster or their group's poster. When editing a group poster, all the members of the group can edit simultaneously, just as one does with a shared Google document. Each student's ink is a different color, so students and teachers can tell who has contributed what.

As students work, teachers can monitor their work and control the class. They carry a tablet around the classroom that displays FACT's dashboard.

To conduct a formative assessment of students' work, FACT has many issue detectors. Most of them compare the students' work to expected work; these are called *product* detectors. FACT also has *process* detectors. These raise issues about the chronological pattern of students' edits, such as failing to collaborate. Similar collaboration detectors were quite accurate when used in a lab study [27].

FACT constantly decides which active issues are most important and shows them as alerts on the teacher's dashboard. When teachers peek at a student's work (i.e., view the student's poster on their dashboard), they see the top priority issue in a sidebar. They can scroll to view other issues in the sidebar. Each issue has both an explanation of it and questions that teachers can ask the student in order to open a visit discussing it. Alternatively, the teacher can push a Send button next to one of the questions, and it will appear in the student's inbox.

In order to help design FACT, video data from 14 trials of paper-based CCs were collected. During the iterative development of FACT, video data from 52 trials were collected. The later videos were collected as a formative evaluation rather than a summative evaluation. That is, they were collected to help us understand and redesign FACT. Nonetheless, we compared the videos of Paper and FACT trials and found:

1. FACT students wasted less time than the Paper students (5.9% for FACT vs. 10.4% for Paper;  $p = 0.013$ ). This was clearly due to replacing paper with electronic documents.
2. FACT students spent more time off-task than the Paper students (5.7% for FACT vs. 2.9% for Paper;  $p = 0.011$ ), probably due to the novelty of the stylus-tablet user interface.
3. FACT groups and Paper groups did not differ in how they worked together. Both FACT and Paper groups worked silently most of the time (53.8% for FACT vs. 67.7% for Paper). Groups rarely engaged in the most desirable form of collaboration, called co-construction or transactivity (2.8% for FACT vs. 4.0% for Paper).
4. FACT students self-corrected 47% of their errors, whereas Paper students self-corrected 67% of their 12 errors ( $p < .001$ ). Other errors were either left incorrect or corrected with the aid of the teacher. This suggests that productive struggle was more frequent for Paper students, contrary to our expectations.

5. When pairs were classified according the amount of self-correction of errors, 39% of the FACT pairs were struggling productively vs. 63% of the Paper pairs.
6. The mean number of teacher visits per lesson did not differ (27.8 for FACT vs. 27.2 for Paper;  $p = 0.890$ ), nor did the mean time between visit starts (4:25 for FACT vs. 5:40 for Paper;  $p = 0.182$ ).

The figures above indicate that many groups were not productively struggling and almost all were not collaborating properly. Yet teachers visited few groups (one per 4 or 5 min). Thus, when a teacher finished one visit and was deciding whom to visit next, *almost every group in the class needed to be visited*. Even if FACT helps the teacher make an optimal choice of whom to visit and what to say, there are many other groups left without a visit. Perhaps it would help if FACT could “visit” groups, too.

## 2 Auto-Sending and Its Pilot Test

As mentioned earlier, when teachers Peek at a student, they see a sidebar that shows questions that the teachers can use to initiate a visit. Teachers can also push a Send button to send a question directly to students. It then appears as a message in the student’s inbox.

In order to increase its effectiveness, FACT was modified to, so to speak, push the Send button pushes itself. After an activity began, it waited 5 min so students could get well started. It would then send students a message from their highest priority issue. It would always wait at least 2 min between sending messages to a group. We called this policy “auto-sending.”

As a pilot test of auto-sending, we conducted an AB evaluation in the middle school math classes of 2 teachers. Three classes had the full FACT system. Two classes had FACT with its detectors turned off, which meant that the teachers saw alerts neither on the dashboard nor when Peeking, and FACT auto-sent no messages.

We used the same methods and measures as in the formative evaluation reported earlier. The pattern of results during this pilot test were similar to those reported earlier, except for the most critical outcome, productive struggle, so we report just those results.

In order to help determine what encouraged students to be differentially productive, we divided all errors into four categories:

- The teacher visited the group when the error was being corrected or within the preceding 30 s.
- The students read a message in their inbox during the 30 s preceding correction of the error. The message could have been sent either by the teacher or by FACT.
- The students corrected the error without having consulted their inbox or the teacher during the preceding 30 s.
- The error had not yet been corrected when the activity ended.

Table 1 shows the error counts per condition per category. Comparing the On vs. Off conditions, the error distributions are reliability different (Chi-square,  $p < .001$ ). Students in the detectors On condition corrected significantly more errors without help from FACT or the teacher.

As in the formative evaluation, we classified pairs that corrected more than 50% of their errors by the end of the activity as productive. By this somewhat arbitrary criterion, all 7 pairs in the Analysis On classes were productive, while in the Analysis Off classes, only 3 of the 6 pairs were productive. This difference is reliable (Chi-sq,  $p = 0.004$ ). This is consistent with the hypothesis that turning the detectors on increased productive struggle.

### 3 Discussion

*Summary:* While iteratively developing FACT with aid of teachers, students and classroom observers, we recorded videos of 52 FACT classes and 14 paper-based CC classes. Video analyses suggest that although FACT made the workflow more efficient, there appeared to be little change in group interaction and teacher behavior. Contrary to our ambitions, FACT decreased productive struggle in the groups. The problem appeared to be simply that there weren't enough teacher visits to students because there is only one teacher but almost all groups need visits. Thus, we modified FACT to automatically send the messages that teachers could send. We compare two versions of FACT, with detectors turned either On or Off. Groups in the On condition more frequently struggled productively than groups in the Off condition. This is consistent with our hypothesis that the bottleneck in our classes is that more groups need to be visited, and that FACT's auto-send feature can at least partially fill the gap.

Although we refer to the conditions as detectors On vs. Off, many other factors co-varied with the manipulation including the classes, the time of day and the familiarity of the teachers with FACT. Thus, we cannot conclude that turning the detectors on *caused* students to correct more errors. Better-controlled experiments with more classes and teachers are needed.

A second problem is that errors are only one sign of struggle. Struggle could also show up as slow speed or extensive discussion.

In future work, the teacher's visits and the system's messages should be coordinated closely in order keep the teacher in charge of the class and yet maximize the impact on students. FACT will need a new kind of intelligence in order to support this sort of coordination. A larger, better-controlled evaluation would also be important.

**Table 1.** Errors

Correction type	Off	On
Teacher-assisted	0	0
FACT-assisted	0	4
Self-corrected	3	13
Uncorrected	4	1

**Acknowledgements.** This research was supported by the Bill and Melinda Gates Foundation under OPP1061281, the Diane and Gary Tooker Chair for Effective Education in Science, Technology, Engineering and Math and NSF grant 1840051. We gratefully acknowledge the contributions of all the members of the FACT project, past and present.

## References

1. Dillenbourg, P., Jermann, P.: Technology for classroom orchestration. In: Khine, M.S., Saleh, I.M. (eds.) *New Science of Learning: Cognition, Computers and Collaboration in Education*. Springer, New York (2010). [https://doi.org/10.1007/978-1-4419-5716-0\\_26](https://doi.org/10.1007/978-1-4419-5716-0_26)
2. Prieto, L.P., Dlab, M.H., Abdulwahed, M., Balid, W.: Orchestrating technology enhanced learning: a literature review and conceptual framework. *Int. J. Technol. Enhanced Learn.* **3** (6), 583–598 (2011)
3. Holstein, K., McLaren, B.M., Aleven, V.: Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10947, pp. 154–168. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_12](https://doi.org/10.1007/978-3-319-93843-1_12)
4. Molenaar, I., Knoop-van Campen, C.A., Hasselman, F.: The effects of learning analytics empowered technology on students' arithmetic skill development. In: *Proceedings of the Learning Analytics and Knowledge LAK 2017*, Vancouver, BC, Canada, pp. 614–515 (2017)
5. Molenaar, I., Knoop-van Campen, C.A.: Teacher dashboards in practice: usage and impact. In: *Proceedings European Conference on Technology Enhanced Learning: EC-TEL 2017*, pp. 125–138 (2017)
6. Håklev, S., Faucon, L., Hadzilacos, T., Dillenbourg, P.: FROG: rapid prototyping of collaborative learning scenarios. In: *Proceedings of the EC-TEL 2017* (2017)
7. Haklev, S., Faucon, L., Hadzilacos, T., Dillenbourg, P.: Orchestration graphs: enabling rich social pedagogical scenarios in MOOCs. In: *Proceedings of the Fourth (2017) ACM Conference on Learning@Scale*, pp. 261–264 (2017)
8. van Alphen, E., Bakker, S.: Lernanto: using an ambient display during differentiated instruction. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2334–2340. ACM (2016)
9. Mercier, E.: Teacher orchestration and student learning during mathematics activities in a smart classroom. *Int. J. Smart Technol. Learn.* **1**(1), 33–52 (2016)
10. Martinez-Maldonado, R., Yacef, K., Kay, J.: TSCL: a conceptual model to inform understanding of collaborative learning processes at interactive tabletops. *Int. J. Hum. Comput. Stud.* **83**, 62–82 (2015)
11. Martinez-Maldonado, R., Clayphan, A., Yacef, K., Kay, J.: MTFeedback: providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Trans. Learn. Technol.* **8**(2), 187–200 (2015)
12. Berland, M., Davis, D., Smith, C.P.: AMOEBA: designing for collaboration in computer science classrooms through live learning analytics. *Int. J. Comput. Support. Collaborative Learn.* **10**, 425–447 (2015)
13. Prieto, L.P., Asensio-Perez, J.I., Munoz-Cristobal, J.A., Jorrin-Abellan, I.M., Dimitriadis, Y., Gomez-Sanchez, E.: Supporting orchestration of CSCL scenarios in web-based distributed learning environments. *Comput. Educ.* **73**, 9–25 (2014)
14. Balestrini, M., Hernandez-Leo, D., Nieves, R., Blat, J.: Technology-supported orchestration matters: outperforming paper-based scripting in a jigsaw classroom. *IEEE Trans. Learn. Technol.* **7**(1), 17–30 (2014)
15. Higgens, S., Mercier, E., Burd, E., Joyce-Gibbons, A.: Multi-touch tables and collaborative learning. *Br. J. Educ. Technol.* **43**(6), 1041–1054 (2012)
16. Do-Lenh, S.: Supporting reflection and classroom orchestration with tangible tabletops. Ecole Polytechnique Federale de Lausanne (2012)

17. Alavi, H., Dillenbourg, P.: An ambient awareness tool for supporting supervised collaborative problem solving. *IEEE Trans. Learn. Technol.* **5**(3), 264–274 (2012)
18. Looi, C.-K., Lin, C.-P., Liu, K.-P.: Group scribbles to support knowledge building in a jigsaw method. *IEEE Trans. Learn. Technol.* **1**(3), 157–164 (2008)
19. VanLehn, K., Cheema, S., Wetzel, J., Pead, D.: Some less obvious features of classroom orchestration systems. In: Lin, L., Atkinson, R.K. (eds.) *Educational Technologies: Challenges, Applications and Learning Outcomes*, pp. 73–94. Nova Scientific Publishers (2016)
20. VanLehn, K., et al.: The effect of digital versus traditional orchestration on collaboration in small groups. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 369–373. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_69](https://doi.org/10.1007/978-3-319-93846-2_69)
21. VanLehn, K., Burkhardt, H., Cheema, S., Pead, D., Schoenfeld, A.H., Wetzel, J.: How can FACT encourage collaboration and self-correction?. In: Millis, K., Long, D., Magliano, J., Wiemer, K. (eds.) *Multi-Disciplinary Approaches to Deep Learning*, pp. 114–127. Routledge (2018)
22. Wetzel, J., et al.: A preliminary evaluation of the usability of an ai-infused orchestration system. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 379–383. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_71](https://doi.org/10.1007/978-3-319-93846-2_71)
23. VanLehn, K., et al.: Can an orchestration system increase collaborative, productive struggle in teaching-by-eliciting classrooms?. *Interactive Learning Environments*, in press
24. <http://map.mathshell.org/index.php>
25. Herman, J., et al.: The implementation and effects of the Mathematics Design Collaborative (MDC): early findings from Kentucky ninth-grade algebra 1 courses (CRESST Report 845). University of California at Los Angeles, National Center for Research on Evaluation, Standards and Student Testing (2015)
26. Black, P., Wiliam, D.: Assessment and classroom learning. *Assess. Educ. Principles Policy Pract.* **5**(1), 7–75 (1998)
27. Viswanathan, S.A., VanLehn, K.: Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Trans. Learn. Technol.* **11**(2), 230–242 (2018)



# Adaptive Learning Material Recommendation in Online Language Education

Shuhan Wang<sup>1</sup>(✉), Hao Wu<sup>2</sup>, Ji Hun Kim<sup>1</sup>, and Erik Andersen<sup>1</sup>

<sup>1</sup> Department of Computer Science, Cornell University, Ithaca, USA  
[{sw788,jk2227,ela63}@cornell.edu](mailto:{sw788,jk2227,ela63}@cornell.edu)

<sup>2</sup> Department of Computer Science, George Washington University, Washington,  
D.C., USA  
[fqq11679@gmail.com](mailto:fqq11679@gmail.com)

**Abstract.** In online language education, it is challenging to recommend learning materials that match the student’s knowledge since we typically lack information about the difficulty of materials and the abilities of each student. We propose a refined hierarchical structure to model vocabulary knowledge in a corpus and introduce an adaptive algorithm to recommend reading texts for online language learners. We evaluated our approach with a Japanese learning tool, finding that adding adaptivity into material recommendation significantly increased engagement.

## 1 Introduction

Engaging students with personalized content in online language learning presents two key challenges. First, we must prepare a corpus of learning materials that are organized by difficulty. Although we would like to utilize materials collected from the Internet, it is prohibitively expensive to ask experts to measure the difficulty of those materials. Second, we must assess each student’s competency level and recommend content that is appropriate for that student. Most existing content recommender systems for language learning are designed for formal learning scenarios and make recommendations based on standardized pre-assessment results. However, these systems may not scale easily to informal learning scenarios such as online learning, where we usually do not have accurate and standardized information of a student’s prior knowledge.

Existing assessment and recommendation systems [1,3,5] generally use unidimensional measurements for student ability and content difficulty, which is incomprehensive [2]. Ideally, a unified system could multidimensionally evaluate a student’s ability and the relative difficulty of learning materials in order to prepare future lessons for that student, without requiring prior information from the student or significant expert labor. Moreover, previous work on multidimensional knowledge structuring for grammar knowledge uses strict constraints to specify the relative difficulty between two texts [8]. However, this does not

scale to teaching vocabulary with a large online corpus since these strict constraints yield too few edges in the structure. To this end, we propose the *fuzzy partial ordering graph*, a refined hierarchical knowledge structure with relaxed constraints, which significantly increases the density of the knowledge structure.

We also present a material recommender system for online language learning that incorporates adaptive knowledge assessment. It collects authentic and up-to-date learning materials from the Internet and organizes them with a fuzzy partial ordering graph. It also uses a probabilistic function to balance assessment and recommendation throughout the learning process in order to improve student engagement. We evaluated our approach through *JRec*, an online Japanese language learning tool that recommends appropriate reading texts from the Internet based on the student's prior knowledge. Our user study demonstrates that our adaptive recommendation system led users to read 62.5% more texts than a non-adaptive recommendation version. This suggests that our multidimensional assessment can improve engagement in material recommendation.

## 2 Approach

**Fuzzy Partial Ordering Graphs.** In order to multidimensionally assess a student's knowledge and make recommendation accordingly, we need to measure the difficulty of each learning material and organize the corpus into a hierarchical structure. In our model, a reading text  $t_1$  is considered *fuzzily harder than* another text  $t_2$  if  $t_1$  covers a *majority* of vocabulary words in  $t_2$ . This also implies that students who understand  $t_1$  will also be able to understand  $t_2$ . Based on this fuzzy partial ordering, we model the vocabulary knowledge within a corpus of texts using a *fuzzy partial ordering graph*, in which each node denotes a text, and a directed edge from  $t_1$  to  $t_2$  indicates  $t_1$  is fuzzily harder than  $t_2$ .

This model improves our previous work in hierarchical knowledge structures [8] by increasing the number of partial ordering edges within the structure (the density). This previous work was based on a strict partial ordering, meaning that there is an edge from  $t_1$  to  $t_2$  only if  $t_1$  covers *all* knowledge in  $t_2$ . This strict partial ordering works well for grammar learning but may not scale well to vocabulary, since it is not common in an authentic corpus that a text covers all vocabulary knowledge of another text. Consequently, the strict partial ordering yields a vocabulary-based knowledge structure that is too sparse. The fuzzy partial ordering, however, addresses this issue by increasing the number of edges in the vocabulary-based knowledge structure to make it dense enough for assessment and recommendation.

To avoid unacceptable loss of confidence in our fuzzy partial orderings, we conducted a series of case studies in our corpus of 4,269 Japanese texts. We selected the fuzzy parameter  $\alpha = 0.8$ , meaning that  $t_1$  is fuzzily harder than  $t_2$  if  $t_1$  covers at least 80% of the vocabulary words in  $t_2$ . The fuzzy partial ordering graph with  $\alpha = 0.8$  has 71% more edges than the strict version.

**Adaptive Learning Material Recommendation.** Based on the fuzzy partial ordering graph, we seek to build a recommender system that carefully balances the trade-off between assessment and recommendation: in order for recommendations to be appropriate, the system needs to accurately assess each student; however, excessive assessment can potentially harm engagement because students might need to respond to too many problems that are far outside of their comfort zone. Our heuristics for assessment and recommendation are:

*The Assessment Heuristic:* Select the problem that maximizes the *expected* amount of information gained on the student’s prior knowledge. Formally, the assessment heuristic selects the problem  $s^*$  such that:

$$s^* = \arg \max_s [p_s n_s^+ + (1 - p_s) n_s^-] \quad (1)$$

where  $p_s$  indicates the probability that the student can solve  $s$ . If the student can solve  $s$ ,  $n_s^+$  represents how many problems we know that he/she can solve. Otherwise, if the student cannot solve  $s$ ,  $n_s^-$  represents how many problems we know that he/she cannot solve. Both  $n_s^+$  and  $n_s^-$  include  $s$  itself and exclude the problems we already know the student can/cannot solve before presenting  $s$ . The probability  $p_s$  can be estimated:  $p_s = N^+/(N^+ + N^-)$ , where  $N^+$  and  $N^-$  denote the number of presented problems that the student can/cannot solve.

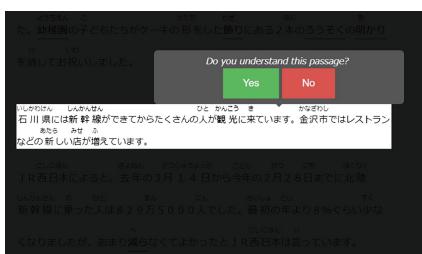
*The Recommendation Heuristic:* Select the problem that is directly harder than some problem that the student can solve. This heuristic is based on Vygotsky’s Zone of Proximal Development (ZPD) theory [7].

Since we believe that students are more engaged while solving a problem relevant to their experience, if there are multiple problems satisfying this requirement, pick the one that is *most relevant* to the student prior knowledge. Practically, the relevance is measured as the number of edges from that problem’s node to any solvable problem’s node in the fuzzy partial ordering graph.

*Balancing Assessment and Recommendation:* Our system uses a probabilistic function to balance the assessment and recommendation heuristics. To select the next problem, our system chooses the assessment heuristic with probability  $p = \#Prob/M$  and chooses the recommendation heuristic with probability  $1 - p$ . Here  $\#Prob$  represents the number of the problems that the student has experienced, regardless of whether he/she has solved those problems.  $M$  is a pre-set parameter that controls how fast our system transitions from assessment-favoring to recommendation-favoring. It also indicates that our system will always choose the recommendation heuristic after the student has experienced  $M$  problems.

### 3 Evaluation of Adaptive Recommendation

We evaluate our adaptive learning material recommender system in *JRec* (Fig. 1), a Japanese reading text recommendation tool. Our corpus of 380 articles was collected from NHK Easy [4], a Japanese news website for language learners. In order to accommodate beginners, our tool split those articles into 4,267 sentences and paragraphs so that students do not have to read the whole article.



**Fig. 1.** Screenshot of *JRec*, which draws texts from NHK Easy [4].

**Table 1.** Wilcoxon Rank-sum tests for all pairs of our four groups.

Comparison	Results
A.R. vs N.R.	$p = .035, Z = 2.109$
A.O. vs A.R.	$p = .766, Z = 0.298$
A.O. vs N.R.	$p = .022, Z = 2.287$
Rand vs N.R.	$p = .547, Z = 0.603$
A.O. vs Rand	$p = .294, Z = 1.049$
A.R. vs Rand	$p = .389, Z = 0.861$

Afterwards, it analyzed the hierarchical structure of vocabulary knowledge in the corpus and built a fuzzy partial ordering graph. When using this tool, users are directed to an NHK Easy webpage, read a recommended text (a paragraph or a sentence), and respond whether or not they understand it. Our tool highlights the recommended text and grays out the rest of the webpage. We recruited 368 users from the Japanese Learning Sub-reddit [6].

**Adding Adaptivity Improved Engagement Significantly.** We tested four different versions: (1) adaptive recommendation (which balances recommendation and assessment using  $M = 50$ ) and (2) non-adaptive recommendation (with no assessment incorporated), as well as (3) assessment-only and (4) random selection as additional baselines. We particularly wanted to see if adaptive recommendation is more engaging than non-adaptive recommendation, since this would demonstrate that adaptive assessment can enhance learning material recommendation.

In order to measure engagement, we recorded the number of texts each user read before leaving. 131 randomly selected users used adaptive recommendation (A.R.), 91 users used non-adaptive recommendation (N.R.), 115 users used assessment-only (A.O.) and 31 users used the random algorithm (Rand.). Users were assigned to these conditions at a ratio of 3:3:3:1, respectively, but the tool only recorded when a user responded to a text and some users may have quit before responding to the first problem. As a result, the number of recorded users in each group differs somewhat from the expected ratio.

Since our data was not normally distributed, we ran Wilcoxon Rank-sum tests for all pairs of the four groups (Table 1). We observed that the median user in the adaptive recommendation group (*Median* = 13) read 62.5% more texts than those in the non-adaptive recommendation group (*Median* = 8), and the difference between these two groups was statistically significant ( $p = .035$ ), which indicates that adaptive recommendation led users to read more texts than non-adaptive recommendation. In addition, the median user in the assessment-only group read 12 texts, which was also significantly more than that in the non-adaptive recommendation group ( $p = .022$ ). The median user in the random group read 8 texts and we did not find a statistically significant difference

compared to the other three groups, possibly because the random group had too few users. Overall, our results show that incorporating adaptive assessment can significantly enhance learning material recommendation in online learning.

**Acknowledgements.** This material is based upon work supported by the National Science Foundation under Grant No. IIS-1657176.

## References

1. Chen, C.M., Hsu, S.H., Li, Y.L., Peng, C.J.: Personalized intelligent m-learning system for supporting effective English learning. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2006, vol. 6, pp. 4898–4903. IEEE (2006)
2. Falmagne, J.-C., Cosyn, E., Doignon, J.-P., Thiéry, N.: The assessment of knowledge, in theory and in practice. In: Missaoui, R., Schmidt, J. (eds.) ICFCA 2006. LNCS (LNAI), vol. 3874, pp. 61–79. Springer, Heidelberg (2006). [https://doi.org/10.1007/11671404\\_4](https://doi.org/10.1007/11671404_4)
3. Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems. Routledge, London (1980)
4. NHK: NEWS WEB EASY (2019). [www3.nhk.or.jp/news/easy/](http://www3.nhk.or.jp/news/easy/)
5. Rasch, G.: Probabilistic models for some intelligence and attainment tests. ERIC (1993)
6. Reddit: Learn Japanese (2019). <https://www.reddit.com/r/LearnJapanese/>
7. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1980)
8. Wang, S., He, F., Andersen, E.: A unified framework for knowledge assessment and progression analysis and design. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 937–948. ACM (2017)



# Deep Knowledge Tracing with Side Information

Zhiwei Wang<sup>1</sup>, Xiaoqin Feng<sup>2</sup>, Jiliang Tang<sup>1</sup>, Gale Yan Huang<sup>2</sup>,  
and Zitao Liu<sup>2</sup>(✉)

<sup>1</sup> Data Science and Engineering Lab, Michigan State University, East Lansing, USA

{wangzh65,tangjili}@msu.edu

<sup>2</sup> TAL AI Lab, Beijing, China

{fengxqin,galehuang,liuzitao}@100tal.com

**Abstract.** Monitoring student knowledge states or skill acquisition levels known as knowledge tracing, is a fundamental part of intelligent tutoring systems. Despite its inherent challenges, recent deep neural networks based knowledge tracing models have achieved great success, which is largely from models' ability to learn sequential dependencies of questions in student exercise data. However, in addition to sequential information, questions inherently exhibit side relations, which can enrich our understandings about student knowledge states and has great potentials to advance knowledge tracing. Thus, in this paper, we exploit side relations to improve knowledge tracing and design a novel framework DTKS. The experimental results on real education data validate the effectiveness of the proposed framework and demonstrate the importance of side information in knowledge tracing.

## 1 Introduction

Knowledge tracing - where machine monitors students' knowledge states and their skill acquisition levels - is essential for personalized education and a fundamental part of intelligent tutoring systems [1, 7, 12, 15]. However, tracing student knowledge states is inherently challenging because of the complexity of human learning process, which involves a variety of factors from diverse domains such as neural science [3, 4], psychology [10], and education [8]. Meanwhile, the large amount of data produced by a growing number of online education platforms and recent advances of machine learning technology provide us with unprecedented opportunities to build advanced models for accurate knowledge tracing. Consequently, it has garnered widespread attention from researchers in both education and artificial intelligence communities [12, 14, 16]. Recently, one framework named Deep Knowledge Tracing (DKT) that is based on deep neural networks has shown superior performance over previously proposed knowledge tracing models [12]. Specifically, based on student historical answered questions, it is

---

Z. Wang and X. Feng—Work was done when the authors did internship in TAL AI Lab.

able to predict student performance on future questions with high accuracy. The key reason of the success of DKT is its ability to capture the sequential dependencies among questions embedded in the question answer sequences.

In fact, in addition to the sequential dependencies, questions naturally exhibit side relations due to their intrinsic properties. For example, questions are typically designed to improve certain concepts or skills. Thus, questions with similar underlying concepts or skills are inherently related. These relations can be represented as a question-question graph where nodes are questions and an edge exists in two questions if they are designed to examine similar sets of skills and concepts. The question-question graph provides rich information that can lead us to a better understanding of student knowledge states and exploiting such information has the great potential to improve the knowledge tacking performance.

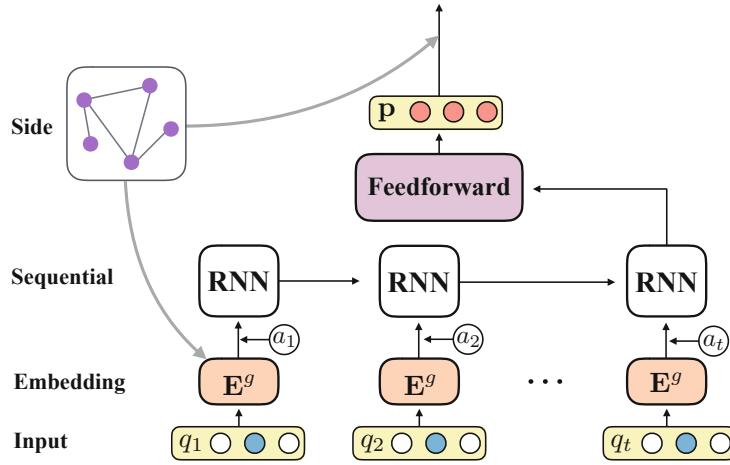
In this work, we exploit question relation information for better knowledge tracing and propose a framework DTKS that can capture both sequential dependencies and intrinsic relations of questions simultaneously. In summary, the contributions of this work are: (1) We identify the importance to incorporate side relations of questions into knowledge tracing; (2) We design a novel framework DTKS that provides a principled approach to capture both sequential and side relation information to model the student knowledge states and accurately predict their performance; and (3) We demonstrate the effectiveness of the proposed framework with real data.

## 2 Related Work

In this section, we briefly review the related works. Knowledge tracing is a long established research question and an essential task for computer assisted education [1, 2, 7, 12, 15, 16]. Previously, Bayesian Knowledge Tracing (BKT) based approach has been in predominate use [1, 15]. It represents the student knowledge state with a set of binary variables and each variable corresponds to student understanding of a single concept [7]. Other approaches such as Learning Factors Analysis [5] and ensemble methods [2] have also been proposed and achieved comparable performance with BKT. Recently, deep neural network based approach has become increasingly popular [12, 16]. Models in this line such as DKT [12] represent student knowledge state with continuous and expressive latent vectors and are able to capture the complexity of knowledge state. However, few of them incorporates the question relation information, which could be very helpful for knowledge tracing tasks.

## 3 The Proposed Framework

In this section, we introduce our proposed model DTKS that is able to incorporate question relations in modeling student knowledge state. The overall structure of the proposed model is shown in Fig. 1. Before detailing each layer next, we first introduce the notations. Vectors and matrices are represented with bold



**Fig. 1.** The network architecture of the proposed framework.

lower-case letters such as  $\mathbf{h}$  and bold upper-case letters such as  $\mathbf{W}$ . In addition, the  $i^{th}$  entry of vector  $\mathbf{h}$  is denoted as  $\mathbf{h}(i)$  and the entry at the  $i^{th}$  row and  $j^{th}$  column of matrix  $\mathbf{W}$  as  $\mathbf{W}(i, j)$ .

**Input and Embedding Layers:** The input of the framework is the student past question answer sequence  $S = (x_1, x_2, \dots, x_n)$ , where  $x_j = (q_j, a_j)$  involves a question  $q_j$  and the correctness of the student answer denoted as  $a_j \in \{0, 1\}$ . We represent  $x_j$  as  $\mathbf{x}_j$  using an embedding layer.

**The Sequential Layer:** We take advantage of RNN models to trace student knowledge states. Specifically, at time step  $t$ , RNN maintains a latent vector  $\mathbf{h}_t \in \mathbb{R}^{n_h}$  representing student knowledge state through the following cell structure:  $\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b})$ . Thus, this recursive structure naturally describes the evolution of student knowledge state  $\mathbf{h}_t$  that is driven by the previous knowledge state  $\mathbf{h}_{t-1}$  and current observation  $\mathbf{x}_t$ . In practice, more advanced recurrent cells such as long short-term memory unit (LSTM) and gated recurrent unit cell (GRU) [6, 11] often achieve better performance than original cell. We investigate both of them in this work. After sequential layer, we design a feedforward layer to predict the student future's response to each question based on the final knowledge state representation  $\mathbf{h}$  by following equation:  $\mathbf{p} = \sigma(\mathbf{h}\mathbf{W}^p + \mathbf{b}^p)$ , where  $\mathbf{p}(i)$  indicates the probability that the student can answer the  $i^{th}$  question correctly.

**The Side Layer:** In side layer, two model components are designed to capture the question relation. Firstly, instead of using embedding layers, we apply graph embedding algorithms such as LINE [13] and Node2Vec [9] to the question-question relation graphs to obtain the question representations that preserve the question relations. Secondly, to impose the intuition that if a pair of questions (e.g.,  $i^{th}$  and  $j^{th}$  questions) requires similar skills or involves similar concepts, the probability for a given student answering the two questions correctly should also be similar, we design the following regularization term  $\mathcal{L}_r = \frac{1}{2}\mathbf{p}^T \mathbf{L} \mathbf{p}$ , where

$\mathbf{L}$  is the Laplacian matrix of adjacent matrix  $\mathbf{A}$  representing the question relation graph.

**The Loss Function:** With the prediction  $\mathbf{p}$  obtained from sequential layer and the relation regularization term  $\mathcal{L}_r$ , we define the loss function of the proposed framework DKTS for each training data as  $\mathcal{L} = \mathcal{L}_p + \alpha\mathcal{L}_r$ , where  $\alpha$  is adopted to control the contribution of relation regularizer and  $\mathcal{L}_p$  is the binary cross-entry loss that is defined as:

$$\mathcal{L}_p = -a_{t+1} \log(\mathbf{p}^T \mathbf{q}_{t+1}) - (1 - a_{t+1}) \log(1 - \mathbf{p}^T \mathbf{q}_{t+1}) \quad (1)$$

where  $\mathbf{q}_{t+1}$  is the one-hot encoding of the question at time step  $t + 1$ .

## 4 Experiment

In this section, we conduct experiments on real education data to verify the effectiveness of the proposed model.

**Table 1.** Performance comparison Results. ‘NA’ indicates not applicable.

Method	Question embedding		
	Gaussian	LINE	Node2Vec
RNN	0.6527	0.7015	0.6988
LSTM	0.6999	0.7152	0.7140
GRU	0.7074	0.7173	0.7165
DKTS	NA	0.7338	0.7340

**Dataset:** We collect a student question answer behavior dataset from one of the most popular GMAT preparation mobile applications in China. It contains 8,684 questions and 90831 anonymized students and is cleaned by a filtering process. For each student, we collect her question answer behaviors and form a sequence of behaviors ordering by time information. A question relation graph is constructed according to the underlying knowledge and skills.

**Baselines:** In baselines, we use RNN, LSTM, and GRU to model the students knowledge state and represent question by embedding vectors that are sampled from a Gaussian distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  (Gaussian), or obtained through graph embedding algorithms (LINE, Node2Vec). Note that previously proposed DKT model uses LSTM to learn student knowledge state with question representation vectors sampled from Gaussian distribution [12].

**Experimental Results:** We evaluate the prediction performance by area under the curve (AUC) and a higher AUC indicates better performance. The results are shown in Table 1. We observe that (1) The embedding vectors that preserve the

question relation information significantly improve the prediction performance, which clearly demonstrates the importance of question relation information for knowledge tracing tasks; and (2) The proposed framework DKTS outperforms all other methods by a large margin. We contribute the superior performance of the proposed model to its ability to incorporate question relation information.

## 5 Conclusion

In this work, we exploit question relation information for knowledge tracing tasks. Specifically, we design a novel deep neural network based framework that is able to capture the sequential dependencies and intrinsic relations of questions to trace the student knowledge state. Moreover, we evaluate the proposed framework with real education data on student future interaction prediction task. The experimental results have clearly demonstrated the importance of the question relation information and the proposed framework outperforms state-of-the-art baselines significantly.

**Acknowledgements.** Zhiwei Wang and Jiliang Tang are supported by the National Science Foundation (NSF) under grant numbers IIS-1714741, IIS-1715940, IIS-1845081 and CNS-1815636, and a grant from Criteo Faculty Research Award.

## References

1. Baker, R.S.J., Corbett, A.T., Aleven, V.: More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In: Woolf, B.P., Aimeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 406–415. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69132-7\\_44](https://doi.org/10.1007/978-3-540-69132-7_44)
2. Baker, R.S.J., Pardos, Z.A., Gowda, S.M., Nooraei, B.B., Heffernan, N.T.: Ensembling predictions of student knowledge within intelligent tutoring systems. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 13–24. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-22362-4\\_2](https://doi.org/10.1007/978-3-642-22362-4_2)
3. Bassett, D.S., Porter, M.A., Mucha, P.J., Carlson, J.M., Grafton, S.T.: Dynamic re-configuration of human brain networks during learning. PNAS **108**(18), 7641–7646 (2011)
4. Caine, R.N., Caine, G.: Understanding a brain-based approach to learning and teaching. Educ. Leadersh. **48**(2), 66–70 (1990)
5. Cen, H., Koedinger, K., Junker, B.: Learning factors analysis – a general method for cognitive model evaluation and improvement. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 164–175. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_17](https://doi.org/10.1007/11774303_17)
6. Cho, K., et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) (2014)
7. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **278**(4), 4–253 (1994)

8. Felder, R.M., Silverman, L.K., et al.: Learning and teaching styles in engineering education. *Eng. Educ.* **78**(7), 674–681 (1988)
9. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: KDD (2016)
10. Hilgard, E.R.: Theories of learning (1948)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Piech, C., et al.: Deep knowledge tracing. In: NIPS (2015)
13. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: WWW (2015)
14. Wang, L., Sy, A., Liu, L., Piech, C.: Deep knowledge tracing on programming exercises. In: L@S (2017)
15. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized bayesian knowledge tracing models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 171–180. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_18](https://doi.org/10.1007/978-3-642-39112-5_18)
16. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: WWW (2017)



# Analysis of Holistic Interactions Between Lecturers and Students in Lectures

Eiji Watanabe<sup>1</sup>(✉) , Takashi Ozeki<sup>2</sup>, and Takeshi Kohama<sup>3</sup>

<sup>1</sup> Konan University, Kobe 658-8501, Japan

[e\\_wata@konan-u.ac.jp](mailto:e_wata@konan-u.ac.jp)

<sup>2</sup> Fukuyama University, Fukuyama 729-0292, Japan

<sup>3</sup> Kindai University, Kinokawa 649-6493, Japan

**Abstract.** This paper proposes modeling methods (i) for the interaction between behaviors of the lecturer and students, (ii) for the interaction between the lecturer and holistic behaviors of students. Moreover, we discuss modeling results based on experimental results.

**Keywords:** Lecture · Lecturer · Student · Behaviors · Interaction · Holistic interaction · Time-series analysis

## 1 Introduction

In lectures delivered using blackboards, the lecturer provides explanations and writing on the blackboards. Furthermore, students take notes and listen to the lecturer. The timing of behaviors by some students differs from other students depending on their interests and level of understanding. Therefore, to control a lecture, the lecturer must evaluate such interests and understandings based on the non-verbal behaviors of students. Cheng et al. [1] discussed the prediction of lecture ratings by using non-verbal behaviors of lecturers. Moreover, Rosati et al. [4] discussed the interactions between student learning styles, teaching presentation modes and student performance. Raca et al. [3] have proposed the methods for the extraction of the student behavior by cameras and the detection of “sleepers’ lag”. On the other hand, in [2], a conceptual model TSCL (Tabletop-Supported Collaborative Learning) has been proposed for understanding of the collaborative learning process.

The authors have already proposed a method to model the influence of lecturer’s behavior on students by using both image processing and multilayered neural networks [5–7]. However, the lecturer is also influenced by the behaviors of students; thus, we must discuss this influence and the interactions between behaviors of the lecturer and students. Furthermore, we have to consider the influence of the holistic behaviors of students on the lecturer. In this paper, we propose modeling methods for (i) the interaction between the lecturer and all students, (ii) the interaction between the lecturer and holistic behaviors of all students. Furthermore, we discuss analysis results for the two lectures.

---

Supported by JSPS KAKENHI Grant Number 16K00499, 19K12261 and 19K03095.

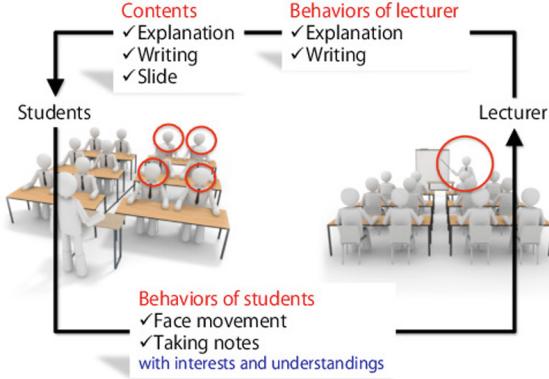
© Springer Nature Switzerland AG 2019

S. Isotani et al. (Eds.): AIED 2019, LNAI 11626, pp. 309–313, 2019.

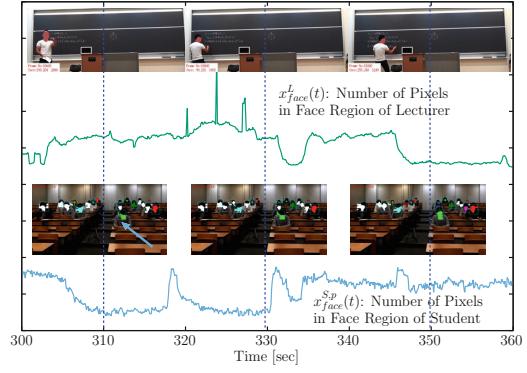
[https://doi.org/10.1007/978-3-030-23207-8\\_57](https://doi.org/10.1007/978-3-030-23207-8_57)

## 2 Modeling of Behaviors of Lecturers and Students

In lectures using the blackboard, the lecturer has the following behaviors; (i) looking at students, (ii) writing on a blackboard, and (iii) explanation (Fig. 1). On the other hand, the students have the following behaviors; (i) looking at the blackboard, the lecturer and other students, (ii) taking notes, and (iii) listening to the explanation. We adopt the number of pixels in the face as the feature for the behaviors of lecturers  $\{x^L(t)\}$  and students  $\{x^{S,p}(t)\}$  [5] (Fig. 2).



**Fig. 1.** Interaction



**Fig. 2.** Features of behaviors

### 2.1 Modeling of Interactions Between Lecturers and Students

First, we define the number of pixels in the facial region of the  $p$ -th student as  $x^{S,p}(t)$  and the number of pixels in the facial region of the lecturer as  $x^L(t)$  by using image processing [5]. Since the behavior  $x^L(t)$  of a lecturer is influenced by students, we can model the behavior  $x^L(t)$  of the lecturer by Eq. (1)

$$x^L(t) = \alpha^L f \left( \sum_{n=1}^N w_n^L x^L(t-n) \right) + \sum_{p=1}^P \alpha^{S,p} f \left( \sum_{n=1}^N w_n^{S,p} x^{S,p}(t-n) \right) + e(t). \quad (1)$$

where  $e(t)$  denotes Gaussian noise. Here, the weights  $\alpha^L$  and  $\alpha^{S,p}$  denote the influence of the behavior of the lecturer and the  $p$ th student. Moreover, the weights  $w_n^L$  and  $w_n^{S,p}$  represent the time correlations of the behaviors of the lecturer and the  $p$ th student.  $f(\cdot)$  denotes the sigmoid function  $f(x) = \tanh x$ .

Next, we can model the behavior  $x^{S,q}(t)$  of the  $q$ th student by Eq. (2).

$$x^{S,q}(t) = \beta^q f \left( \sum_{n=1}^N w_n^L x^L(t-n) \right) + \sum_{p=1}^P \beta^{q,p} f \left( \sum_{n=1}^N w_n^{S,p} x^{S,p}(t-n) \right) + e(t). \quad (2)$$

where the weights  $\beta^q$  and  $\beta^{q,p}$  denote the influence of the behavior of the  $q$ th student and the  $p$ th student. Similarly, the weights  $w_n^L$  and  $w_n^{S,p}$  represent the time correlations of the behaviors of the lecturer and the  $p$ th student.

## 2.2 Modeling of Holistic Interactions Between Lecturers and Students

When the number of students becomes large, it is difficult for the lecturer to monitor each student. In this case, the lecturer tends to grasp the holistic behaviors of students. We discuss the interactions between lecturers and holistic behaviors (the sum  $X^S(t) = \sum_{p=1}^P x^{S,p}(t)$ ) of students. First, we introduce the model for the behavior  $x^L(t)$  of the lecturer. Here, we assume that the behavior  $x^L(t)$  of the lecturer is influenced by the past behaviors of the lecturer oneself and the past holistic behaviors  $X^S(t - n)$ .

$$x^L(t) = A^L f \left( \sum_{n=1}^N W_n^L x^L(t-n) \right) + A^S f \left( \sum_{n=1}^N W_n^S X^S(t-n) \right) + e(t). \quad (3)$$

where  $A^L$  and  $A^S$  denote the influence of the behavior of the lecturer and the holistic behaviors of all students. Moreover, the weights  $W_n^L$  and  $W_n^S$  represent the time correlations of the behaviors of the lecturer and all students. Next, we introduce the following model for the holistic behavior  $X^S(t)$  of all students. Here, the holistic behaviors  $X^S(t)$  of all students are influence by the past behaviors of other students and the past behavior  $x^L(t - n)$  of the lecturer.

$$X^S(t) = B^L f \left( \sum_{n=1}^N W_n^L x^L(t-n) \right) + B^S f \left( \sum_{n=1}^N W_n^S X^S(t-n) \right) + e(t). \quad (4)$$

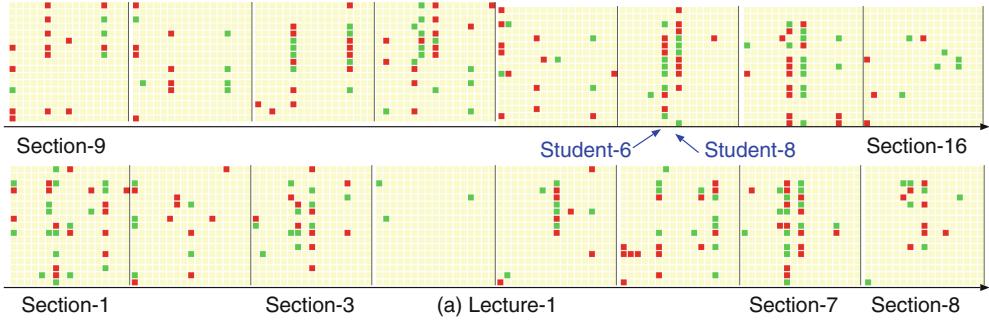
where  $B^L$  and  $B^S$  denote the influences of the behaviors of the lecturer and the holistic behaviors of all students. Similarly, the weights  $W_n^L$  and  $W_n^S$  can represent the time correlations of the behaviors of the lecturer and all students.

## 3 Analysis Results

We evaluated two lectures under the following conditions; (i) content: the derivation of formulas for some trigonometric functions (about 20 [min]), (ii) lecturer: one, students: 16 undergraduates, (iii) cameras: resolution:  $960 \times 540$  [dot], frame rate: 10 [fps], and (vi) the length of each section for modeling  $T$ : 10 [sec].

### 3.1 Modeling of Interactions Between Lecturers and Students

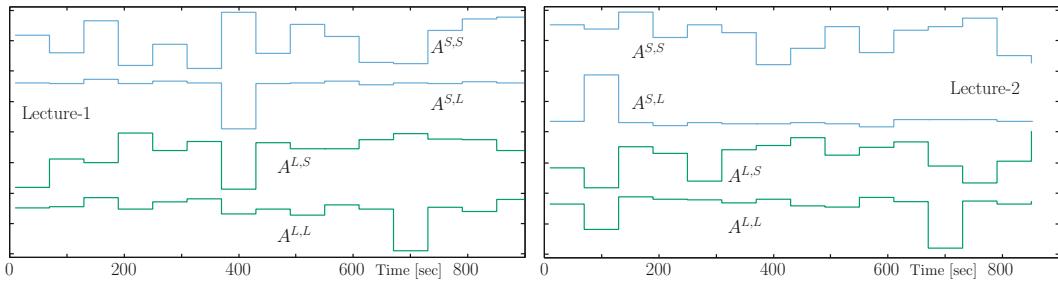
Each lecture is divided into the section having one minute and the length  $L$  is set as 10 [sec]. In Fig. 3, we show the weights  $\alpha^L$ ,  $\alpha^{S,p}$ ,  $\beta^q$  and  $\beta^{q,p}$  in Lecture-1 and -2. Here, the length of the divided section is 60 [sec]. In this figure, we can see the followings; (i) Lecture-1: In Section-3, 7, 12, 14 and 15, the weights  $\beta^{q,p}$  for Student-6 and 8 are larger than other sections. From video images, we can confirm that Student-6 and 8 have similar behaviors with each other and have different behaviors from others. (ii) Lecture-2 (not shown): In Section-1 and 3, the weights  $\beta^{q,p}$  for Student-5 are larger than those for other students.



**Fig. 3.** Modeling results of interactions between behaviors of lecturers and students (the weights of Eqs. (1) and (2); ■: weights  $\leq -1$ , □:  $|weights| < 1$ , ■: weights  $\geq 1$ ) (Color figure online)

### 3.2 Modeling of Holistic Interactions Between Lecturers and Students

Figure 4 shows the change of the weights as follows; (i) Lecture-1: From 400 [sec], the weights  $A^{L,S}$ ,  $A^{S,L}$  and  $A^{S,S}$  change largely. In section [0,400] [sec], the lecturer has explained the general contents. From 400 [sec], the lecturer has explained using the blackboard. Therefore, many students have taken notes and the above weights changed largely. In section [670, 730] [sec], the lecturer has begun to explain the solution and the weights  $A^{L,L}$  change largely. (ii) Lecture-2: In section [70, 130] [sec], the weights change largely. In the video, the lecturer has explained the interesting contents for students and the behaviors of students have been active. In section [670,730] [sec], the lecturer has explained using an important keyword and the weights  $A^{L,L}$  change largely.



**Fig. 4.** Modeling results for holistic interactions between lecturers and students.

## 4 Conclusions

In this paper, we have discussed the modeling for the interactions between lecturers and students. From analysis results, we have confirmed the followings; (i) In the interactions between lecturers and students, the weights for specific students become large, and (ii) In the holistic interactions, the relations between the changes of weights and the progress of the lecture.

## References

1. Cheng, D.S., Salamin, H., Salvagnini, P.: Predicting online lecture ratings based on gesturing and vocal behavior. *J. Multimodal User Interfaces* **8**(2), 151–160 (2014)
2. Martinez-Maldonado, R., Yacef, K., Kay, J.: TSCL: a conceptual model to inform understanding of collaborative learning processes at interactive tabletops. *Int. J. Hum. Comput. Stud.* **83**, 62–82 (2015)
3. Raca, M., Tormey, R., Dillenbourg, P.: Sleepers' lag: study on motion and attention. *J. Learn. Anal.* **3**(2), 239–260 (2016)
4. Rosati, P., Dean, R.K., Rodman, S.M.: A study of the relationship between students' learning styles and instructors' lecture styles. *IEEE Trans. Educ.* **31**(3), 208–212 (1988)
5. Watanabe, E., Ozeki, T., Kohama, T.: Extraction of relations between behaviors by lecturer and students in lectures. In: Proceedings of IEEE Conference on Automatic Face & Gesture Recognition and Workshops, pp. 945–950 (2011)
6. Watanabe, E., Ozeki, T., Kohama, T.: Analysis and extraction of behaviors by students in lectures. In: Proceedings of the EDM 2014, pp. 1–2 (2014)
7. Watanabe, E., Ozeki, T., Kohama, T.: Analysis of interactions between lecturer and students. In: Proceedings of the LAK 2018, pp. 1–5 (2018)



# Take the Initiative: Mixed Initiative Dialogue Policies for Pedagogical Agents in Game-Based Learning Environments

Joseph B. Wiggins<sup>1</sup>(✉), Mayank Kulkarni<sup>1</sup>, Wookhee Min<sup>2</sup>,  
Kristy Elizabeth Boyer<sup>1</sup>, Bradford Mott<sup>2</sup>, Eric Wiebe<sup>2</sup>, and James Lester<sup>2</sup>

<sup>1</sup> University of Florida, Gainesville, FL 32601, USA

{jbwiggi3, mayankk91, keboyer}@ufl.edu

<sup>2</sup> North Carolina State University, Raleigh, NC 27695, USA

{wmin, bwmott, wiebe, lester}@ncsu.edu

**Abstract.** Pedagogical agents have been shown to be highly effective for supporting learning in a broad range of contexts, including game-based learning. However, there are key open questions around how to design dialogue policies for pedagogical agents that support students in game-based learning environments. This paper reports on a study to investigate two different agent dialogue policies with regard to conversational initiative, a core consideration in dialogue system design. In the User Initiative policy, only the student could initiate conversations with the agent, while in the Mixed Initiative policy, both the agent and the student could initiate conversations. In a study with 67 college students, results showed that the Mixed Initiative policy not only promoted more conversation, but also better supported the goals of the game-based learning environment by fostering exploration, yielding better performance on in-game assessments, and creating higher student engagement.

**Keywords:** Pedagogical agents · Game-based learning · Initiative

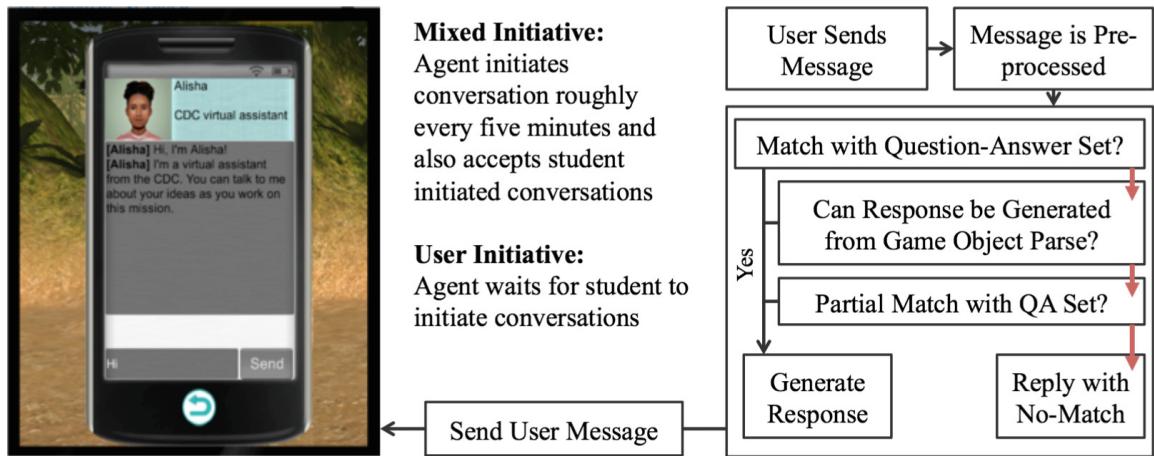
## 1 Introduction

Pedagogical agents have shown great promise for supporting learning in a wide range of domains including literacy [9], mathematics [10], and science [3]. Recent years have seen advances in virtual agents that are capable of conducting multi-party dialogues [2], generating and understanding emotion [4, 5], and producing and interpreting body language [1]. However, previous work has not considered the effects of dialogue initiative policy for pedagogical agents in environments where the conversation is not the central activity. Initiative policy plays a crucial role in defining a pedagogical agent's interaction with students. As defined by Jurafsky and colleagues, the participant who controls the flow of a conversation (through actions such as seeking information or changing the topic) has the initiative [7]. Dialogue systems typically use one of three policies for handling initiative: system initiative, user initiative, or mixed initiative. A system-initiative

policy gives the system the responsibility for controlling and directing the conversation, whereas user-initiative systems support a conversation that the user directs and controls. A mixed-initiative policy combines these approaches: the user can control the topic or direction of the conversation, while the system is responsible for clarifying and asking questions to advance the conversation or complete a task.

## 2 Conversational Pedagogical Agent

The pedagogical agent that is the focus of this paper (Fig. 1) is accessible to students at any time during their gameplay through an in-game smartphone interface. The game-based learning environment, Crystal Island, is an open world with many possible paths for students to take while completing the game. However, it is essential that students explore the game world and gather information, forming and testing hypotheses as they progress. More details about Crystal Island and the agent implementation can be found in prior work [11]. The pedagogical agent, Alisha, plays the role of a virtual assistant from the Center for Disease Control (CDC), the United States' health protection agency. Before describing the architecture of the agent, we first review the context into which she is integrated.



**Fig. 1.** Pedagogical agent's dialogue system design (The thick red arrows represent the flow if the condition is not met). (Color figure online)

We developed two versions of the pedagogical agent, a Mixed Initiative version in which the agent starts a conversation with the student every five minutes during their gameplay session, and a User Initiative version in which the agent never initiates the conversation. The user can initiate conversation at any time in both conditions, and the agent and user can communicate with one another regardless of the player's location in the physical space of the game world.

### 3 User Study and Results

This study took place at a large land-grant university in the southeastern United States. Students were recruited from two introductory computer science courses offered by the college in which they would receive extra credit for participating in a research study. Of the 67 students, 34 were assigned to the Mixed Initiative condition and 33 to the User Initiative condition.

After an hour of gameplay, surveys were used to assess self-reported engagement with the game [8], student experience with the pedagogical agent, and overall student affective experience [6]. A content knowledge post-test (identical to the pre-test) was administered upon the completion of the gameplay session.

**Table 1.** Students' gameplay differences (\* :  $p < 0.05$ ; \*\* :  $p < 0.01$ ).

		Mixed init.	User init.	p-value
Agent interaction	Student utterances	<b>21.59</b>	10.48	0.0004**
	Student words	<b>81.73</b>	48.12	0.0039**
Gameplay	NPC conversations	<b>88.7</b>	64.3	<0.0001**
	Number of books read	8.62	7.09	0.0683
	Book questions missed	4.97	<b>6.75</b>	0.0344*
Student outcomes	Normalized learning gain	0.26	0.21	0.4904
	User engagement	<b>3.97</b>	3.59	0.0189*
	Frustration	26.6	30.0	0.5408

As shown in Table 1, the Mixed Initiative condition's students have significantly more conversations with students interacted with non-player characters (NPCs), a valuable source of information, but there are no significant differences in the number of books read or tests for contaminated objects. However, students missed significantly fewer questions in the embedded assessments given in the Mixed Initiative condition.

After an hour of gameplay, the students completed the post-test and surveys. We hypothesized that there would be differences in the cognitive and affective outcomes of the sessions because of the differences in the conditions. Table 1 displays the differences between the normalized learning gain, user engagement, and frustration that the students experienced during their gameplay. The students in the Mixed Initiative condition had higher engagement with no significant differences in learning or frustration scores.

### 4 Discussion and Conclusion

There are significant differences in dialogue, gameplay, and outcomes across the two dialogue conditions. First, we observed more user utterances and more total

words typed by students in the Mixed Initiative condition. This result is perhaps an expected artifact of the design difference in dialogue policy, since in the Mixed Initiative condition the agent initiated more conversations ( $\mu = 33.06$  versus  $\mu = 14.12$ ), and we would expect to see students respond accordingly. As for differences in gameplay, NPCs more frequently in the Mixed Initiative condition. This is likely because the agent, upon reaching out to the student, would often advise students to seek out NPCs who have essential information for the learning task. The significant difference observed in NPC interactions suggests that students took the pedagogical agent's advice even if they had not specifically solicited it. Another gameplay difference observed between conditions is that students missed fewer questions on in-game reading assessments in the Mixed Initiative condition. It is possible that while interacting more with NPCs, students gained additional content knowledge needed to succeed on the in-game assessments. Rather than just reading the content in the books, the content was also reinforced by the NPCs. Another possibility is that in the Mixed Initiative condition, students were more aware of the pedagogical agents' presence, which may have led to an increased feeling of accountability on the reading tasks and assessments.

Finally, we observed significantly increased self-reported engagement in the Mixed Initiative condition. This increased end-of-game engagement is a promising benefit of the Mixed Initiative condition, as we did not see a significant trade-off with learning gains or increased frustration. This increased engagement may also be a reason for higher conversation levels and interaction with NPCs in the Mixed Initiative condition. The frustration scale and user engagement survey (UES) both include items that measure perceived cognitive load, and the results point to no significant increase in load for the Mixed Initiative policy. We believe that when the agent was taking the initiative, students valued the agent's input more highly and followed the advice more promptly. The Mixed Initiative condition removes some burden from students, providing help incrementally and potentially redirecting disengaged students back onto a productive track, resulting in a greater sense of engagement with the system.

Pedagogical agents hold significant promise for supporting learning and affective outcomes, especially in open learning environments in which students are determining their trajectories through the experience. However, pedagogical agents can become distractions in complex learning environments with learning goals beyond the student-agent interaction. A critical component in facilitating effective agent-student interactions lies in how the agent initiates conversation with the student. In this paper, we reported on a study that investigated the effects of pedagogical agents using different initiative policies in game-based learning. We found that when pedagogical agents utilized a mixed initiative policy, in which both the student and the agent could initiate conversations, the interaction promoted not only more conversation, but also yielded productive in-game behaviors and increased user engagement without increased frustration.

**Acknowledgments.** This research was funded by the National Science Foundation under grants DRL-1721160 and IIS-1409639. Any opinions, findings, and conclusions

or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

1. Abdullah, A., et al.: Pedagogical agents to support embodied, discovery-based learning. In: Intelligent Virtual Agents. LNCS (LNAI), vol. 10498, pp. 1–14. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-67401-8\\_1](https://doi.org/10.1007/978-3-319-67401-8_1)
2. Al Moubayed, S., Lehman, J.: Regulating turn-taking in multi-child spoken interaction. In: Brinkman, W.-P., Broekens, J., Heylen, D. (eds.) IVA 2015. LNCS (LNAI), vol. 9238, pp. 363–374. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-21996-7\\_40](https://doi.org/10.1007/978-3-319-21996-7_40)
3. Borjigin, A., Miao, C., Lim, S.F., Li, S., Shen, Z.: Teachable agents with intrinsic motivation. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 34–43. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_4](https://doi.org/10.1007/978-3-319-19773-9_4)
4. D'mello, S., Graesser, A.: Autotutor and affective AutoTutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. (TiiS) **2**(4), 23 (2012)
5. Girard, S., et al.: Defining the behavior of an affective learning companion in the affective meta-tutor project. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 21–30. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_3](https://doi.org/10.1007/978-3-642-39112-5_3)
6. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988)
7. Jurafsky, D., Martin, J.: Dialog systems and chatbots. In: Speech and Language Processing (2017)
8. O'Brien, H.L., Toms, E.G.: The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inf. Sci. Technol. **61**(1), 50–69 (2010)
9. Panaite, M., et al.: Bring it on! Challenges encountered while building a comprehensive tutoring system using *ReaderBench*. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 409–419. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_30](https://doi.org/10.1007/978-3-319-93843-1_30)
10. Ternblad, E.M., Haake, M., Anderberg, E., Gulz, A.: Do preschoolers ‘Game the System’? A case study of children’s intelligent (mis)use of a teachable agent based play-&-learn game in mathematics. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_41](https://doi.org/10.1007/978-3-319-93843-1_41)
11. Wiggins, J.B., et al.: User affect and no-match dialogue scenarios: an analysis of facial expression. In: Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction, pp. 6–14. ACM (2018)



# Investigating on Discussion for Sharing Understanding by Using Reciprocal Kit-Build Concept Map

Warunya Wunnasri<sup>(✉)</sup>, Jaruwat Pailai, Yusuke Hayashi,  
and Tsukasa Hirashima

Graduate School of Engineering, Hiroshima University, Higashihiroshima, Japan  
{warunya, jaruwat, hayashi,  
tsukasa}@lel.hiroshima-u.ac.jp

**Abstract.** Kit-Build concept map (KB map) is an automatic framework based on the concept map, which is utilized for sharing understanding in pair discussion as Reciprocal Kit-Build concept map (RKB map). In the preliminary experiment, the participants from RKB group can recognize their partner's understanding significantly better than the participants who used the traditional concept map (TCM group) and shows the advantages over the traditional concept map in sharing understanding activity. Therefore, the evidences during discussion will be deeply investigated in this paper to examine the cause of these advantages, especially the relationship between type of talk and the changed propositions.

**Keywords:** Collaborative learning · Shared understanding · Pair discussion · Exploratory talk · Kit-Build concept map

## 1 Introduction

The Kit-Build concept map (KB map) is a framework to realize automatic concept map assessment [1, 2]. It can be utilized as a formative assessment tool for supporting teachers in designing feedback in their class effectively [3]. Consequently, KB map is applied to use as a collaborative learning task for sharing understanding in a pair discussion, is called Reciprocal Kit-Build concept map (RKB map) [4, 5].

In this study, the in detailed analysis is focused to investigate the effects of RKB map through the discussion and concept map of the participants. The factors which support the participants to recognize their partner's understandings were examined through the change of their propositions. Moreover, the characteristic of their discussion is also analyzed to confirm that the exploratory talk is an effective type of talk when the participants want to share their understanding to each other.

## 2 Reciprocal Kit-Build Concept Map

Following the concept maps and the collaborative learning, RKB map is designed for facilitating the collaborators to share their understanding. In RKB map, participants summarize their understanding in the form of the concept map at first. Then, their maps are decomposed to generate kits. The kit of a participant is provided for another participant (the partner), and then, the partner is requested to reconstruct a map by using the kit. Next, two comparison maps are generated by the overlaying between the original map and the reconstructed map then the participants have to share their understanding through the collaborating technique, discussion. The participants are promoted to discuss their same/different understanding based on the two comparison maps that are provided by the RKB map system.

The results of preliminary experiment was published in AIED 2018 [4] firstly. 78 university students were divided into two groups randomly, which contain 20 pairs for RKB group and 18 pairs for TCM group. All participants were requested to construct the concept map three times. The first correspond to the “Comprehension map”, which represents their understanding before discussion. The second was the “Revised map”, which represents their understanding after discussion. The last map was the “Inference map”, which was constructed following the understanding gained from their partner. These three maps were paired and were scored manually by the relational scoring method [6], which evaluates the concept map in propositional level and realize on the meaning of proposition as a high priority. The results show the participants in the RKB group constructed their “Inference map” to be the same as the “Revised map” of their partner more effective than TCM group (RKB:AVG = 61.15, S.D. = 22.16, TCM:AVG = 46.57, S.D. = 29.52), with a statistically significant difference ( $p$ -value < 0.05). These results can explain that RKB map can encourage the participants to recognize their partner’s understanding better than the concept map. This ability will be a strong advantage for the next step of creating collaborative knowledge, as partners that can understand each other can generate collaborative knowledge productively.

In discussion phase, the participants in RKB group could find the same- and different- understanding by using the comparison maps while the participants in TCM group were required to find such different parts from their traditional concept maps by themselves. Most of the participants from the TCM group just read their concept map, asked a few questions then finished the discussion so most of their talks are the non-contributed discussion talk. On the other hand, RKB map requests the participants to reconstruct the kit of their partner, so they have to think deeply about their partner’s understanding. Even if they cannot connect their partner’s kit well, they can ask questions of their partner during the discussion so they created the exploratory talks more than the participants from TCM group obviously.

### 3 Methodology of in Detailed Analysis

The assumption of this study aims to confirm that the activities of RKB map encourage the participants to create exploratory talk, so they will give their opinion, reason, or understanding to each other in their discussion [7–10]. This may be the cause of the advantages of RKB map over the traditional concept map for sharing understanding.

**Categorizing the proposition based on the type of talk.** The talk in discussion phase will be categorized to be exploratory talk, cumulative talk, disputative talk, and non-contributed discussion. To describe the characteristic of each type of talk, the criteria is shown in Table 1.

**Table 1.** The criteria of type of talk's categorization

Type of Talk	Was mentioned	Get response from partner	Consider only their own understanding	Give/answer reasons of understanding
Exploratory talk	✓	✓	✗	✗
Cumulative talk	✓	✓	✗	✗
Disputative talk	✓	✓	✓	✓/✗
Non-contributed discussion	✓	✗	✗	✗

**Counting the Actions on Proposition.** Firstly, the propositions on the “Comprehension map” are compared with the propositions on the “Revised map”. The proposition which is constructed by the same pair of concepts will be investigated the actions on proposition which contain (1) “Not change” action, it means to the propositions which occur in both of the “Comprehension map” and the “Revised map”. It means the participant still keeps their understanding same with before discussion. (2) “Disappear” action refers to the propositions which the participants constructed in the “Comprehension map” but it did not occur in the “Revised map”. It means the participants decide to omit their previous understanding after discussion. (3) “New linking words” action occurs when the participants keep the same connection but change the linking word (label of connection). (4) the proposition which is firstly created in the “Revised map” is called “New proposition” action. The participants got new understanding after discussion then expressed this new understanding in the “Revised map”.

To count the changed propositions same or different from the partner, the proposition which is constructed by the same pair of concepts on participant’s map and their partner’s map will be compared. If their same connected propositions use the same linking word or the synonym word, those propositions will be accepted as the same proposition. In case of “Disappear” action, if there are no the deleted proposition on their partner’s map, this case also will be counted as the same proposition.

## 4 Discussion

To explain the reason why the participants from RKB group can recognize the understanding of their partner better than the participants from TCM group, the differences between the “Comprehension map” and the “Revised map” are focused firstly. The propositions which are categorized as “Change linking word” and “New proposition” actions are counted combinedly as the changed propositions then grouped them based on each type of talk. Table 2 represents the changed and not changed propositions from the “Comprehension map” of RKB group which were recognized by their partner. From this table, we can interpret that the changed propositions, which were mentioned by the contributed discussion especially exploratory talk, can be recognized by their partner well. The comparison between the “Revised map” of their partner and their “Inference map” can be used to confirm the propositions which the participants can recognize. When the participants constructed their “Comprehension map” based on their understanding from the article, they may change their understanding during discussion then express their changing in their “Revised map.” In this case, we assume that if their make their discussion well, they should recognize their partner’s understanding after discussion. The contributed discussion especially exploratory talk encourages the participants to express their understanding through the discussion better than the non-contributed discussion, which their partner can recognize not change proposition 62% and change proposition only 32%.

**Table 2.** Changed propositions which can be recognized by their partner based on type of talk

Type of talk	Action	#props	#props which can be recognized by their partner	Type of Talk	#props	#props which can be recognized by their partner
Contributed discussion	Not change	53	44 (83%)	Exploratory	39	31 (79%)
				Cumulative	8	8 (100%)
				Disputative	6	5 (83%)
	Change	81	32 (40%)	Exploratory	64	29 (45%)
				Cumulative	9	1 (11%)
				Disputative	8	2 (25%)

## 5 Conclusion and Future Work

In this study, the conversation and the concept map of participants which the expressed their understanding are investigated as a deep analysis to examine the cause of RKB map’s advantages. The procedure to categorize the types of talk is specified concretely and the rules for counting the action on propositions are created. The results show that when the participants gave the exploratory talk in their discussion, they gave the reason for their action so their partner can recognize their understanding obviously. The

discussion through the comparison map of RKB map encourages the participants to give a lot of exploratory talk so the participants can share their understanding better than the traditional concept map and can be one of the alternative methods for sharing understanding activity.

**Acknowledgement.** This work was supported by Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP), “Big-data and AI-enabled Cyberspace Technologies” (funding agency: New Energy and Industrial Technology Development Organization).

## References

1. Hirashima, T., Yamasaki, K., Fukuda, H., Funaoi, H.: Kit-Build concept map for automatic diagnosis. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 466–468. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21869-9\\_71](https://doi.org/10.1007/978-3-642-21869-9_71)
2. Hirashima, T., Yamasaki, K., Fukuda, H., Funaoi, H.: Framework of Kit-Build concept map for automatic diagnosis and its preliminary use. Res. Pract. Technol. Enhanced Learn. **10**(1), 1–21 (2015)
3. Pailai, J., Wunnasri, W., Yoshida, K., Hayashi, Y., Hirashima, T.: The practical use of Kit-Build concept map on formative assessment. Res. Pract. Technol. Enhanced Learn. **20**(12), 1–23 (2017)
4. Wunnasri, W., Pailai, J., Hayashi, Y., Hirashima, T.: Reciprocal Kit-Building of concept map to share each other’s understanding as preparation for collaboration. Int. Proc. Artif. Intell. Educ. **2018**, 599–612 (2018)
5. Wunnasri, W., Pailai, J., Hayashi, Y., Hirashima, T.: Reciprocal Kit-Build concept map: an approach for encouraging pair discussion to share each other’s understanding. IEICE Trans. Inf. Syst. **E101**(9), 2356–2367 (2018)
6. McClure, J.R., Sonak, B., Suen, H.K.: Concept map assessment of classroom learning: reliability, validity, and logistical practicality. J. Res. Sci. Teach. **36**(4), 475–492 (1999)
7. Mercer, N.: The quality of talk in children’s collaborative activity in the classroom. Learn. Instr. **6**(4), 359–377 (1996)
8. Barnes, M.: Cumulative and exploratory talk in a collaborative learning classroom. In: 22nd Proceeding on the Mathematics Education Research Group of Australasia, pp. 53–59 (1999)
9. Mercer, N., Dawes, L.: The value of exploratory talk. In: Mercer, N., Hodgkinson, S. (eds.) Exploring Talk in School: Inspired by the Work of Douglas Barnes, pp. 55–72. SAGE Publications Ltd., London (2008). <https://doi.org/10.4135/9781446279526.n4>. <https://www.amazon.com/Exploring-Talk-School-Inspired-Douglas/dp/1847873790>
10. Knight, S., Mercer, N.: The role of exploratory talk in classroom search engine tasks. Technol. Pedagogy Educ. **24**(3), 303–319 (2015)

# **Doctoral Consortium**



# Detection of Collaboration: Relationship Between Log and Speech-Based Classification

Sree Aurovindh Viswanathan<sup>(✉)</sup> and Kurt Vanlehn<sup>(✉)</sup>

Arizona State University, Arizona, USA  
[{sviswalo,kvanlehn}@asu.edu](mailto:{sviswalo,kvanlehn}@asu.edu)

**Abstract.** Research in the field of collaboration shows that students do not spontaneously collaborate with each other. A system that can measure collaboration in real time could be useful by, for example, helping the teacher locate a group requiring guidance. To address this challenge, my research focuses on building and comparing collaboration detectors for different types of classroom problem solving activities, such as card sorting and hand writing. I am also studying transfer: how collaboration detectors for one task can be used with a new task. Finally, we attempt to build a teachers dashboard that can describe reasoning behind the triggered alerts thereby helping the teachers with insights to aid the collaborative activity. Data for building such detectors were collected in the form of verbal interaction and user action logs from students' tablets. Three qualitative levels of interactivity was distinguished: Collaboration, Cooperation and Asymmetric Contribution. Machine learning was used to induce a classifier that can assign a code for every episode based on the set of features. Our preliminary results indicate that machine learned classifiers were reliable.

**Keywords:** Collaborative learning · Machine learning · Learning analytics

## 1 Introduction and Problem Statement

Collaboration is a 21<sup>st</sup> century skill as well as an effective method for learning [1, 2]. However, collaboration between students is not spontaneous and acquiring collaboration skills is not straightforward. Several theoretical frameworks of collaboration [3–5] connect variations of social interactions to effectiveness of learning. Various dimensions of effective collaboration have been identified in the literature [6, 7]. Transactivity has been identified as one of the important characteristics of collaboration grounded in frameworks of Piaget [8] and Vygotsky [9], and it has shown to facilitate acquiring domain knowledge [1]. Chi's ICAP framework [5] includes transactive process in its category Interactive. Of the four categories of overt behavior, Interactive process foster the most learning.

Many projects have worked on the challenge of automating the analysis of interaction among group members. These antecedents will be briefly reviewed by defining two dimensions, *purpose* and *input*, then describing the few systems whose position along these two dimension match the position of the project reported here. The two dimensions are excerpted from several similar multi-dimensional reviews [10, 11].

When a large number of projects could be cited as illustrations of a dimension, only those published most recently will be cited.

The first dimension concerns the purpose or function of the collaboration measure. That is, what does the system do with the output of the collaboration detector? This dimension has the following categories: Clustering, Classification, Mirroring, Metacognitive, Guiding, Orchestration and Restructuring. Our project fits into two of the categories: *Classification* and *Orchestration*. Projects in classification category [12, 13] used human judges to code group interactions into a variety of collaboration categories, then used supervised machine learning methods to induce classifiers (also called detectors). The main research questions is: how accurate is the induced detector? The projects in Orchestration categories [14] display the amount of collaboration per group on a dashboard held by the teacher. This allows the teacher to visit groups that need help collaborating. The main research question is whether such collaboration detection is useful to the teacher and effective at increasing collaboration in the classroom.

The second dimension classifies prior work by input to the detector. All collaboration detection projects so far have students work in a shared workspace, so their detectors take the users' interactions (log data) as one input. Most projects also analyzed some form of communication among group members. The communication input can be classified as:

- Group members communicated in a formal language [15].
- Group members used a small set of buttons to express agreement/disagreement [16].
- Group members communicated by typing natural language and classifying their contribution using a menu of sentence openers or speech acts. Some systems ignored the text and used *only* the students' classifications of their text [17].
- Group members communicated via typing (chat), with or without sentence openers. The text was analyzed by human “wizards” [18], keywords [19, 20] or machine-learned text classifiers [21].
- Group members conversed in unconstrained speech, recorded by individual microphones [12, 20, 22, 23].

The design of our classification codes matches that of Chi's ICAP Framework. My thesis project falls into the *Classification* category of the purpose dimension and *unconstrained speech* in the input dimension. In addition, I am developing collaboration detectors that can generalize across different tasks. Finally, I use the collaboration codes generated by the system and the underlying data to populate a dashboard that not only shows teachers which groups are not collaborating but also explains what evidence supports its assessment.

## 2 Methodology and Progress

The overall dissertation work focuses on building collaboration detectors that measure the quality of collaboration in real time. Laboratory studies were conducted with more than sixty pairs of students working on two different types of tasks. In order to create and evaluate collaboration detectors, the judgments of human coders were used as the ‘gold standard’ classification of the group’s interactions. The coders had both high

quality audio and several videos to aid their judgment. Collaboration detectors were then machine-learned from the human judgments. Their accuracies were measured using 10 fold cross validation.

My thesis project is divided into the tasks briefly described below:

1. The first task involved students collaboratively working on a card-moving task which required interpreting time-distance graphs. Machine learned detectors were built by using speech and log data to measure collaboration [24]. The results were promising with a high level of agreement. However, it has to be noted that the particular task made it relatively easy to measure collaboration. (Complete)
2. The second task involved students working on a collaborative task where they were required to analyze solutions of four hypothetical students. They had to write paragraph long explanations. An in-depth analysis of video tapes and logs of tablets were performed to understand how students write on the surface of the tablets. It also highlighted the fact that superficial measures of collaboration may not be adequately useful for detection of collaboration in hand writing settings. (Complete)
3. The third task involved determining whether collaboration detection could be accurate when student voices are converted to a privacy-preserving binary signal (1 = speaking, 0 = silence) before being transmitted and stored. Data were collected as students wrote paragraphs together and solved problems. A speech signal was processed at a microphone by voice activity detector to produce the binary signal. The results indicate that binary based collaboration detectors yielded only slightly less accuracy than detectors that took the high quality audio signal as input. (Complete)
4. Whereas task 1 above showed that a log-based collaboration detector was just as accurate as speech-based collaboration detection, the card-moving task made such detection easy. This fourth task investigated log-based collaboration detection with a more common task, collaborative writing. Data came from students who analyzed mistaken problem solutions done by four hypothetical students. The students then wrote an analysis of each solution. The results indicate that log-based collaboration detection accuracy was low to moderate for this collaborative writing task. Comparing the features of the collaborative writing task to the card-moving task allows speculations on what task properties facilitate log-based collaboration detection. (Complete)
5. The fifth task will involve creating a general collaboration detector that will function well with multiple collaborative tasks. Features would be extracted from acoustic and prosodic characteristics of audio signal along with its time series characterization. If a generalized collaboration detector is reliable, then it could be used in various tasks to measure collaboration. This would help the researchers avoid the laborious work of annotating the video/audio files manually to understand the process of collaboration. (In progress)
6. Finally, in collaboration with a larger group of students, I am attempting to create a visualization dashboard that will provide insights to teachers about the collaboration based on speech and actions in collaborative group activity. It will also provide the teacher with suggestions for improving the collaboration of specific groups. (in progress).

### 3 Contributions and Impact

The thesis explores methods to automatically measure the types collaboration exhibited by students working together on learning activities. Collaboration detectors are based on building machine learning models of log and/or speech data. Firstly, this work complements research in collaborative learning environments with a goal to classify collaborative activity in MOOCs and other environments where students communicate in text. Secondly, if task-general classification of spoken collaboration is successful, it would reduce the laborious process of human coding required to establish reliability and would potentially allow the researchers to build various systems that utilize the underlying categories of collaboration. Finally, the proposed dashboard would provide insights into student's speech and actions with a goal of reducing teachers' cognitive overload and provides teachers with information to facilitate the classroom.

**Acknowledgements.** This research was funded by the Diane and Gary Tooker chair for effective education in Science Technology Engineering and Math, by NSF grant IIS-1628782, and by the Bill and Melinda Gates Foundation under Grant OP1061281.

### References

1. Chen, J., Wang, M., Kirschner, P.A., Tsai, C.-C.: The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: a meta-analysis. *Rev. Educ. Res.* **88**(6), 799–843 (2018)
2. Vogel, F., Wecker, C., Kollar, I., Fischer, F.: Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. *Educ. Psychol. Rev.* **29**(3), 477–511 (2017)
3. Stahl, G.: Theories of cognition in collaborative learning. In: Hmelo-Silver, C., O'Donnell, A., Chan, C., Chinn, C. (eds.) *The International Handbook of Collaborative Learning*, pp. 74–90. Taylor & Francis, New York (2013)
4. Hartmann, C., Angersbach, J., Rummel, N.: Social Interaction, Constructivism and their Application within (CS) CL Theories. International Society of the Learning Sciences, Inc. [ISLS] (2015)
5. Chi, M.T., Wylie, R.: ICAP: a hypothesis of differentiated learning effectiveness for four modes of engagement activities. *Educ. Psychol.* **49**(4), 219–243 (2014)
6. Kahrimanis, G., et al.: Assessing collaboration quality in synchronous CSCL problem-solving activities: adaptation and empirical evaluation of a rating scheme. In: Cress, U., Dimitrova, V., Specht, M. (eds.) *EC-TEL 2009. LNCS*, vol. 5794, pp. 267–272. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04636-0\\_25](https://doi.org/10.1007/978-3-642-04636-0_25)
7. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. *Int. J. Comput. Support. Collab. Learn.* **2**(1), 63 (2007)
8. Berkowitz, M.W., Gibbs, J.C.: Measuring the developmental features of moral discussion (1982)
9. Schwartz, D.L.: The productive agency that drives collaborative learning. In: Dillenbourg, P. (ed.) *Collaborative Learning: Cognitive and Computational Approaches*, pp. 197–218. Elsevier Science, New York (1999)
10. Soller, A., Martinez, A., Jermann, P., Muehlenbrock, M.: From mirroring to guiding: a review of state of the art technology for supporting collaborative learning. *Int. J. Artif. Intell. Educ.* **15**, 261–290 (2005)

11. VanLehn, K.: Regulative loops, step loops and task loops. *Int. J. Artif. Intell. Educ.* **26**(1), 107–112 (2016)
12. Martinez-Maldonado, R., Kay, J., Yacef, K.: An automatic approach for mining patterns of collaboration around an interactive tabletop. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNNAI), vol. 7926, pp. 101–110. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_11](https://doi.org/10.1007/978-3-642-39112-5_11)
13. Gweon, G., Agarwal, P., Raj, B., Rose, C.P.: The automatic assessment of knowledge integration processes in project teams. In: Proceedings of Computer Supported Collaborative Learning (2011)
14. Martinez-Maldonado, R., Clayphan, A., Yacef, K., Kay, J.: MTFeedback: providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Trans. Learn. Technol.* **8**, 187–200 (2014)
15. Tedesco, P.A.: MArCo: building an artificial conflict mediator to support group planning interactions. *Int. J. Artif. Intell. Educ.* **13**(1), 117–155 (2003)
16. de los Angeles Constantino-Gonzalez, M., Suthers, D.D., de los Santos, J.G.E.: Coaching web-based collaborative learning based on problem solution differences and participation. *Int. J. Artif. Intell. Educ.* **13**(2–4), 263–299 (2003)
17. Baghaei, N., Mitrovic, A., Irwin, W.: Supporting collaborative learning and problem-solving in a constraint-based CSCL environment for UML class diagrams. *Comput. Support. Collab. Learn.* **2**, 159–190 (2007)
18. Tsovaltzi, D., Rummel, N., McLaren, B., Pinkwart, N., Scheuer, O., Harrer, A., Braun, I.: Extending a virtual chemistry laboratory with a collaboration script to promote conceptual learning. *Int. J. Technol. Enhanc. Learn.* **2**(1/2), 91–110 (2010)
19. Dragon, T., Floryan, M., Woolf, B., Murray, T.: Recognizing dialogue content in student collaborative conversation. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6095, pp. 113–122. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13437-1\\_12](https://doi.org/10.1007/978-3-642-13437-1_12)
20. Martinez-Maldonado, R., Yacef, K., Kay, J.: TSCL: a conceptual model to inform understanding of collaborative learning processes at interactive tabletops. *Int. J. Hum Comput Stud.* **83**, 62–82 (2015)
21. Walker, E., Rummel, N., Koedinger, K.R.: Adaptive intelligent support to improve peer tutoring in algebra. *Int. J. Artif. Intell. Educ.* **24**(1), 33–61 (2014)
22. Gweon, G., Jain, M., McDonough, J., Raj, B., Rose, C.P.: Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *Int. J. Comput. Support. Collab. Learn.* **8**(2), 245–265 (2013)
23. Bassiou, N., et al.: Privacy-preserving speech analytics for automatic assessment of student collaboration. In: Privacy-Preserving Speech Analytics for Automatic Assessment of Student Collaboration, pp. 888–892 (2016)
24. Viswanathan, S.A., VanLehn, K.: Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Trans. Learn. Technol.* **11**, 230–242 (2018)



# An Intelligent Tutoring System and Teacher Dashboard to Support Mathematizing During Science Inquiry

Rachel Dickler<sup>(✉)</sup>

Rutgers University, New Brunswick, NJ 08901, USA  
rachel.dickler@gse.rutgers.edu

**Abstract.** Using mathematics is critical to science inquiry at the high school level and is predictive of students' later success in STEM college majors and careers. Inq-ITS (Inquiry Intelligent Tutoring System) has recently added mathematizing functionalities in order to support students in the mathematical practices needed for scientific inquiry, and our teacher alerting dashboard, Inq-Blotter, is being extended to alert teachers in real-time to students' difficulties with this practice. Mathematizing in science can be challenging as students must attend to multiple sources of information (i.e., graphs, data tables), as well as do graphing and modeling. In the present paper, I describe three studies on the use of Inq-Blotter to support students on mathematizing in which I: (1) explore students' eye-movements and think-aloud protocols while mathematizing in Inq-ITS to identify the proportion of mathematizing difficulties that are related to knowledge acquisition processes versus other mathematical competencies (e.g., graph building and modeling), (2) examine if alerts within Inq-Blotter permit teachers to identify students who need help most (relative to teachers without access to alerts), and (3) identify whether teacher support based on alerts leads to improvements on students' next opportunity to engage in mathematizing and examine the corresponding teacher discourse associated with students' gains. These studies will indicate how to support students on mathematizing with intelligent technologies.

**Keywords:** Intelligent tutoring · Dashboard · Science inquiry · Mathematizing

## 1 Introduction

In order to pursue and fulfill future careers in STEM [1], students must master the practices of scientists as emphasized in major policy documents such as the United States' Next Generation Science Standards (NGSS; [2]). A central inquiry practices of scientists is using mathematics and computational thinking. In the present paper, I focus on the component of using mathematics (i.e., mathematizing) because it is deeply intertwined with science [2, 3], can be augmented by computational tools (c.f. [4]), and is fundamental to students' performance in high school science [5–9], STEM college courses [10–12] and STEM careers [2, 13]. While critical, mathematizing is also extremely difficult for students [3, 7, 8] and a barrier to science particularly when taking into account gender, race, and socioeconomic factors [14].

One solution to support students in mathematizing is through the implementation of intelligent tutoring systems. Inq-ITS (Inquiry Intelligent Tutoring System; [15, 16]) has recently added mathematizing functionalities needed for scientific inquiry. While there are other ITSs for mathematics, these systems do not support students on science inquiry practices (cf., [17]). Our teacher alerting dashboard, Inq-Blotter, thus is being extended to alert teachers to students' difficulties with this practice. Inq-Blotter provides actionable alerts so that teachers can scaffold students in real time as students complete investigations in Inq-ITS [15]. While there are several dashboards for online environments including intelligent tutoring systems for mathematics [18, 19], only a few dashboards exist for science [15, 20–22]. Of the dashboards available for science, Inq-Blotter [15] is the only dashboard that currently alerts on students' science inquiry practices and their respective sub-components in real-time. Prior work shows that the use of Inq-Blotter leads to improvement in student performance on the difficult practice of analyzing and interpreting data [23, 24]. Additionally, students improved on inquiry practices (i.e., asking questions, carrying out investigations, analyzing and interpreting data) on their next opportunity after receiving help from their teacher based on Inq-Blotter alerts [25]. These findings are promising in terms of the potential for Inq-Blotter alerts in guiding teacher support on the practice of mathematizing. In the following sections, I present three studies that form the basis of my dissertation on the use of Inq-ITS and Inq-Blotter to support students on the practice of mathematizing.

## 2 Study 1: Competencies Underlying Mathematizing in Inq-ITS

In the first study, students' eye-movements are examined as they engage in the mathematizing stages of the Inq-ITS Ramp with Graphing Virtual Lab. Students need support when acquiring knowledge from visual information sources in science because all information is presented simultaneously [26]. In a pilot study [27], I examined students' eye-movements in Inq-ITS and revealed that students who performed lower on the practice of analyzing and interpreting data did not attend to the data they collected or to graphs of their data. In the present study, the goal is to use eye-tracking and think-aloud protocols as students' use Inq-ITS to separate out the aspects of mathematizing that are difficult due to problems with knowledge acquisition.

### 2.1 Methods

**Participants and Materials.** 20–30 high school students will participate in the study outside of school during the summer of 2019. Students will complete the Inq-ITS Ramp with Graphing Virtual Lab with Tobii portable eye-trackers on their desktop computers while thinking aloud [28]. There are three Ramp Lab activities that each include stages of: asking questions, carrying out investigations, graphing, modeling, analyzing and interpreting data, and explaining findings. For the purposes of this study, I am focusing on the stages of Graphing (i.e., students build a graph from scratch by selecting axes and relevant data points) and Modeling (i.e., students develop and examine the fit of a mathematical model in the form of an equation to their graphs) in each activity.

**Measures.** Students' performance on mathematizing within Inq-ITS is captured at the sub-practice level based on patented educational data mined algorithms [15, 16, 29]. The sub-practices for mathematizing in science include: constructing graphs, translating trends in graphs into equations, recognizing and using appropriate units, and identifying variables and constants in mathematical models. The Tobii eye-tracker (EyeX model) device captures eye-movement data that is stored within xml files with coordinates of fixation points, and software automatically generates data about the location of students' fixations and time spent looking at these specific locations (we have a patent on our technology to direct learners' attention based on their knowledge gaps in science [30]). Since the Graphing and Modeling stages are divided into regions based on the different components on each stage (i.e., data table, graph, equation), eye tracking allows us to identify how students are attending to various sources of information.

**Analyses.** Students' mathematizing performance, eye-movements, and think-aloud data will be triangulated. I will identify students who demonstrated low performance on each sub-component of mathematizing and examine relationships between their eye-movements and verbal protocols, as was done in our pilot study [27]. I will then identify whether students' difficulties with particular mathematizing sub-practices could be explained by their knowledge acquisition processes (i.e., using eye movements; [31]), other competencies related to mathematical understandings (i.e., using think-aloud protocols), or some combination of both. These findings will inform both the competencies underlying mathematizing during inquiry, as well as the design of Inq-Blotter alerts.

### 3 Study 2: Full Inq-Blotter Versus Minimal Inq-Blotter

Based on the results of Study 1, Inq-Blotter [15, 32] alerts will provide information to teachers to support students on the practice of mathematizing. Study 2 focuses on whether the alerts are effective in driving teacher support of students who need help most on the practice of mathematizing. Alerting has the potential to support students in greatest need of support, which is important when taking into account issues of equity related to math and science [14]. Particularly, I am interested in the effectiveness of this alerting dashboard (relative to using a dashboard without alerts) when implemented in schools in areas of varying socio-economic status (SES).

#### 3.1 Methods

**Participants and Materials.** 8 high school teachers (4 teachers from a high school in a high SES area and 4 teachers from a high school in a low SES area) and their students ( $\sim 300$  students) will participate in the study in the Fall of 2019 during their regular science class periods. All students will complete the Inq-ITS Ramp Lab (as in Study 1). Two teachers from each high school will be randomly assigned to a Full Inq-Blotter condition (i.e., access to alerts on student mathematizing and tips on how to support their students) and the other two teachers will be randomly assigned to a Minimal Inq-

Blotter condition (i.e., access to a dashboard with a list of student names that teachers can click-on on when they help a student, but teachers do not receive alerts).

**Measures.** Students' performance on mathematizing will be captured in log files [16, 29]. Both versions of Inq-Blotter will store teacher actions in log files.

**Analyses.** I will triangulate the log files from Full and Minimal Inq-Blotter with evaluations of students' inquiry within Inq-ITS [15]. Within each classroom, I will examine student performance on mathematizing activity-by-activity and identify whether or not the student was helped by the teacher (based on the teacher logs). I will then examine the effectiveness of alerts (Full Inq-Blotter) for helping teachers identify the students who needed help most versus teachers who did not receive alerts (Minimal Blotter) across schools. It will then be important to examine whether teacher support based on alerts for mathematizing improves students' performance.

## 4 Study 3: Effects of Alert-Based Feedback on Performance

In order to determine whether teacher feedback based on alerts improves student performance on the practice of mathematizing, Study 3 examines students' scores on their next opportunity to use the mathematizing practice after being helped. To understand how teacher feedback helped students on mathematizing, I will also analyze recordings of teachers' discourse as they respond to Inq-Blotter alerts [33].

### 4.1 Methods

**Participants and Materials.** 8 high school teachers and their students (~300 students) will participate in the study in the winter of 2019–2020 during their regular science class periods. Students will complete the same Ramp Lab as in the prior studies and all teachers will have access to the full version of Inq-Blotter with alerts on mathematizing. When teachers respond to alerts, they will click a “record” button and all discourse exchanged following the alert will be captured through the teachers' devices.

**Measures.** Inq-ITS and Inq-Blotter log files will be used as in the prior studies. Teachers' voice data will be captured, time-stamped, and stored in log files any time the teachers press the “record” button when they receive an alert for mathematizing.

**Analyses.** Students' Inq-ITS log files will be triangulated with the teachers' Inq-Blotter log files and voice recordings. I will identify the students who completed a second activity in the Ramp Lab (3 activities total) after being helped by the teacher. This will allow us to determine if students improved on the practice of mathematizing on their next opportunity after receiving help from the teacher. I will then examine the teacher discourse following an alert and code for the types of scaffolds that teachers provided to students. For example, in prior pilot work [25] I coded teacher feedback for particular scaffolds in order to ascertain how teacher help, in turn, led to students' improvement.

## 5 Discussion

Overall, the findings from the three studies proposed in the present paper will provide valuable information on the competencies (as well as knowledge acquisition) underlying mathematizing during inquiry and how to best support students on the practice of mathematizing using innovative technologies such as teacher dashboards. Additionally, this work seeks to ensure that intelligent technologies benefit students who are at risk of falling behind in high school, college, and STEM careers due to poor competencies at mathematizing.

## References

1. Bybee, R.W., Fuchs, B.: Preparing the 21st century workforce: a new reform in science and technology education. *J. Res. Sci. Teach.* **43**(4), 349–352 (2006)
2. States, N.G.S.S.L.: Next Generation Science Standards: For States, by States. National Academies Press, Washington (2013)
3. Basson, I.: Physics and mathematics as interrelated fields of thought development using acceleration as an example. *Int. J. Math. Educ. Sci. Technol.* **33**(5), 679–690 (2002)
4. Wilkerson, M., Fenwick, M.: The practice of using mathematics and computational thinking. In: Schwarz, C.V., Passmore, C., Reiser, B.J. (eds.) *Helping Students Make Sense of the World Using Next Generation Science and Engineering Practices*. National Science Teachers' Association Press, Arlington (2016)
5. Hansson, L., Hansson, Ö., Juter, K., Redfors, A.: Reality-theoretical models—mathematics: a ternary perspective on physics lessons in upper-secondary school. *Sci. Educ.* **24**, 615–644 (2015)
6. Hudson, H.T., Rottmann, R.M.: Correlation between performance in physics and prior mathematics knowledge. *J. Res. Sci. Teach.* **18**(4), 291–294 (1981)
7. McDermott, L.C., Rosenquist, M.L., Van Zee, E.H.: Student difficulties in connecting graphs and physics: examples from kinematics. *Am. J. Phys.* **55**(6), 503–513 (1987)
8. Potgieter, M., Harding, A., Engelbrecht, J.: Transfer of algebraic and graphical thinking between mathematics and chemistry. *J. Res. Sci. Teach.* **45**(2), 197–218 (2008)
9. Sadler, P.M., Tai, R.H.: Success in introductory college physics: the role of high school preparation. *Sci. Educ.* **85**(2), 111–136 (2001)
10. Gottfried, M.A., Bozick, R.: Supporting the STEM pipeline: linking applied STEM course-taking in high school to declaring a STEM major in college. *Educ. Financ. Policy* **11**, 177–202 (2016)
11. Hoban, R.A., Finlayson, O.E., Nolan, B.C.: Transfer in chemistry: a study of students' abilities in transferring mathematical knowledge to chemistry. *Int. J. Math. Educ. Sci. Technol.* **44**(1), 14–35 (2013)
12. Sadler, P.M., Tai, R.H.: The two high-school pillars supporting college science. *Science* **317**, 457–458 (2007)
13. Sadler, P.M., Sonnert, G., Hazari, Z., Thi, R.: The role of advanced high school coursework in increasing STEM career interest. *Sci. Educ.* **23**(1), 1 (2014)
14. Quinn, D.M., Cooc, N.: Science achievement gaps by gender and race/ethnicity in elementary and middle school: trends and predictors. *Educ. Res.* **44**(6), 336–346 (2015)
15. Gobert, J., Moussavi, R., Li, H.: Sao Pedro, M., Dickler, R.: Real-time Scaffolding students' on-line data interpretation during inquiry with Inq-ITS. In: Auer, M., Azad, A., Edwards, A.,

- de Jong, T. (eds.) Cyber-Physical Laboratories In Engineering And Science Education. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-76935-6\\_8](https://doi.org/10.1007/978-3-319-76935-6_8)
- 16. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S.: From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *J. Learn. Sci.* **22**, 521–563 (2013)
  - 17. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. *J. Learn. Sci.* **4**(2), 167–207 (1995)
  - 18. Aleven, V., Roll, I., McLaren, B.M., Koedinger, K.R.: Help helps, but only so much: research on help seeking with intelligent tutoring systems. *Int. J. Artif. Intell. Educ.* **26**(1), 205–223 (2016)
  - 19. Xhakaj, F., Aleven, V., McLaren, B.M.: Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 315–329. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_23](https://doi.org/10.1007/978-3-319-66610-5_23)
  - 20. Acosta, A., Slotta, J.D.: CKBiology: an active learning curriculum design for secondary biology. In: *Frontiers in Education*, p. 52. Frontiers (2018)
  - 21. Matuk, C.F., Linn, M.C., Eylon, B.S.: Technology to support teachers using evidence from student work to customize technology-enhanced inquiry units. *Instr. Sci.* **43**(2), 229–257 (2015)
  - 22. Vitale, J., Linn, M., Gerard, L.: Visualizing data from automated scores to help teachers guide inquiry with scientific visualizations in diverse classes. Presented at: 12th Annual International Conference of the Learning Sciences as Part of a Symposium, Real-Time Visualization of Student Activities to Support Classroom Orchestration, Singapore (2016)
  - 23. Sao Pedro, M.A., Gobert, J., Dickler, R.: Can an alerting teacher dashboard improve how teachers help their students learn science inquiry practices? To be Presented at: American Educational Research Association (AERA): Learning and Instruction, Toronto, Canada (2019)
  - 24. Sao Pedro, M., Gobert, J., Dickler, R.: Improving science teaching and learning through rigorous and relevant education technology interventions. Paper to be Presented at: National Association of Research on Science Teaching, Baltimore, MD (2019)
  - 25. Dickler, R., Gobert, J., Sao Pedro, M., Li, H.: Teacher scaffolds mediated by a science inquiry dashboard. Submitted to: *European Association of Research on Science Education*, Bologna, Italy (submitted)
  - 26. Gobert, J.D.: Leveraging technology and cognitive theory on visualization to promote students' science. In: Gilbert, J.K. (ed.) *Visualization in Science Education*, pp. 73–90. Springer, Dordrecht (2005). [https://doi.org/10.1007/1-4020-3613-2\\_6](https://doi.org/10.1007/1-4020-3613-2_6)
  - 27. Dickler, R., Gobert, J., Yasar, O.: Exploring the use of eye-tracking as a method to capture student knowledge acquisition in a virtual science inquiry investigation. Presented at: Rutgers STEM Community Outreach Symposium, New Brunswick, NJ (2018)
  - 28. Van Gog, T., Paas, F., Van Merriënboer, J.J.: Uncovering expertise-related differences in troubleshooting performance: combining eye movement and concurrent verbal protocol data. *Appl. Cogn. Psychol.* **19**(2), 205–221 (2005)
  - 29. Gobert, J.D., Baker, R.S., Sao Pedro, M.A.: Inquiry skills tutoring system. U.S. Patent No. 9,373,082. U.S. Patent and Trademark Office, Washington, DC (2014)
  - 30. Gobert, J., Toto, E.: Instruction system with eyetracking-based adaptive scaffolding. U.S. Patent No. 9,230,022. U.S. Patent and Trademark Office, Washington, DC (2014)
  - 31. Mason, L., Pluchino, P., Tornatora, M.C., Ariasi, N.: An eye-tracking study of learning from science text with concrete and abstract illustrations. *J. Exp. Educ.* **81**(3), 356–384 (2013)

32. Gobert, J., Sao Pedro, M., Betts, C., Baker, R.S. Inquiry skills tutoring system (child patent for alerting system). U.S. Patent No. 9,564,057. U.S. Patent and Trademark Office, Washington, DC (2016)
33. Gobert, J., Sao Pedro, M.: Inq-Blotter: a real time alerting tool to transform teachers' assessment of science inquiry practices (NSF-IIS-1629045). Awarded from the National Science Foundation (2016)



# Towards Adaptive Hour of Code

Tomáš Effenberger<sup>(✉)</sup>

Masaryk University, Brno, Czech Republic  
[tomas.effenberger@mail.muni.cz](mailto:tomas.effenberger@mail.muni.cz)

**Abstract.** Hour of Code activities became a de facto standard for the first encounter with programming, reaching millions of children every year. These activities are typically not personalized and offer the same sequence of tasks to everybody, which leads to too slow pace for some students, while too fast for others. We aim to improve upon the current state of the art in teaching introductory programming by providing insight into how adaptive learning techniques can make the Hour of Code activities more efficient and engaging.

## 1 Introduction

Block-based programming games like Hour of Code are a popular way to introduce children into programming [16]. These activities combine several strategies to support learning and motivation, e.g., block-based programming interface to avoid syntax errors and visualization of program execution in a game world [8].

However, these activities are not personalized—they offer the same sequence of 10–20 tasks to everybody, independently of the prior skills and speed of learning. Some tasks are thus too easy for some students, while too difficult for others, which leads to a suboptimal learning experience and negative emotions like boredom or frustration. Our goal is to improve Hour of Code activities by incorporating adaptive behavior, specifically a personalized task recommendation. By giving students tasks of optimal difficulty, the system helps them to achieve a complete immersion into the problem-solving activity, known as the state of flow [4]. The state of flow supports both the learning and intrinsic motivation by fulfilling needs of effectiveness, progress, and mastery [9].

To explore the possibilities of adaptation in Hour of Code activities, we developed RoboMission, a web application for learning introductory programming. The system contains over 80 tasks, which are a variation on a traditional *robot in a grid* theme, with some novel features to make the block-based programs more compact [6]. To make sure the results generalize to other types of activities (e.g., turtle graphics [3], short textual programming), we will also analyze data from other systems for teaching introductory programming.

## 2 Adaptive Learning of Introductory Programming

Adaptation in learning systems can be performed at different time scales, ranging from an offline adaptation of the system to the entire population of students

(*design loop*), to an online adaptation to an individual student during her practice, either between tasks (*outer loop*), or even after each step (*inner loop*) [1].

The personalized task recommendation happens in the outer loop by asking a tutor model for the next task the student should practice. The tutor model relies on a student model for predicting student's performance on considered tasks, which in turn depends on a performance measure that assess the student's performances on previously attempted tasks. These three models usually share the same underlying structure, which is sometimes referred to as a domain model. For example, the domain model can specify which tasks belong to the same problem set and the other three models have some parameters for each of the problem sets. All these models are iteratively improved in the design loop [2].

## 2.1 Domain Model

**RQ1:** *How to organize tasks for a personalized Hour of Code activity?* A task recommendation algorithm requires a large pool of tasks for a single programming game with a wide range of difficulties. In RoboMission, the tasks are grouped into levels according to the concepts they practice, like sequences of commands, loops, and conditional commands. These levels are further refined into sublevels, which contain tasks of approximately the same difficulty, which simplifies performance measurement and task recommendation. To explore whether linearly ordered hierarchical problem sets (levels and sublevels) are a universally appropriate domain model for all Hour of Code activities, we will analyze data from multiple systems and exercises for teaching introductory programming. We have already done preliminary work on measuring the similarity of programming tasks [14], and analysis of collected performance data from RoboMission [5].

## 2.2 Performance Measure

**RQ2:** *How to measure students' performance on programming tasks?* Most student modeling approaches consider only binary correctness [11]. However, introductory programming tasks usually take several minutes to solve; moreover, nearly all attempts are eventually successful, rendering the binary success too infrequent and weak signal of students' skills. Instead, we propose to use a few discrete performance levels, e.g., failed  $\leq$  poor  $\leq$  good  $\leq$  excellent.

We will investigate whether it is enough to use just summary statistics, such as solving time and number of executions, or whether it is necessary to consider complete time series of students' edits and executions. We will also explore methods for combining multiple summary statistics and setting thresholds. We will use data from several systems to analyze the agreement between different performance measures and compare their measurements against ground truth obtained through manual labeling.

Evaluation of performance measures is complicated by an interaction between the performance measure and the domain model. For example, if a given task has too many performances measured as poor, it may be caused either by too strict performance measure, or by the task being in an inappropriate problem

set. Therefore, we need to develop a method for joint evaluation of domain model and performance measure.

### 2.3 Student Model

**RQ3:** *How to predict a future performance of a student on introductory programming tasks?* The task recommendation algorithm can either use the measured performances directly, or depend on a student model predicting future performances. Having an intermediate student model is useful for visualizing progress towards mastery (by transforming the estimated distribution of performances into a single number). We will adapt traditional student models predicting future success, such as PFA, BKT, and LogisticHMM [11, 12] to the discrete performance and compare them on historical data from multiple systems for introductory programming.

We will investigate utilizing programming concepts and their combination; for example, using hierarchical Bayesian Network for conjunctive knowledge modeling previously used in Java and SQL courses [7]. We hypothesize that if the problem sets are homogeneous with respect to the practiced concepts and task difficulties, then a single-skill model for each of the homogeneous problem sets is enough.

Fair evaluation is complicated by several biases present in the collected data, such as a personalized recommendation of the next task, learning, self-selection bias, and attrition bias [13]. To minimize the impact of these biases, we will incorporate some randomization into data collection in RoboMission, which is described in Sect. 2.5.

### 2.4 Tutor Model for Task Recommendation

**RQ4:** *How to recommend the next task to practice in Hour of Code activities?* Which recommendation algorithm to use, how to optimize its parameters, how to balance exploration and exploitation, and how to evaluate its impact?

We propose the following decomposition of the task recommendation problem: (1) select a problem set, (2) select a task from the problem set, (3) use a mastery criterion to decide when to move to a next problem set. For a short Hour of Code tutorial, it is reasonable to assume that problem sets can be linearized (as described in Sect. 2.1) and that students should not skip a whole problem set without solving at least one task from it. Enforcing homogeneity of problem sets (tasks of the same difficulty, practicing the same concepts) allows to select the task from the problem set uniformly at random, maximizing the exploration.

The tutor model depends on a student model and specifies a transformation from the predicted performance probabilities to a single number (progress towards mastery) and a threshold for the mastery. To evaluate the suitability of the proposed recommendation algorithm and the impact of the parameters, we will use performance measurement as a proxy for correct recommendation—the medium (good) performance indicates an appropriate recommendation (neither

too easy nor too difficult task). The same caveats as for the evaluation of the student models applies (Sect. 2.3), but a further challenge is imposed by the sparsity of the collected data [15]. Thus, to compare a few most promising candidates properly, we will perform a randomized control trial in RoboMission. Section 2.5 describes a proposed data collection method and objective.

## 2.5 Evaluation of Learning

The standard way to evaluate the impact of an intervention on students' learning is via a post-test, which, however, does not fit well in an online learning system. Post-tests can be conducted in in-classroom experiments, but these are costly and may be contrived. On the other hand, data collected by a learning system includes many biases, which complicates a fair comparison of evaluated student and tutor models.

We propose the following method for collection of significantly less biased data, inspired by randomly chosen *reference questions* already used in other domains [10]. Before the last sublevel of each level, the student would solve a *control task*, chosen uniformly at random from all tasks from already mastered levels. Choosing a task from all levels would often lead to a way too difficult task, and long, frustrating experience, while solving a task from previously mastered levels reinforces the skills through interleaved practice. It also fulfills students' need for learning and mastery, giving them occasionally opportunity to solve a task with excellent performance. However, it also introduces a bias into the data collection, which can be further amplified by the imperfections of the domain model. We will perform simulated experiments to explore the potential impact of this bias and ways to mitigate it.

While the adaptive recommendation aims at medium performance of a student on the task (measured by means described in Sect. 2.2), for the control tasks, the better performance the better. However, because the average difficulty of subsequent control tasks is increasing, the objective should also take into account the number of control tasks the student passed, e.g., as a sum of performances on control tasks.

## 3 Expected Contributions

We aim to make Hour of Code activities adaptive through a suitable domain model, performance measure, student model, and tutor model for task recommendation. To this end, we analyze data from multiple systems for introductory programming, perform simulation experiments, and design online experiments in RoboMission. Last year, RoboMission was used by 4 thousand students and collected 60 thousand task sessions and 1 million of program snapshots. Incorporating our findings into RoboMission helps us to validate them quickly. Nevertheless, we strive to make our research directly applicable to all systems for teaching introductory programming, because personalization of the Hour of Code can positively impact the lives of millions of children every year.

## References

1. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction based on adaptive learning technologies. In: Mayer, R.E., Alexander, P. (eds.) *Handbook of Research on Learning and Instruction*. Routledge, London (2016)
2. Baker, R.S.J.: Stupid tutoring systems, intelligent humans. *Int. J. Artif. Intell. Educ.* **26**(2), 600–614 (2016)
3. Caspersen, M.E., Christensen, H.B.: Here, there and everywhere - on the recurring use of turtle graphics in CS1. In: *ACM International Conference Proceeding Series*, vol. 8, pp. 34–40 (2000)
4. Csikszentmihalyi, M.: *Flow: The Psychology of Optimal Experience*. Harper & Row, New York (1990)
5. Effenberger, T.: Adaptive system for learning programming. Master's thesis, Masaryk University (2018)
6. Effenberger, T., Pelánek, R.: Towards making block-based programming activities adaptive. In: *Proceedings of Learning at Scale*, p. 13. ACM (2018)
7. Huang, Y., Hollstein, J.D.G., Brusilovsky, P.: Modeling skill combination patterns for deeper knowledge tracing. In: *UMAP (Extended Proceedings)* (2016)
8. Kelleher, C., Pausch, R.: Lowering the barriers to programming: a taxonomy of programming environments and languages for novice programmers. *ACM Comput. Surv. (CSUR)* **37**(2), 83–137 (2005)
9. Malone, T.W.: Making learning fun: a taxonomic model of intrinsic motivations for learning. In: *Conative and Affective Process Analysis* (1987)
10. Papoušek, J., Stanislav, V., Pelánek, R.: Evaluation of an adaptive practice system for learning geography facts. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 134–142. ACM (2016)
11. Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Model. User-Adapt. Interact.* **27**(3), 313–350 (2017)
12. Pelánek, R.: Conceptual issues in mastery criteria: differentiating uncertainty and degrees of knowledge. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10947, pp. 450–461. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_33](https://doi.org/10.1007/978-3-319-93843-1_33)
13. Pelánek, R.: The details matter: methodological nuances in the evaluation of student models. *User Model. User-Adapt. Interact.* **28**, 207–235 (2018)
14. Pelánek, R., Effenberger, T., Vaněk, M., Sassmann, V., Gmiterko, D.: Measuring item similarity in introductory programming. In: *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, p. 19. ACM (2018)
15. Ricci, F., Rokach, L., Shapira, B. (eds.): *Recommender Systems Handbook*. Springer, Boston, MA (2015). <https://doi.org/10.1007/978-1-4899-7637-6>
16. Wilson, C.: Hour of code-a record year for computer science. *ACM Inroads* **6**(1), 22 (2015)



# Leaving No One Behind: Educating Those Most Impacted by Artificial Intelligence

Laura Gemmell<sup>1,2</sup>(✉), Lucy Wenham<sup>1</sup>, and Sabine Hauert<sup>1,2</sup>

<sup>1</sup> University of Bristol, Bristol, UK

[laura.gemmell@brl.ac.uk](mailto:laura.gemmell@brl.ac.uk)

<sup>2</sup> Bristol Robotics Laboratory, Bristol, UK

**Abstract.** This paper describes the methodology to gather background information towards creating an educational framework for retraining in Artificial Intelligence (AI). The framework will be designed for those often missed by other educational efforts, such as mainstream education, university and large international companies. The intended participants are those often predicted to be most displaced by AI. The first step will involve conducting semi-structured interviews with individuals and companies to create a baseline of requirements for such an educational framework.

**Keywords:** Education · AI · Retraining

## 1 Problem and Previous Work

With advances in robotics and Artificial Intelligence (AI), there is a lot of uncertainty and fear about the future of work. Many figures have surfaced about how many people will be affected by these technological advances. Some sources believe one fifth of British jobs could be ‘displaced’ by 2030 [1] and, in 2018, McKinsey reported that up to 14% of workers worldwide (375 million people) will need to change jobs and reskill because of automation and AI by 2030 [18]. At the same time, other reports predict an increase in jobs overall, although the nature of these jobs may be different [4]. Whatever the figure, there is an apparent need to retrain many people from different backgrounds. This is a challenging problem, and there is concern large portions of the population could be left behind because they are not currently targeted in educational efforts.

A lot of excellent reports have been produced by companies, researchers and institutions, including the House of Lords [12], the Royal Society [30], KPMG [13] and PwC [25]. These reports discuss the need for retraining and reskilling if the UK is to avoid job losses and stay competitive in the global AI market (often this is referred to as the ‘skills gap’). The Royal Society Report on Machine Learning [30] recommends we ensure advances in AI benefits all of society. Programs dedicated to closing the skills gap focus on different levels of education. The government is taking steps to address this with a new National

Computing Centre training teachers to deliver the new computing curriculum [9], a new industry-funded masters degree [10] and 200 new PhD places in AI. While such measures are absolutely necessary, they miss the need to retrain a large portion of the population who are already working - many of whom may have left education many years ago and who may have limited mathematical or technological skills.

The government has accepted the concept of “lifelong learning” and announced a National Retraining Scheme which is due to launch later in 2019. The House of Lords Report, *AI in the UK: ready, willing and able?* [12], points to this as the type of pilot that is needed to ensure people have the skills they require. It also points out that lessons must be learnt from previous failed Government initiatives on adult learning (Adult Learning Accounts which were plagued by fraud [28] and the recent apprenticeship schemes which have been met with criticism [2,29]). Reskilling has been successful in a number of large companies, examples include employees of Marks and Spencers who are undertaking apprenticeships in Data Science [19] and Accenture (a professional services company) who automated 17,000 jobs yet had no job losses (which the CEO accredited to reskilling) [3]. When giving evidence to the House of Lords Report Select Committee on AI [12], Professor Richard Susskind discussed the retraining of truck drivers (who, according to a 2013 report by Frey et al., have a 69–90% chance of their jobs being taken by autonomous vehicles [8]). Susskind went on to say that the people displaced in these industries will not be retrained in the same way as others, such as software engineers, due to their educational background. We need to find new ways to empower displaced workers to become AI literate, users of technology, developers of this technology or to focus on finding new areas of work.

An example of such a retraining is being trialled in Finland as part of their country-wide strategy for AI [17] and a working group on continuous learning [11]. There are many schemes being piloted throughout the country - one was training 1% of the population in the basics of AI to ensure the electorate knows what they are voting for [6,7] and another by the Laurea School of Applied Science to educate the citizens of the City of Espoo on AI [16,26]. Whilst a lot may be learnt from these initiatives and can be used to build a framework for such pilots being rolled out in other countries care is needed whenever ideas are transferred to another context. It is worth noting, for example, that the educational attainment in Finland is higher than that of the UK according to the OECD [20,21]. It is particularly interesting to note that 76% of Finish adults already participate in some form of formal adult learning. Such differences will likely impact how policy initiatives play out in practice in such different settings.

## 2 Proposed Solution

The aim of this project is to create an educational framework which addresses the retraining needs of people potentially missed by other initiatives (education, academia and workplace retraining generally found in large international service

based companies, like Accenture). The educational framework can be considered to be split into the three stages (*Using*, *Understanding* and *Changing*). However, based on the initial stage of the project this can be pivoted to reflect what is required by the users. According to a report *How the UK can win the AI race produced* by KPMG [13], the UK public is very positive about the use of AI in the NHS. This will be used as a starting point for the content of the educational framework.

## 2.1 Proposed Methodology

To create the desired framework, the project has been divided into three main stages - Background, Education and Policy. This methodology will focus on the first stage which will involve conducting interviews with potential stakeholders in the educational framework - experts (including government and thought leaders), companies most affected by AI (including truck driving, retail and telecommunications), learning companies and, most importantly, individuals. The aim of this stage of the project is to explore people's attitudes which will help develop a hypothesis around what is needed and where the starting point for such an educational framework needs to be. This will provide a building block for more in depth interviews and hypothesis testing at later stages.

Semi-structured interviews with open-ended questions have been designed for the four groups of stakeholders. Interviews with individuals and learning companies aim to understand the baseline of requirements needed for such an educational framework. Guidelines in terms of educational level, and current English, mathematics and digital literacy skills of the intended audience will provide an excellent grounding as how to shape the framework. These interviews will study attitudes, such as fear of AI and willingness to retrain, which can be blockers to retraining in AI being well received or even attended. Samani et al [27] carried out interviews to determine similar guidelines when examining the gap between high and low performing students. Understanding the attitudes of the government and companies is important to allow a retraining scheme to be successful on a larger scale. An ideal outcome from these interviews is to find out if companies, local council or government departments are open to collaboration on a potential pilot of the framework.

The main reason for choosing interviews over questionnaires is the need for interviewees to feel invested in the project rather than studied [24]. When individual experience, perspective and opinion are sought, the interview is a common qualitative approach to access such data and to take listen to the voices of those directly involved. These interviews will also allow the interviewer to build a rapport with the interviewee, who may go on to participate further in shaping the educational framework. They also provide an opportunity to gain support for such a project in the local community, where a pilot is likely to take place.

Other reasons for choosing interviews over other methods, such as questionnaires, include them tending to have a higher response rate due to the immediacy of the answers being collected [5]. Some of those most relevant at this stage of the research may not have the necessary literacy or English reading skills to

complete a questionnaire. Using interviews will allow those with only conversational English skills to be included. Interviews will also allow terms and questions to be clarified, or tailored where necessary. For example, when trying to gauge how familiar individuals are with AI, different terms (AI, Artificial Intelligence, Machine Learning) or examples (Alexa, autonomous cars, Amazon recommending items) can be used. The Royal Society Report [30] found 9% of people hadn't heard of the term Machine Learning, but 89% had heard of particular applications of Machine Learning. At this stage, it would be helpful to provide the chance for respondents to give explanations and more depth. In an interview this can be guided by the interviewer and if the questions uncover unexpected information this can be discussed.

However, there are many known limitations with using interviews in research [15]. Firstly, they require more time and effort than questionnaires. They also must be transcribed and the data produced is harder to analyse. As these are exploratory interviews, the point of such interviews, as described by Oppenheim [22], is not necessarily data collection. They tend to be more heuristic - which is exactly what is needed at this stage to ensure it is a learning opportunity. Thus, there is a trade off for allowing more flexibility in the questioning. The number of interviewees will also be small enough that time is not a major issue.

Interviewer bias is an issue when conducting such research [5, 14, 15]. As are any inconsistencies that occur with any social interaction (such as changing questions when speaking to different interviewees and interviewer energy). In this case, the interviewer will be passionate about the subject being researched and this should not affect the responses from the interviewees who may be wary or annoyed with the concept of AI. The design of the interview has been done with this in mind. Introductions, explanations of key terms, main questions and follow ups have been scripted to ensure uniformity where needed. Another known limitation of interviews is reaching enough respondents. Working with local community points of interest (such as football stadiums, charities and community centres) should help recruit respondents. A partnership like this could also provide additional access to otherwise unreached members of the public.

Interviews will build on the work done by the Royal Society [30] and KPMG [13] in studying attitudes to robotics and AI. Focusing on individuals, whom this retraining will be benefiting, will ensure any education is user centric and built for those who need it most. Such an approach has been put forward in a report *Shaping the new National Retraining Scheme* [23] to ensure it works for individuals rather than just for companies.

### 3 Ask of the AIED Community

Advice from the collective expertise of the AIED community on best practices when carrying out such educational research would be of particular value at this point in the project. Guidance on how to scale this type of educational framework beyond the pilot, including how to make research like this internationally relevant when countries vary so much in attitudes and levels of education would also

be extremely useful. Finally, we welcome any pointers to existing efforts done around the world.

## References

1. Cities Outlook 2018. Centre for Cities (2018)
2. BBC: Apprenticeship levy is not working, employers say (2018)
3. Brinded, L.: Automation killed 17,000 roles at a huge tech and services firm - but no one actually lost their job. Business Insider, January 2017. <https://www.businessinsider.com/accentures-richard-lumb-davos-interview-robots-jobs-skills-leadership-training-2017-1>
4. Centre for European Economic Research: Robots create jobs - new research (2018). <https://ifr.org/ifr-press-releases/news/robots-create-jobs-new-research>
5. Cohen, L., Manion, L., Morrison, K.: Research Methods in Education. Education, Research methods. Routledge, New York (2011). <https://books.google.co.uk/books?id=p7oifuW1A6gC>
6. Delcker, J.: Finland's grand AI experiment (2019). <https://www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/>
7. FCAI: Elements of Artificial Intelligence free online course (2018). <https://www.elementsofai.com/>
8. Frey, C.B., Osborne, M.: The Future of Employment: How susceptible are jobs to computerisation? (2013)
9. Gov: Tech experts to provide National Centre for Computing Education (2018)
10. Gov: Next generation of artificial intelligence talent to be trained at UK universities (2019). <https://www.gov.uk/government/news/next-generation-of-artificial-intelligence-talent-to-be-trained-at-uk-universities>
11. Heinivirta, K.: Työryhmä parantamaan jatkuvan oppimisen mahdollisuksia (2019). [https://minedu.fi/artikkeli/-/asset\\_publisher/tyoryhma-parantamaan-jatkuvan-oppimisen-mahdollisuksia](https://minedu.fi/artikkeli/-/asset_publisher/tyoryhma-parantamaan-jatkuvan-oppimisen-mahdollisuksia)
12. House of Lords Select Committee on Artificial Intelligence: AI in the UK: ready, willing and able? House of Lords, 181, March 2018. <https://publications.parliament.uk/pa/ld201719/ldselect/l dai/100/100.pdf>
13. KPMG: How the UK can win the AI race: What we know, what the public think and where we go from here, 1–16 September 2018. <https://assets.kpmg.com/content/dam/kpmg/uk/pdf/2018/09/how-the-uk-can-win-the-artificial-intelligence-ai-race.pdf>
14. Kvale, S.: Doing Interviews. Qualitative Research Kit, SAGE Publications (2008). <https://books.google.co.uk/books?id=x7lXd08rD7IC>
15. Kvale, S.: Ten standard objections to qualitative research interviews. *J. Phenomenol. Psychol.* **25**(2), 147–173 (1994). <https://doi.org/10.1163/156916294X00016>. <https://www.sfu.ca/~palys/Kvale-TenStandardObjectionsToQualInterviews.pdf>
16. Lehtinen, T.: AI experiment, phase 1: Helping artificial intelligence (2017). [https://www.espoofi/en-US/Jobs\\_and\\_enterprise/A\\_dynamic\\_city/Glimpses\\_into\\_the\\_future/AI\\_experiment\\_phase\\_1\\_Helping\\_artificial\(133974\)](https://www.espoofi/en-US/Jobs_and_enterprise/A_dynamic_city/Glimpses_into_the_future/AI_experiment_phase_1_Helping_artificial(133974))
17. Lintilä, M.: AI Finland - Background (2017). <https://www.tekoalyaika.fi/en/background/>

18. Manyika, J., et al.: DONE-Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation. McKinsey Global Institute, pp. 1–160, December 2017. <https://doi.org/10.1002/lary.20616>, <https://www.mckinsey.com/featured-insights/future-of-organizations-and-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>
19. Noble, L.: Marks & Spencer creates Data Science Academy using Apprenticeship levy (2018). <https://www.hrdconnect.com/2018/08/03/marks-spencer-creates-data-science-academy-using-apprenticeship-levy/>
20. OECD: Education at a Glance 2018. Education at a Glance, OECD, September 2018. <https://doi.org/10.1787/eag-2018-en>, [https://www.oecd-ilibrary.org/education/education-at-a-glance-2018\\_eag-2018-en](https://www.oecd-ilibrary.org/education/education-at-a-glance-2018_eag-2018-en)
21. OECD: Finland - Overview of the education system (EAG 2018). Technical report, OECD (2018)
22. Oppenheim, A.N., Oppenheim, A.N.: Questionnaire Design, Interviewing, and Attitude Measurement. Pinter Publishers; Distributed exclusively in the USA and Canada by St. Martin's Press, New York (1992)
23. Pember, S.: Shaping the new National Retraining Scheme, March (2018). <https://feweek.co.uk/wp-content/uploads/2018/03/Shaping-the-new-National-Retraining-Scheme.pdf>
24. Plas, J.M., Kvale, S.: InterViews: An Introduction to Qualitative Research Interviewing. SAGE Publications, Thousand Oaks (1996)
25. PwC: The economic impact of artificial intelligence on the UK economy. Technical report, PwC. <https://www.pwc.co.uk/services/economics-policy/insights/the-impact-of-artificial-intelligence-on-the-uk-economy.html>
26. Ristimäki, M.: The City of Espoo: a unique experiment with AI (2018). <https://www.tieto.com/en/success-stories/2018/the-city-of-espo-o-a-unique-experiment/>
27. Samani, T., Porayska-Pomsta, K., Luckin, R.: Bridging the gap between high and low performing pupils through performance learning online analysis and curricula. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 650–655. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_82](https://doi.org/10.1007/978-3-319-61425-0_82)
28. Sparrow, A.: Labour adult learning vow ‘led to £100m fraud fiasco’, April 2003. <https://www.telegraph.co.uk/news/uknews/1426543/Labour-adult-learning-vow-led-to-100m-fraud-fiasco.html>
29. Stancliffe, M.: What’s going wrong with the apprenticeship levy? (2013). <https://www.personneltoday.com/hr/whats-going-wrong-with-apprenticeship-levy/>
30. The Royal Society: Machine learning: the power and promise of computers that learn by example, vol. 66 (2017). <https://doi.org/10.1126/scitranslmed.3002564>, <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>



# Modeling Students' Behavior Using Sequential Patterns to Predict Their Performance

Mehrdad Mirzaei<sup>(✉)</sup> and Shaghayegh Sahebi

University at Albany – SUNY, Albany, NY 12203, USA  
[{mmirzaei,ssahebi}@albany.edu](mailto:{mmirzaei,ssahebi}@albany.edu)

**Abstract.** Online learning environments generate educational data that can be used to model students' behavior and predict their performance. In online learning environments, in which students are free to choose their next activity, various factors such as time spent on individual tasks and the choice of next learning material may impact students' performance. The main goal of this research is to enhance student learning by modeling students' behavior and testing whether these behavioral patterns correlate with their performance. Using sequential pattern mining methods, we will identify the most frequent patterns in students' online learning activities and test whether/which patterns correlate with higher or lower performance. By identifying which student behavioral patterns correlate with higher or lower performance, this study has the potential to inform redesign of online learning platforms and study guidelines that help students learn more and perform better.

**Keywords:** Sequential pattern mining · Student performance · Matrix factorization

## 1 Introduction

Previous research has shown that given the choice, students may take on repetitive and non-productive behaviors in solving problems [2]. Also, it has been shown that students can be grouped into various clusters according to their studying patterns, but these clusters are not directly related to students' performance. Although insightful, the past research has mostly focused on one type of learning material in analyzing student behaviors. Moreover, these studies have not drawn clear distinctions between non-productive and productive behavioral patterns. The goal of this research is to examine different factors in forming students' behavior and to draw conclusions that can improve students' learning and performance. The performance is defined as the grades of the students or learning gain in a course. Both educators and students can take advantage of these findings. Students can adopt patterns that are useful for learning and avoid patterns that may inhibit learning. Educators may also be able to use this behavioral data to identify struggling students earlier in a course and intervene

to improve their behavior before important assessments. In this work, we propose a general approach to find frequent patterns of students' behavior using online educational platforms.

## 2 Related Work

A group of studies aim to group students based on their behaviors. Researchers have employed data mining algorithms such as clustering in these works. In [1], they have used two different approaches to extract frequent sequences of actions in a collaborative learning environment to distinguish high achieving students from low achieving ones in small groups. In [2] patterns of student behavior with parameterized exercises are modeled and analyzed. In this work, micro patterns are extracted using a frequent mining algorithm and are used to build macro patterns to cluster students in groups with similar patterns. We extend this method by combining students' behavioral patterns on multiple types of learning materials (e.g., worked examples and problems). A number of researches have used matrix factorization to find latent factors as patterns in student data. In [5] non-negative matrix factorization is used to cluster gene expressions and find factorization with the same gene expression profiles by estimating offsets for individual genes. In [4] a method is proposed to identify common and discriminative topics among a set of given documents according to their keywords. We will adapt and extend this work to distinguish between useful and harmful patterns of students' interactions with online learning material.

## 3 Research Questions

Various factors may impact student performance, such as time spent on individual tasks and how a student chooses to engage with the online platform. The goal of this research is to examine such factors and draw conclusions that could improve the efficiency of the students and efficacy of online learning tools. Student activities and decision-making while functioning in a computer-based learning environment are underutilized and could be used to guide students with effective patterns in studying. The information obtained in this analysis will be used to answer the following questions:

**Question 1:** Do individual students exhibit stable behavioral patterns in their work with learning content, or does their learning approach depend on factors, such as time of the semester or learning material complexity?

**Question 2:** Are student behavioral patterns associated with their learning performance?

**Question 3:** How accurately can we discriminate between students' productive behavioral patterns vs. the non-productive ones?

## 4 Previous Work

We extracted students' behavior patterns while interacting with an online learning environment. These patterns are consecutive actions in the sequence of students' activities. To extract the patterns, we used a sequential pattern mining method (CM-SPAM) [3]. Then extracted patterns are used to build a vector for each student that contains the frequency of all different patterns and model the student's behavior. Clustering the pattern vectors, we discovered three clusters with distinct patterns. We call these clusters: "Confirmers", "Thinkers", and "Readers". "Confirmers" mostly tend to confirm their success by repeating to solve a problem again and again. "Thinkers" are the group that achieve success after some failed attempts, and have longer activities than other student groups. "Readers" usually spend more time on reading the worked examples.

## 5 Future Work

To cluster students' behavioral patterns, we propose a matrix factorization method (MF), extending the work by Kim et. al [4]. Having two sets of documents, the model in [4] finds topics from each document set, among which some topics are common between the two document sets and the rest of the topics are different between them. Another possible direction is to embed social networks of students in online courses to enhance the performance prediction [6].

### 5.1 Proposed Method

Our proposed method is based on the model in [4]. We will use MF to find common patterns and distinct patterns between two groups of students. To have similar patterns in each group, we will minimize the differences between patterns. Students' performance will be used to distinguish them in two different groups: high and low-performers. The common patterns will be considered as ordinary patterns that represent students in both groups, but distinct patterns in each group are specific to that group.

### 5.2 Problem Formulation

We have the pattern vectors of high-performance students and low-performance students extracted in  $X_1$  and  $X_2$ . The purpose is to find  $k$  pattern clusters such that  $k_c$  of pattern cluster are common between two groups of students and  $k_d$  of them are different between students. So there are two matrices that should be decomposed:

$$X_1 \approx W_1 H_1^T \quad X_2 \approx W_2 H_2^T \quad (1)$$

We split  $W$  and  $H$  to have common and discriminative pattern clusters. The matrices are split in this way:

$$W_1 = [W_{1,c} \quad W_{1,d}], \quad W_2 = [W_{2,c} \quad W_{2,d}] \quad (2)$$

$$H_1 = [H_{1,c} \quad H_{1,d}], \quad H_2 = [H_{2,c} \quad H_{2,d}] \quad (3)$$

$W_{1,c}$  and  $W_{2,c}$  are similar pattern clusters and  $W_{1,d}$  and  $W_{2,d}$  are distinct ones. We should define functions to calculate how common or distinctive the patterns are and add them to the formulation.

The model in [4] is proposed to find common and discriminative topics in two document sets. We replace documents and words with pattern vectors and patterns respectively to find patterns that are different between two groups. Moreover, we use pattern similarity matrix in the model, since we expect to have similar patterns in each group.

## References

1. Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., Kharrufa, A.: Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In: Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), pp. 111–120 (2011)
2. Guerra, J., Sahebi, S., Brusilovsky, P., Lin, Y.R.: The problem solving genome: analyzing sequential patterns of student work with parameterized exercises. In: Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pp. 153–160 (2014)
3. Fournier-Viger, P., Gomariz, A., Campos, M., Thomas, R.: Fast vertical mining of sequential patterns using co-occurrence information. In: Tseng, V.S., Ho, T.B., Zhou, Z.-H., Chen, A.L.P., Kao, H.-Y. (eds.) PAKDD 2014. LNCS (LNAI), vol. 8443, pp. 40–52. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-06608-0\\_4](https://doi.org/10.1007/978-3-319-06608-0_4)
4. Kim, H., Choo, J., Kim, J., Reddy, C.K., Park, H.: Simultaneous discovery of common and discriminative topics via joint non-negative matrix factorization. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 15), pp. 567–576 (2015)
5. Badea, L.: Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In: Pacific Symposium on Biocomputing, pp. 267–278 (2008)
6. Doan, T., Lim, E.: Modeling location-based social network data with area attraction and neighborhood competition. Data Min. Knowl. Disc. **33**, 58–95 (2019)



# Personalization in OELEs: Developing a Data-Driven Framework to Model and Scaffold SRL Processes

Anabil Munshi<sup>(✉)</sup> and Gautam Biswas

Institute for Software Integrated Systems, Vanderbilt University,  
Nashville, TN 37212, USA  
[{anabil.munshi,gautam.biswas}@vanderbilt.edu](mailto:{anabil.munshi,gautam.biswas}@vanderbilt.edu)

**Abstract.** This research focuses on developing a data-driven framework for modeling and scaffolding learners' self-regulated learning (SRL) processes in open-ended learning environments (OELE). The aim of this work is to offer a personalized and productive learning experience by adapting scaffolds to help learners develop self-regulation skills and strategies. This research applies mining techniques on data collected from multiple channels to track learners' *cognitive, affective, metacognitive and motivational* (CAMM) processes as they work in Betty's Brain, a computer-based OELE. The CAMM information is used to derive online models of learners' SRL processes. These learner models inform the design of personalized scaffolds that help students develop the required SRL process and become more proficient learners. The significance of this research lies in developing and using data-driven learner SRL models to personalize and contextualize the scaffolds provided to learners within the OELE.

**Keywords:** Personalization · Self-regulated learning · Metacognition · Adaptive scaffolding · CAMM processes · Multimodal data mining · Open-ended learning environments

## 1 Introduction

Researchers have emphasized the need for personalizing computer-based learning environments (CBLE) to support complex learning and problem-solving goals of learners. Support for personalization in a CBLE can be incorporated into the design of the system. For example, open-ended learning environments (OELE) are designed to provide learners with a choice in how they use the tools provided in the system to construct solutions to complex problems, thus promoting the use of higher order thinking skills, such as the use of strategies and self-regulated learning (SRL) processes. However, novice learners often have difficulties in invoking effective regulation strategies, and therefore, run into obstacles and fail to accomplish their learning goals. This motivates the need for designing adaptive scaffolds, so that learners can be provided individualized help at appropriate times, i.e., when they need them most. To accomplish this form of adaptivity and personalization, it is important to derive learner

models that can keep track of learners' progress in their learning and problem-solving tasks, as well as the SRL processes they apply when working on these tasks.

A learner model typically seeks to represent the evolving knowledge, as well as difficulties and misconceptions learners have as they work towards their learning goals. SRL theories explore how these knowledge acquisition processes unfold in recursive cycles - where learners set goals, form plans, apply strategies, monitor their performance, and then reflect on outcomes [9]. Self-regulated learners who are successful in this cyclical learning process are capable of monitoring and controlling their *cognitive, affective, metacognitive, and motivational (CAMM) processes* [1]. Therefore, to generate efficient learner models (and inform subsequent scaffold design), it becomes important to track the temporal evolution of learners' SRL, i.e., the evolution of their dynamic *CAMM* states.

Our research aims to derive SRL models of learners solving complex science problems in Betty's Brain, [3], an OLE, where students learn by teaching a *virtual pedagogical agent* called Betty. Students teach Betty by constructing causal models of scientific processes. The learning environment provides tools and resources to develop learners' cognitive and metacognitive processes such as information acquisition, solution construction and solution evaluation.

Open-ended learning environments like Betty's Brain encourage exploration and strategic thinking. But this very open-ended nature of OLEs can make the process of tracking learners' strategic and self-regulatory (SRL) behaviors a challenging task. Our research uses CAMM data collected from multiple parallel channels (*system logs, video logs, self-report, eye-trackers, etc.*) to track and model SRL in Betty's Brain. Interpreting this data using data mining techniques and generating learner SRL models helps us to design and personalize our framework for scaffolding learners. The scaffolds are offered as conversational feedback via the virtual agents in Betty's Brain.

## 2 Background

SRL theories explore learners' behavioral, cognitive and metacognitive processes in learning, with emphasis on the roles of self and external feedback on the regulation of these processes. While SRL has been studied in a wide variety of contexts, computer-based learning environments (CBLEs) present unique opportunities for fostering SRL in science learning [2]. CBLEs can represent information in many ways, and it is often up to the learners to decide which representations are most helpful, based on their motivational factors, prior knowledge, task definitions, goals, and strategic knowledge [9]. Research has indicated that learners' SRL processes may facilitate the theorized positive correlations between CBLEs and learner performance [2]. So, it is crucial to model learners' SRL processes in CBLEs. In case of the more open-ended CBLEs (*OLEs like Betty's Brain*), the availability of multiple learning paths gives more agency to the learner – in turn, providing them with opportunities for developing relevant self-regulation strategies. So, deriving SRL models in OLEs can be leveraged to provide external scaffolds to the learner during key moments in their learning behavior. Such interventions can help more learners take advantage of the unique

learning affordances provided by these learning environments, while also ensuring a more personalized and productive learning process.

### 3 Proposed Methodology

The first step in our research is to measure SRL from learner data. Panadero et al. [7] captured the “three waves” in SRL measurement techniques – (i) self-report, (ii) “online” measures that trace actual learner activities, and (iii) combining intervention and assessment. Our approach employs a combination of the three techniques to measure SRL in classrooms where learners are working with Betty’s Brain. Multimodal data channels facilitate the tracking and validation of fine-grained *CAMM* information –

- i. ***Cognitive*** – Learners’ logged activity is the primary source of cognitive information. This log data can be mined using pattern mining techniques in real time to track *cognitive “inflection points”* which indicate changes in goal or *active strategy use* by the learner (*e.g.: when a learner shifts from constructing solutions to evaluating their current solution*). A second source of cognitive information is obtained from eye-tracking devices. Learners’ eye-gaze features (*e.g.: gaze fixation while performing a particular learning activity*) provide the basis for deriving fine-grained classifiers that identify individualized cognitive information.
- ii. ***Affective*** – Affect detectors built within the Betty’s Brain system, trained on affect data collected by human researchers using the BROMP tool [6], track learners’ affective inflection points (*e.g.: when a learner shifts from a state of engagement to a state of boredom*). Real-time processing of learners’ facial videos using facial affect recognition software can be a second data source to track learners’ affective states and validate the BROMP-driven affect data.
- iii. ***Metacognitive*** – Learners’ internal metacognition cannot be captured directly from observed activity data. So, our framework involves *real time human audio-interviews* at specific *cognitive-affective inflection* points to obtain metacognitive information.
- iv. ***Motivational*** – Self-report questionnaires track information on students’ motivation (self-efficacy, task value, etc.) during the learning process.

In addition, we have summative (*pre-to-post learning gains*) and formative (*assessments of learner solutions in Betty’s Brain*) measures of learner performance.

The multimodal data channels described above allow us to track and interpret learners’ CAMM processes in Betty’s Brain as they work on the system. We leverage this information to constantly update and refine the SRL model of each learner.

Our scaffolding framework is informed by the values of CAMM parameters of each individual learner’s SRL model. So, as a learner’s CAMM processes evolve temporally based on their interactions with the learning environment, the scaffolding framework adapts to the current self-regulatory needs of the learner. The scaffolds are triggered as conversational feedback initiated by one of the virtual agents in the system – student responses to each piece of conversation inform the agent’s corresponding response, allowing for a dynamic and personalized learning experience. When a stable version of

our feedback is in place, we intend to analyze learner interactions with the provided feedback to refine future iterations of the scaffolding framework.

## 4 Current Work and Future Plans

As a first step towards our goal, we conducted classroom experiments with 99 middle school students of an urban public school in Nashville, USA, who built causal models of climate change in Betty's Brain over 4 days. We collected data on learner cognition (*log traces*), affect (*affect detector, discussed above*) and motivation (*self-report*). We applied mining techniques (*sequential pattern mining in [4]; temporal log analysis in [5]; process mining in [8]*) on the data to derive empirical measures of learners' CAMM processes and how they are influenced by interactions with the agents in Betty's Brain. Our findings have implications on shaping future work towards our research goal. They help us understand how students' regulatory skills unfold in OLEEs, and show the ability of our proposed data-driven methodology to track learners' SRL and performance in OLEEs. We are currently using the findings from our initial experimental studies to refine and shape our learner model. Once that is developed, our plan is to design personalized scaffolds in the form of agent conversations embedded within the system. We intend to conduct more classroom studies to analyze the impact of the designed scaffolds on individual learners, and thereby use our findings to enhance the efficiency of our scaffolding framework. We hope that the external regulations provided through these scaffolds will help learners gain awareness of their SRL processes, thereby helping them regulate said processes and take control of their own learning.

## References

1. Azevedo, R., Behnagh, R., Duffy, M., Harley, J., Trevors, G.: Metacognition and self-regulated learning in student-centered leaning environments. In: Theoretical Foundations of Student-Centered Learning, pp. 171–197 (2012)
2. Lajoie, S.P., Azevedo, R.: Teaching and learning in technology-rich environments. Handbook of Educational Psychology, 2nd edn, pp. 803–821. Erlbaum, Mahwah (2006)
3. Leelawong, K., Biswas, G.: Designing learning by teaching agents: the Betty's Brain system. IJAIED **18**(3), 181–208 (2008)
4. Munshi, A., Rajendran, R., Moore, A., Ocumpaugh, J., Biswas, G.: Studying the interactions between components of self-regulated learning in open ended learning environments. In: ICLS 2018, London (2018)
5. Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R.S., Paquette, L.: Modeling learners cognitive and affective states to scaffold SRL in open ended learning environments. In: UMAP 2018, Singapore (2018)
6. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T.: Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 technical and training manual. Technical report 2015
7. Panadero, E., Klug, J., Järvelä, S.: Third wave of measurement in the self-regulated learning field: when measurement and intervention come hand in hand. Scand. J. Educ. Res. (2015). <https://doi.org/10.1080/00313831.2015.1066436>

8. Rajendran, R., Munshi, A., Emara, M., Biswas, G.: A temporal model of learner behaviors in OLEEs using process mining. In: Yang, J.C. et al. (eds.) Proceedings of ICCE 2018, Manila, Philippines, pp. 276–285 (2018)
9. Winne, P.H., Hadwin, A.F.: The weave of motivation and self-regulated learning. In: Motivation and Self-Regulated Learning: Theory, Research and Applications, pp. 297–314 (2008)



# Analyzing Engagement in an On-Line Session

Vandana Naik<sup>(✉)</sup> and Venkatesh Kamat

Department of Computer Science and Technology, Goa University, Taleigao, Goa,  
India

{dcst.vandana,vvkamat}@unigoa.ac.in

<https://www.unigoa.ac.in/department-of-computer-science-technology.html>

**Abstract.** It is well known that if the learning strategies align with learning outcomes, learner well engaged in the session is likely to make progress in acquiring knowledge. However, it is challenging to ascertain learner's engagement in an online environment and to guess their grasp on particular topics. The objective of this work is to check for relations between the engagement and the performance. Firstly, log traces for each learner in a session depending on their interaction will be labeled. These features are analyzed to calculate engagement indicators that represent the level of learner's involvement and engagement levels per activity and session. This will help to identify the less engaged learners as well as to inform about the low engaging sessions or a particular activity in the sessions. It could be used in an adaptive learning environment to update the learning process by providing more engaging activities. Using the quantified traces, the prediction of the performance based on the interactions of the learner will be attempted. The training dataset from completed courses with labeled performance will be used to develop a model that can effectively predict the performance well in advance. This can help to prescribe techniques like extra help through more exercises, reference material for whom the predicted performance is below the threshold level. Supervised machine learning algorithms like neural networks, random forest and support vector machines will be explored to understand the prominent indicators of performance and to compare and find the most efficient algorithm for the purpose.

**Keywords:** Engagement · Performance · Log tracing

## 1 Introduction

Learner engagement refers to the depth of consideration, inquisitiveness, interest, optimism, and passion that learners show when they are learning or being

---

This publication is an outcome of the R&D work undertaken under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation.

taught, which extends to the level of their progress [20]. Engagement is essential to study for mainly three reasons [23]. It is a necessary condition to learn. Next, it decides student's everyday experiences in school, both psychologically and socially. Third, engagement is a critical contributor to student's long term academic development. A thorough study of the literature reveals that the dynamics of engagement derives from more than one dimension. The longitudinal study of the long term effect of these engagement levels on the conduct of the individuals in the future has also shown positive results. A careful study of this literature brings out the most commonly occurring categorization where it is divided into four dimensions. Martin [13] proposes a two-dimensional model comprising mainly cognitive and behavioral dimensions. Cognitive engagement incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills. Behavioral engagement includes persistence, effort, attention, participation, involvement [7]. The work of Fredricks [5] adds a third dimension to the above called the emotional component. Emotional engagement includes interest, boredom, happiness, anxiety, and other affective states. Lastly, the four-dimensional model proposed by [1] adds the fourth dimension, academic component and includes time on task, credits earned, and homework completion. For this study, we are going to focus on cognitive, behavioral and academic dimensions as the techniques like webcams or physiological sensors that can help detect emotional engagement but suffer from various problems as discussed in next section are not being used. The logs would be used to define the behavior within the system and would involve aspects like time spent, number of times accessed. Some of these along with the depth of interaction, the difficulty level of the activity may help to define the cognitive engagement. Their performance in both formative and summative assessments could be used to explain the academic dimension. Such log traces can then be used to judge the involvement level of the students. The earlier psychological research has drawn up these dimensions that have allowed checking of various aspects of engagement and helped to predict their effect on student performance [2, 11, 22]. Therefore the calculated involvement level is used to reestablish this connection and then for prediction.

## 2 Related Work

### 2.1 Measuring Engagement Level

Engagement measurement can be carried out in several ways. Self-reporting questionnaires are used widely, wherein the students themselves choose from options or provide answers that help to judge the engagement level. Another way is through teacher rating or field observations where the questions are answered by teachers or observers [6]. Some of the disadvantages of these methods include biasing, false reporting and unscalability. Another type of detection technique that is device controlled based, involves tracking through external devices like webcam that can be used for eye tracking or for capturing the face, body posture and hand gestures [8]. These are hardware dependent and need to continuously

run to track the activities and also need the learner to be screen tied [14]. Another method that includes external devices is one where a physiological sensor that capture features like EEG, blood pressure, heart rate for prediction [3] is used. These devices again are hardware dependent and invasive and therefore can interfere with the results. A better method that does not require any extra hardware or is free from problems like false reporting is automatic inference through the logs of the learner created in an online environment. The current work also falls in this category of research. In most of the studies in this category, not all the engagement indicators are used, for example, [10] focus only on time aspect, while in [15] only participation time and frequency are used. In others, only specific activities from all those available in an online environment are used. For instance, [19] the authors use only forums activities, and in [7] it is limited to an experimental type of content. In our study, however we propose to use every type of activity and all user actions within it irrespective of the content type.

## 2.2 Automated Techniques of Engagement Detection

One of the most closely related works to our study is [12]. It uses algorithm derived models. Here an algorithm first determines optimal parameters of engagement, restricted to three indicators namely assessment, forum, and log in. It calculates the parameters and weightings, with some initial guess that is improved by maximizing the inverse correlation between total risk rating and final course grade for each student. These algorithm derived ratings are compared with the ones calculated manually by teachers by conceptualizing what they expected of a good student. The findings recommend the use of a human intervention for better prediction. However, the current work differs with this work in two aspects. Firstly there is no constraint on the indicators. All activities set up by the instructor in the session will be used. The information inferable from logs related to them, like time spent, no of attempts, the effort as defined by the depth of interaction within each activity will be used to get more meaningful interpretations. Secondly, there will be no human input in these predictions. Another automated technique uses machine learning algorithms. A related work that uses such algorithms is that of Analytics Moodle [18]. Here, using the Community of Inquiry paradigm the depth of social and cognitive interactions are evaluated to predict the risk of dropout for each student. The proposed research is similar to this work, and as a first step calculates the engagement indicators using the supervised learning framework for learning management systems (LMS), as described in it. However, this framework has not been used for engagement analysis to predict performance but has much potential of being used for the same. We feel it can provide a novel way of predicting performance with higher accuracy as compared to other works like [4, 16, 21] that use only limited indicators like the forum, assignment, and quiz. Unlike the Analytics feature that uses only linear indicators right now, we intend to incorporate binary and discrete indicators as well for the calculations. Another vital contribution of our

work is that it intends to provide reports of engagement at the various levels like activity, session and per student.

### 3 Proposed Research

Our research aims to use the supervised learning framework for LMS [18] to calculate the engagement indicators. The experiments will be carried out in the LMS provided by the Consortium for Indian Information Technology Education (CIITE – LMS). The dataset of already completed courses with added indicators along with the performance (as a label) will be used for training. Based on the accuracy of the model it can then be deployed to predict the performance of the learners as per their activity interactions. Reports will be generated per session, activity, and student. A list of prominent indicators will be decided through a pilot study, and then for every action taken by the learner within the system, their values are calculated. This calculation quantifies the logs corresponding to indicators (features in supervised learning) into three values: (a) linear - a floating number (value: -1 to 1), for example, duration of video watched within the session. (b) binary (value: 0 or 1), for example, has the student attempted the assignment. (c) discrete (value: closed list, one hot encoded), for example, the difficulty level of the questions. These features are added to the tables of student, activity, and session. Further processing can help to obtain the next level of information like depth of interaction, time spent, etc. For example, a student viewing a thread in the forum will be at level one. If they are posting on a thread, they are at level two. If they are helping others by providing an answer, they are at level three and so on. This calculation is based on the technique in [9]. This process is repeated for every activity accessed by the user within the session, and the overall value is calculated for the learner that indicates his/her engagement level. The activity engagement level can be also be calculated to indicate to the faculty which sessions/activities are not engaging enough or could be used as input to the adaptive systems that can then change the activities based on the learner's needs and preferences. Another offshoot of this research work is to verify whether the engagement parameter framework proposed in [17] helps to improve the engagement level if the activities are classified as per the framework.

### 4 Conclusion

The current research work aims to help in calculating the engagement levels in the online sessions and then predict performance based on these levels. It is proposed to use CIITE – LMS as a platform to track the interactions in the form of logs in the system and calculate the engagement levels for each session, activity and student. This could be used to inform faculty about the less engaging sessions/activities and also in adaptive systems to modify the environment to improve the learner experience. The same engagement indicators can then be used to predict the performance of the student and take action to avoid lower grades.

## References

1. Appleton, J.J., Christenson, S.L., Kim, D., Reschly, A.L.: Measuring cognitive and psychological engagement: validation of the student engagement instrument. *J. Sch. Psychol.* **44**(5), 427–445 (2006)
2. Carini, R.M., Kuh, G.D., Klein, S.P.: Student engagement and student learning: testing the linkages\*. *Res. High. Educ.* **47**(1), 1–32 (2006)
3. Chaouachi, M., Chalfoun, P., Jraidi, I., Frasson, C.: Affect and mental engagement: towards adaptability for intelligent. In: FLAIRS Conference (2010)
4. Dascalu, M., Popescu, E., Becheru, A., Crossley, S., Trausan-Matu, S.: Predicting academic performance based on students' blog and microblog posts. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) EC-TEL 2016. LNCS, vol. 9891, pp. 370–376. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45153-4\\_29](https://doi.org/10.1007/978-3-319-45153-4_29)
5. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**(1), 59–109 (2004)
6. Fredricks, J.A., McColskey, W.: The measurement of student engagement: a comparative analysis of various methods and student self-report instruments. In: Christenson, S.L., Reschly, A.L., Wylie, C. (eds.) *Handbook of Research on Student Engagement*, pp. 763–782. Springer, Boston (2012). [https://doi.org/10.1007/978-1-4614-2018-7\\_37](https://doi.org/10.1007/978-1-4614-2018-7_37)
7. Gobert, J.D., Baker, R.S., Wixon, M.B.: Operationalizing and detecting disengagement within online science microworlds. *Educ. Psychol.* **50**(1), 43–57 (2015)
8. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: predicting engagement and frustration. In: Educational Data Mining 2013 (2013)
9. Henrick, P.B.G.: Moodle analytics plans: project inspire. Moodle Moot (2018)
10. Joseph, E.: Engagement tracing: using response times to model student disengagement. In: Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology, vol. 125, p. 88 (2005)
11. Lee, J.S.: The relationship between student engagement and academic performance: is it a myth or reality? *J. Educ. Res.* **107**(3), 177–185 (2014)
12. Liu, D.Y.T., Richards, D., Dawson, P., Froissard, J.-C., Atif, A.: Knowledge acquisition for learning analytics: comparing teacher-derived, algorithm-derived, and hybrid models in the moodle engagement analytics plugin. In: Ohwada, H., Yoshida, K. (eds.) PKAW 2016. LNCS (LNAI), vol. 9806, pp. 183–197. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-42706-5\\_14](https://doi.org/10.1007/978-3-319-42706-5_14)
13. Martin, A.J.: Enhancing student motivation and engagement: the effects of a multidimensional intervention. *Contemp. Educ. Psychol.* **33**(2), 239–269 (2008)
14. Miller, B.W.: Using reading times and eye-movements to measure cognitive engagement. *Educ. Psychol.* **50**(1), 31–42 (2015)
15. Morris, L.V., Finnegan, C., Wu, S.S.: Tracking student behavior, persistence, and achievement in online courses. *Internet High. Educ.* **8**(3), 221–231 (2005)
16. Mwalumbwe, I., Mtebe, J.S.: Using learning analytics to predict students' performance in moodle learning management system: a case of Mbeya University of Science and Technology. *Electron. J. Inf. Syst. Dev. Countries* **79**(1), 1–13 (2017)
17. Naik, V., Kamat, V.: Predicting engagement using machine learning techniques. In: 26th International Conference on Computers in Education Doctoral Student Consortium Proceedings, pp. 17–20. Asia-Pacific Society for Computers in Education (APSCE), Taiwan, November 2018

18. Olivé, D.M., Huynh, D.Q., Reynolds, M., Dougiamas, M., Wiese, D.: 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, October 2018
19. Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., Getoor, L.: Uncovering hidden engagement patterns for predicting learner performance in MOOCs. In: Proceedings of the First ACM Conference on Learning @ Scale Conference, pp. 157–158. ACM, March 2014
20. Rodríguez, C.A.V., Lavalle, M.M., Elías, R.P.: Modeling student engagement by means of nonverbal behavior and decision trees. In: 2015 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE), pp. 81–85, November 2015
21. Romero, C., Espejo, P.G., Zafra, A., Romero, J.R., Ventura, S.: Web usage mining for predicting final marks of students that use moodle courses. *Comput. Appl. Eng. Educ.* **21**(1), 135–146 (2013)
22. Shulman, L.S.: Making differences: a table of learning. *Change Mag. High. Learn.* **34**(6), 36–44 (2002)
23. Skinner, E.A., Pitzer, J.R.: Developmental dynamics of student engagement, coping, and everyday resilience. In: Christenson, S.L., Reschly, A.L., Wylie, C. (eds.) *Handbook of Research on Student Engagement*, pp. 21–44. Springer, Boston (2012). [https://doi.org/10.1007/978-1-4614-2018-7\\_2](https://doi.org/10.1007/978-1-4614-2018-7_2)



# A Machine Learning Grading System Using Chatbots

Ifeanyi G. Ndukwe<sup>1</sup>(✉), Ben K. Daniel<sup>1</sup>(✉), and Chukwudi E. Amadi<sup>2</sup>(✉)

<sup>1</sup> University of Otago, Dunedin, New Zealand  
[{glory.ndukwe,ben.daniel}@otago.ac.nz](mailto:{glory.ndukwe,ben.daniel}@otago.ac.nz)

<sup>2</sup> Federal University of Technology, Owerri, Nigeria  
[emmanuel.amadi@futo.edu.ng](mailto:emmanuel.amadi@futo.edu.ng)

**Abstract.** The ability to provide students with timely and accurate feedback is critical to learning. However, grading written essays is demanding, and can be challenging to conduct in large classes. We explore an automated grading system involving the use of a Chatbot that asks students questions, requiring written responses. We implemented unsupervised machine learning techniques for the task of automated grading and conducted an experiment to assess the performance of the Chatbot as compared to human grading. The experiment involved posting questions to 15 students, requiring short written answers. To analyse the performance of the Chatbot, we used a combination of term-frequency inverse-document function (tfidf) with cosine Euclidean distance, and online semantic text analytics (Dandelion API), trained with neural networks on a large bank of questions and answer dataset. We then used Cohen's kappa agreement. The result shows a good inter-rater agreement level between the automated grading and the human instructor. The work presented in the paper presents open up opportunities for using Chatbots in providing automated assessment and at the same time fosters engagement with students.

**Keywords:** Chatbot · Machine learning · Similarity · Short answer grading

## 1 Introduction and Motivation

Research to date has focused on two fundamental subtasks of computer-assisted assessment such as essay grading carried out based on spelling check, grammatical expressions, essay coherency and style [1, 13, 14], and the assessment of short answer texts [6, 8, 12, 15]. Chatbots are designed to hold conversations with users using natural language [11]. Goel [5] introduced a virtual teaching assistant in an online course to answer questions in the classroom, such that students thought that they were interacting with a human. In recent times, Chatbots are used to help automate the grading of assignments in Massive Open Online Courses

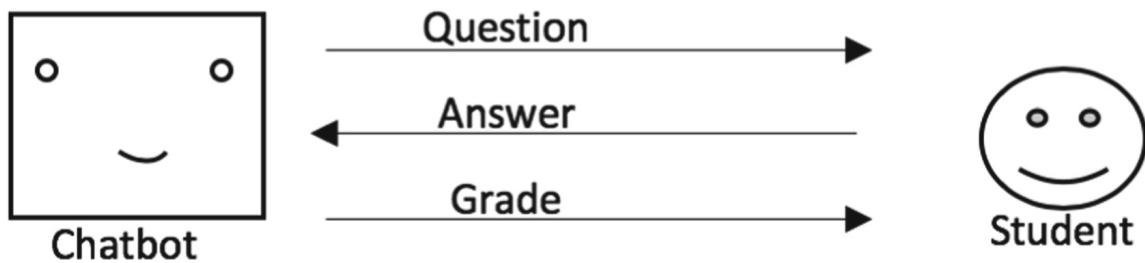
---

Supported by University of Otago, New Zealand.

(MOOC) [2]. However, Chatbots in education are in their early stages [9]. As the number of students taking courses online is increasing, carrying out timely and meaningful assessments remains a challenging undertaking. For example, conducting assessment in MOOCs with a large number of students poses challenges to the instructor. In such a situation, it is necessary to rely on a computer-assisted assessment. In this paper, we explore the idea of an automated grading system involving the use of a Chatbot that asks students questions, requiring short written responses; graded by a Chatbot. We believe that combining Chatbots with a computer-assisted assessment can provide students with meaningful feedback and motivates engagement [4]. This research project aims to answer questions such as; to what extent can a Chatbot provide a consistent and useful assessment of short text answers compared to an assessment carried out by a human instructor? How can Chatbots be used to engage students in learning?

## 2 Proposed Method

We implemented a Chatbot that uses natural language understanding and text similarity to allocate grades or scores to short answer text provided by a student through a match with at least one correct answer (see Fig. 1). Cohen’s Kappa method [3] was applied to a small sample dataset to show that similarity matrix can be used to power the Chatbots to grade short texts. We are motivated by the research that shows that 70% of what students retain happens by what they say and write [7, 10]. We believe that implementing Chatbots to answer students’ questions and grade assignments, will give students the opportunity to practice for their exams, and ultimately enhance learning experience.



**Fig. 1.** The methodology for an automatic short answer grading conversation agent.

## 3 Experiment

The experiment was conducted at a University in Nigeria. First, questions were generated on the subject of Network Security and emailed to 15 students. Students then provided answers in written text format. The teacher independently graded the answers comparing scores to the standard answer. We used tfidf cosine similarity to compute the scores. Also, we used an online semantic text analytics as a service (Dandelion API), trained using neural networks on a vast

bank of questions and answers dataset. The results from both methods were combined to generate grades to the answers. Finally, we used Cohen's Kappa to find out the level of inter-rater agreement between the human instructor and the automatic short answer grading Chatbot.

The following is the question that was asked by the instructor, the standard answer and two random answers selected from the answers that students' provided.

**Question:** What is Network Security?

**Standard Answer:** Network security is the process of protecting underlying network infrastructure from unauthorized access, misuse, malfunction, modification, destruction, or improper disclosure using physical and software preventive measures, so as to create a secured networked environment.

**Student3:** Network security is the level to which a network is safe from unauthorized use.

**Student11:** Network Security are steps taken to ensure the integrity of a network.

## 4 Result

Results show that our automated grading system can perform relatively well, especially when two or more grading techniques are combined compared to a human instructor. Applying Cohen's Kappa measure [3], to calculate the level of agreement in Table 1, we arrived at a kappa value of 0.6. Hence, we can say that there is a good inter-rater agreement level between the Instructor and the automated grading Chatbot.

**Table 1.** Cohen's Kappa agreement measure between the Instructor and the Chatbot using a combination of tfidf cosine similarity and dandelion neural network API

Instructor					
Chatbot	Scores	1	2	3	Total
1	5	0	0	5	
2	1	5	2	8	
3	0	1	1	2	
Total	6	6	3	15	

Cohen's Kappa formula:

$$k = \frac{n_\alpha - n_\epsilon}{n - n_\epsilon} = 0.6 \quad (1)$$

where  $k$  is kappa's value,  $n$  is the number of students,  $n_\alpha$  is the number of agreements and  $n_\epsilon$  is number of agreements due to chance.

## 5 Discussion, Conclusion and Future Work

We explored the idea of automating short text grading using a Chatbot to scale this process as well as making it interactive and engaging for students. Our preliminary results of the experiment suggest that a combination of tfidf cosine similarity with the dandelion neural network API gave a better outcome in grading students. While we are motivated to explore the idea of an automated grading system in large classes and courses such as MOOCs, more work needs to be done to improve the general scope of the kinds of questions Chatbots can answer and its accuracy level. In the future, we plan to carry out large-scale experiments to validate work presented in this paper.

## References

1. Alsied, S.M., Ibrahim, N.W., Pathan, M.M.: Errors analysis of libyan EFL learners' written essays at Sebha University (2018)
2. Bollweg, L., Kurzke, M., Shahriar, K.A., Weber, P.: When robots talk-improving the scalability of practical assignments in MOOCs using chatbots. In: EdMedia+ Innovate Learning, pp. 1455–1464. Association for the Advancement of Computing in Education (AACE), June 2018
3. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Measur.* **20**(1), 37–46 (1960)
4. Eicher, B., Polepeddi, L., Goel, A.: Jill watson doesn't care if you're pregnant: grounding AI ethics in empirical studies. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 88–94. ACM, December 2018
5. Goel, A.K., Polepeddi, L.: Jill watson: a virtual teaching assistant for online education. Report, Georgia Institute of Technology (2016)
6. Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., Linn, M.C.: Validation of automated scoring of science assessments. *J. Res. Sci. Teach.* **53**(2), 215–233 (2016)
7. Molenda, M.: Cone of experience. Educational technology: An encyclopedia (2003)
8. Sam, A.H., et al.: Very-short-answer questions: reliability, discrimination and acceptability. *Med. Educ.* **52**(4), 447–455 (2018)
9. Sandoval, Z.V.: Design and implementation of a chatbot in online highereducation settings. *Issues Inf. Syst.* **19**(4), 447–455 (2018)
10. Seels, B.: The relationship of media and ISD theory: the unrealized promise of dale's cone of experience (1997)
11. Shawar, B.A., Atwell, E.: Chatbots: are they really useful? In: LDV Forum, vol. 22, no. 1, pp. 29–49, January 2007
12. Sijimol, P., Varghese, S.M.: Short answer scoring system using neural networks (2018)
13. Somasundaran, S., Chodorow, M., Tetreault, J.: System and method for automated scoring of textual responses to picture-based items. U.S. Patent 9,959,776. Educational Testing Service (2018)
14. Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., Trausan-Matu, S.J.C.: Automated essay scoring in applied games problem in online training: reducing the teacher bandwidth. *Comput. Educ.* **123**, 212–224 (2018)
15. Yang, X., Huang, Y., Zhuang, F., Zhang, L., Yu, S.: Automatic Chinese short answer grading with deep autoencoder. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 399–404. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_75](https://doi.org/10.1007/978-3-319-93846-2_75)



# Evidence-Based Recommendation for Content Improvement Using Reinforcement Learning

Machi Shimmei<sup>(✉)</sup> and Noboru Matsuda

North Carolina State University, Raleigh, NC 27695, USA  
mshimme@ncsu.edu

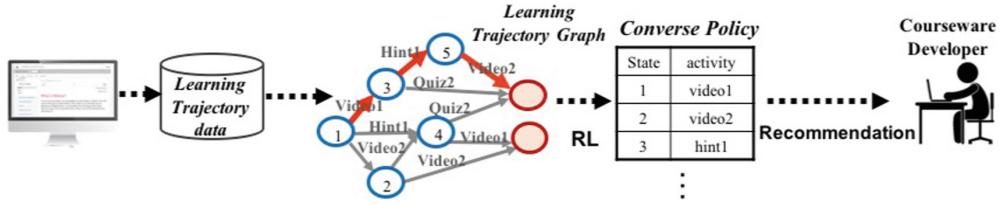
**Abstract.** One of the most challenging issues for online-courseware engineering is to maintain the quality of instructional elements. However, it is hard to know how each instructional element on the courseware contributes to students' learning. To address this challenge, an evidence-based learning-engineering method for validating the quality of instructional elements on online courseware is proposed. Students' learning trajectories on particular online courseware and their final learning outcomes are consolidated into a state transition graph. The value iteration technique is applied to compute the worst actions taken (a converse policy) to yield the least successful learning. We hypothesize that the converse policy reflects the quality of instructional elements. As a proof of concept, this paper describes an evaluation study where we simulated online learning data on three hypothetical pieces of online courseware. The result showed that our method can detect more than a half of the ineffective instructional elements on three types of courseware containing various ratios of ineffective instructional elements.

**Keywords:** Learning-engineering · Self-improving online courseware · Reinforcement learning

## 1 Introduction

Building a practical online courseware is extremely costly. It requires extensive knowledge and expertise in theories of learning and teaching [1]. On the other hand, the demand for effective online courseware has been increasing [2]. Therefore, developing a technological assistance to iteratively improve courseware is a critical need. In this paper, we propose the RAFINE method as a step towards an automated evidence-based learning engineering. Our ultimate goal is to develop self-improving online courseware that automatically detect ineffective parts and fix them. Some work has been done to automate process in learning engineering. Learning Factor Transfer Analysis, for example, is used to automatically detect linked model of domain [3]. Automated grading and adaptive intervention have been studied as well [4, 5]. RAFINE especially focuses on evidence-based validation of courseware content and detects ineffective instructional elements on online courseware. Figure 1 shows an overview of the method. Given students' learning data, RAFINE provides courseware developers with recommendation on instructional elements that need refinements. As a future work, we

will also focus on other aspects of learning engineering, such as an automated refinement of instructional elements.



**Fig. 1.** An overview of the Rafine method.

## 2 RAFINE Method

The unit of analysis of the RAFINE method is an instructional element that constitutes online courseware: (1) videos, (2) formative assessments (aka quizzes), and (3) hint messages associated with formative assessments. We assume a presence of a skill model that contains a set of skills each representing a unit of knowledge that students have to learn, aka knowledge components [6]. A single application of the RAFINE method identifies ineffective instructional elements relative to a particular skill. To simplify explanations, we assume that there is only one skill in our target online courseware. Figure 1 shows how RAFINE detects ineffective instructional elements. We first collect learning trajectories from students who learned on the online courseware. A learning trajectory is defined as a chronological record of learning activities that shows instructional elements taken by a particular student. All of the learning trajectories in a given log data are consolidated into a single learning trajectory graph (LTG). In the LTG, states represent learning status and edges represent learning activities taken that caused a change in learning status. Value iteration technique is then applied to compute the worst learning activity at each state in the LTG to achieve the predefined learning goal.

We define a learning status for student  $i$  at time  $T$  as an intermediate state of learning represented as a pair of Action History and Mastery Level. Action History is a binary vector showing which instructional elements student  $i$  has taken by time  $T$ . Note that in the LTG, student and time data are abstracted. In an LTG, the states where the value of the Mastery Level is greater than a threshold (which is usually 0.85) are called terminal states. All outgoing edges at terminal states are discarded.

### 2.1 Rewards and Converse Policy

A reward value of a particular state depends on the Mastery Level both at a current,  $ml(s)$ , and a successor state,  $ml(s')$ . A reward at the state  $s$  becomes the greatest (0.95) when the successor state is a terminal state.

In general, a policy suggests an action to be taken in a certain state to maximize the value function [7]. However, for the purpose of RAFINE, we need to know which instructional elements should not be taken—i.e., we need to know which action has the least expected reward. Therefore, the action that minimizes the value function needs to

be identified. We shall call this policy the *converse policy*. Given the reward function  $R$ , a value function for state  $s$  is updated through the value iteration as follows:

$$V(s) \leftarrow \min_{a \in A(s)} \sum_{s' \in S} T(s, a, s')(R(s, a, s') + \gamma V(s'))$$

Where  $S$  is a set of all states in a given LTG, and  $A(s)$  shows a set of actions appearing in outgoing edges at state  $s$ . The discount factor  $\gamma$  is arbitrarily set to be 0.9.

A transition model  $T(s, a, s')$  is derived from the learning trajectory data collected from actual students. It shows the ratio of transitions from  $s$  to  $s'$  among all transitions when action  $a$  is taken at  $s$  in the given learning trajectories. After the value function converged, the action that minimizes  $V(s)$  is identified as converse policy.

### 3 Evaluation Study

Our hypothesis is that those instructional elements that frequently appear as a converse policy across different states are likely to be ineffective and hence should be revised. To test this hypothesis, we conducted an evaluation study with hypothetical learning trajectories generated by simulated students. Although any instructional element can be selected as a converse policy, in the current study, we had RAFINE exclude assessment quizzes when making a refinement recommendation from a converse policy with an assumption that other known quantitative methods, e.g. item response theory [8], could be used to evaluate the quality of assessment items.

**Data:** Three instances of online courseware were created to control the “quality” of courseware with varying ratios of a number of effective instructional elements to all instructional elements on the courseware. All instructional elements on the mock courseware (9 videos and 9 hints total) except assessment quizzes were coded as either effective or ineffective. The high-quality courseware (H) had a 8:1 split (8 effective video/hint); the moderate-quality (M) had a 4:5 split; and the low-quality (L) had a 1:8 split. The student’s latent proficiency that indicates a probability of answering a quiz correctly is simulated with a logistic regression model.

Simulated students’ learning trajectories were randomly generated. For each quality of courseware, 100 instances of mock courseware were created with 1,000 simulated students. Each of the learning trajectory datasets was then converted into an LTG. In an LTG, Action History was encoded as a 27-bit binary vector (3 types of instructional elements, 9 each); and the Mastery Level is a decimal number (a multiple of 0.05). The latent proficiency described above was used as an estimate for Mastery Level (instead of actually applying a student model technique). For each of the 300 LTG’s, the value iteration technique was applied to compute a converse policy. As a result, 300 sets of converse policy were created, each suggesting which instructional elements were ineffective on the corresponding online courseware.

**Results:** We first tested if the frequency of being selected as a converse policy can be used as a criterion to detect ineffective instructional elements among the converse policy. The average frequency of each instructional element being selected as a

converse policy was computed by aggregating frequency values across 100 datasets. On average, each *ineffective* instructional element was selected as a converse policy 28.2 times in L, 30.6 in M, and 33.0 in H per dataset whereas each *effective* instructional element was selected 8.6 times in L, 10.0 in M, and 11.5 in H. The difference between ineffective and effective instructional elements was statistically significant for all three qualities of courseware: for L,  $t(99) = 84.67, p < 0.05$ ; for M,  $t(99) = 98.18, p < 0.05$ ; for H,  $t(99) = 37.71, p < 0.05$ . *These results suggest that frequency can be used as a filter to indicate ineffective instructional elements among a converse policy.* This implies that we should be able to find a frequency cut-off to determine which instructional elements must be classified as ineffective. We shall call this heuristic as the *frequency heuristic*. We therefore compared two different cut-off thresholds—mean  $\pm$  standard deviation ( $M \pm SD$ ). The mean and the standard deviation of the frequency that individual instructional elements were selected as a converse policy were computed. Those instructional elements that appeared as a converse policy more than the cut-off are considered as ineffective. Table 1 compares precision, recall, and F1 ( $2 * \text{precision} * \text{recall}/(\text{precision} + \text{recall})$ ) scores for two different cut-off thresholds crossed with the quality of the courseware (L, M, and H). The table shows that when the quality of courseware is low (L) to moderate (M), the  $M - SD$  cut-off yields better F1 score than the  $M + SD$  cut-off. However, when the quality of courseware is high (H), the  $M + SD$  cut-off outperforms  $M - SD$ . This implies that *at the beginning of the iterative courseware engineering, the  $M - SD$  cut-off is better, but as the courseware gets improved, the  $M + SD$  cut-off should be used.* We would want to detect as many inefficient instructional elements as possible even at a cost of false positives.

**Table 1.** A Comparison for Precision, Recall and F1 of the frequency cut-off;  $M \pm SD$

Courseware	Precision		Recall		F1	
	M-SD	M+SD	M-SD	M+SD	M-SD	M+SD
L	0.98	1.00	0.93	0.24	L	<b>0.96</b>
M	0.60	1.00	1.0	0.41	M	<b>0.75</b>
H	0.12	0.90	1.0	0.51	H	0.20
						<b>0.65</b>

## 4 Conclusion

We found that the worst policy (the converse policy) computed from students' learning trajectories graph reflects effectiveness of the instructional elements and the frequency heuristic has a high potential to detect ineffective instructional elements on online courseware. The proposed method, RAFINE, provides online courseware developers with an evidence-based recommendation to iteratively improve the courseware content. For a future study, it is crucial to measure the actual effectiveness of the proposed method in authentic learning settings.

## References

1. Fishman, B., et al.: Creating a framework for research on systemic technology innovations. *J. Learn. Sci.* **13**(1), 43–76 (2004)
2. Shapiro, H.B., et al.: Understanding the massive open online course (MOOC) student experience: an examination of attitudes, motivations, and barriers. *Comput. Educ.* **110**, 35–50 (2017)
3. Pavlik Jr, P.I., Cen, H., Koedinger, K.R.: Learning factors transfer analysis: using learning curve analysis to automatically generate domain models. *Online Submission* (2009)
4. Teusner, R., Hille, T., Staubitz, T.: Effects of automated interventions in programming assignments: evidence from a field experiment. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale. ACM (2018)
5. Sarwate, A., et al.: Grading at scale in earsketch. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale. ACM (2018)
6. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cognit. Sci.* **36**, 757–798 (2012)
7. Wiering, M., van Otterloeds, M.: Reinforcement Learning: State-of-the-Art 2012. ALO, vol. 12. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-27645-3>
8. Baker, F.: The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, College Park (2001)



# A Virtual Counselor for Genetic Risk Communication

Shuo Zhou<sup>(✉)</sup> and Timothy Bickmore

Northeastern University, Boston, MA 02115, USA  
[{zhous06,bickmore}@ccs.neu.edu](mailto:{zhous06,bickmore}@ccs.neu.edu)

**Abstract.** A pedagogical embodied conversational agent that plays the role of a genetic counselor is being developed to improve individuals' comprehension of genetic risks related to hereditary cancers. Genetic risk communication is increasingly important for disease prevention and treatment, yet many individuals lack the basic health literacy and numeracy required to understand this information. The virtual genetic counselor will address the challenges in communicating complex genetic risks by dynamically adapting its teaching strategies to an individual's knowledge state.

**Keywords:** Embodied conversational agent · Pedagogical agent · Intelligent tutoring system · Genetic counseling · Genetic risk communication · Health literacy · Health numeracy

## 1 Introduction

Genetic risk communication is increasingly important for disease prevention and treatment [1]. Understanding and acting on genetic risk information can be very difficult due to the complexity of the information. This can be especially difficult for the one third of U.S. adults with limited health literacy [2]—the ability to find, read, and act on written health information—and the one half of U.S. adults with limited numeracy [3]. Specifically, health numeracy is defined as the ability to access, understand, communicate, and act on numerical, graphical, biostatistical, and probabilistic health information needed to make effective health decisions [4]. Genetic risks are usually communicated by a genetic counselor, but many individuals cannot meet with genetic counselors due to logistical barriers exacerbated by a shortage of genetic counselors [5]. In my dissertation research, I am developing a pedagogical embodied conversational agent (ECA) that plays the role of a genetic counselor, to improve individuals' comprehension of genetic risks related to hereditary cancers.

Successful and effective communication of genetic risks is the key to appropriate health decision making and behavior change. Individuals who are at risk of hereditary cancers may need to make major decisions regarding genetic tests, medications, screening procedures, or preventive surgeries, based on understanding and analysis of complex risk information [6]. Communicating genetic risks is a challenging task, as it often involves conveying complex genetic concepts, uncertainty and randomness, and a large amount of numerical information, including relative risks, absolute risks, probabilities, and frequencies [1, 7, 8]. Because of this, communication of risks, specifically

numerical risk information, has been studied in health communication intensively over the past decades [6, 9, 10].

Health literacy and numeracy play crucial roles in genetic risk communication. Individuals with limited health literacy and numeracy often have less knowledge and lower comprehension of risks, and more difficulty using health information to make informed decisions [7, 11]. In particular, health numeracy crucially affects individuals' abilities to interpret complex graphs, assess risks, and determine preference of treatment based on risks [4]. In order to address the above challenges, the virtual genetic counselor will use techniques from the field of intelligent tutoring systems to dynamically adapt to an individual's comprehension, as well as his/her health literacy and numeracy levels.

## 2 Related Work

Intelligent tutoring system techniques are especially important to the task of genetic risk communication, as a significant portion of genetic counseling sessions are spent educating patients about genetics, hereditary risks, and risk-reducing behaviors. Prior research in genetic risk communication suggests that a genetic counselor should always gauge and confirm the patient's comprehension, and make adjustments in his/her explanations [6, 8], similar to what a human tutor usually does.

The AutoTutor system [12–14] developed by Graesser and colleagues, is most relevant to the pedagogical agent I'm developing. AutoTutor is an animated conversational agent that helps students construct answers to deep reasoning questions through dialogues. AutoTutor can hold a conversation with the learner in natural language, simulating the discourse patterns and pedagogical strategies of a human tutor. A recent version of AutoTutor is able to construct a cognitive model of students' knowledge levels by analyzing their typed or spoken answers, dynamically tailoring the interaction based on an individual student's development [15].

BRCA Gist is the closest ITS system to the one I am developing. This system, developed by Wolfe et al. [16, 17], is a web-based intelligent tutor built on the Shareable Knowledge Objects (SKO) platform, similar to AutoTutor. BRCA Gist teaches women general concepts related to breast cancer risks, using natural-language dialogues. However, the BRCA Gist does not provide the specific risk rates or recommendations that a woman would receive during a real genetic counseling session. Previously, we videotaped genetic counseling sessions, developed a prototype for explaining genetic risks and genetic testing to cancer patients, and tested the system with patients at Dana-Farber Cancer Institute. The virtual genetic counselor I am developing emulates these observed genetic counseling sessions, and provides genetic risk information tailored to the user's knowledge state.

## 3 Prior Work in Embodied Conversational Agents (ECA)

Past research in our group has demonstrated that embodied conversational agents can work effectively as health educators and health counselors [18, 19]. ECAs are animated computer characters designed to simulate face-to-face interaction between an individual

and a human counselor, using speech, facial expressions, hand gestures, and other non-verbal behaviors. ECAs are capable of expressing empathy [20], and they have been shown to be very effective with individuals of limited health literacy [18, 19]. For example, Wang et al. [21] developed a virtual genetic counselor capable of documenting family history with patients, and demonstrated the system highly feasible and effective by evaluating with 70 participants from an underserved patient population. However, this system did not focus on educating patients about genetic risk information.

Previously, we developed a conversational agent system capable of explaining complex medical documents such as clinical trial consent forms [22], and found that participants were more satisfied with the process when the agent's pedagogical contents were tailored based on their knowledge. In addition to pedagogical systems, I've developed a virtual counselor that provides alcohol misuse screening and brief intervention to U.S. veteran patients (Fig. 1). Preliminary results from the clinical trial demonstrated that veterans were able and willing to disclose to the agent about their alcohol use [23]. I've also developed a virtual nurse for care transition intervention with veteran patients (Fig. 1). A clinical trial evaluating this virtual nurse system is currently ongoing.



**Fig. 1.** Left: a virtual counselor for alcohol misuse. Right: a virtual hospital discharge nurse.

## 4 Proposed Work

The proposed virtual genetic counselor will educate its users about genetic risks related to a type of hereditary cancer, dynamically adapting to their knowledge state. The system will consist of four main components based on the standard architecture of an intelligent tutoring system [24]. The first component is a domain knowledge model, which contains the basic concepts and facts required to teach genetic risk information, including concepts of genetics, risks related to hereditary cancers, underlying implications of numerical risk information, potential health outcomes, and recommended risk-reducing behaviors. The agent will use both narrative explanations and applicable

graphics when explaining genetic risks. The second component is a student model that keeps track of the user's knowledge state, by periodically asking quiz questions, as well as the user's health literacy and numeracy captured at enrollment. The third component is a pedagogical module that dynamically chooses its teaching strategies and explanation methods based on user characteristics, the discourse context, and the system's assessment of the user's current state. The agent's pedagogical dialogues will be designed based on the actual genetic counseling sessions we've previously videotaped, the principles recommended in the field of genetic risk communication, as well as successful intelligent tutoring tactics including providing goal setting hints, providing short immediate feedback, and explaining errors, etc. The last component is an agent interface that simulates face-to-face counseling between the user and the virtual counselor.

I plan to evaluate the proposed system in a between-subjects study, comparing the adaptive agent with an agent that only provides a fixed amount of information not tailored to individual characteristics. The virtual counselor will educate users about genetic risks related to hereditary breast cancer. The main outcome measure is participants' comprehension of genetic risk information, assessed using a knowledge test. Other outcome measures include participants' satisfaction with the counseling experience, their working alliance with the agent, as well as their behavioral intention of following the recommended guidelines for breast cancer screening.

As with our prior systems, the agent will speak using a speech synthesizer, synchronized with a variety of nonverbal behaviors generated using BEAT [25], including facial displays of emotions, head nods for acknowledgment, hand gestures for emphasis, gaze shifts to signal turn-taking, and body posture shifts to signal topic changes. Users will be able to converse with the agent by selecting utterance options from a multiple-choice menu on the screen, updated at each turn of the conversation.

The proposed work will contribute to the AIED community, by applying intelligent tutoring techniques to simulate face-to-face genetic counseling, and to improve individuals' comprehension of complex genetic risk information. In particular, the proposed pedagogical agent system aims to address the current challenges in communicating genetic risks to individuals with low health literacy and numeracy.

## References

1. Lea, D.H., Kaphingst, K.A., Bowen, D., Lipkus, I., Hadley, D.W.: Communicating genetic and genomic information: health literacy and numeracy considerations. *Publ. Health Genomics* **14**, 279–289 (2011)
2. Kutner, M., Greenberg, E., Jin, Y., Paulsen, C.: *The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy*. National Center for Education Statistics, Washington, DC (2006)
3. Kutner, M., Greenberg, E., Baer, J.: *A First Look at the Literacy of America's Adults in the 21st Century*. National Center for Education Statistics, Washington, DC (2006)
4. Golbeck, A.L., Ahlers-Schmidt, C.R., Paschal, A.M., Dismuke, S.E.: A definition and operational framework for health numeracy. *Am. J. Prev. Med.* **29**, 375–376 (2005)
5. Peterson, E.B., et al.: Communication of cancer-related genetic and genomic information: a landscape analysis of reviews. *Transl. Behav. Med.* **8**, 59–70 (2018)

6. O'Doherty, K., Suthers, G.K.: Risky communication: pitfalls in counseling about risk, and how to avoid them. *J. Genet. Couns.* **16**, 409–417 (2007)
7. Ancker, J.S., Kaufman, D.: Rethinking health numeracy: a multidisciplinary literature review. *J. Am. Med. Inform. Assoc.* **14**, 713–721 (2007)
8. Apter, A.J., et al.: Numeracy and communication with patients: they are counting on us. *J. Gen. Intern. Med.* **23**, 2117–2124 (2008)
9. Lautenbach, D.M., Christensen, K.D., Sparks, J.A., Green, R.C.: Communicating genetic risk information for common disorders in the era of genomic medicine. *Annu. Rev. Genomics Hum. Genet.* **14**, 491–513 (2013)
10. Visschers, V.H.M., Meertens, R.M., Passchier, W.W.F., de Vries, N.N.K.: Probability information in risk communication: a review of the research literature. *Risk Anal.* **29**, 267–287 (2009)
11. Peters, E., Hibbard, J., Slovic, P., Dieckmann, N.: Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Aff.* **26**, 741–748 (2007)
12. Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A.: AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Trans. Educ.* **48**, 612–618 (2005)
13. Graesser, A.C., et al.: AutoTutor: a tutor with dialogue in natural language. *Behav. Res. Methods Instrum. Comput.* **36**, 180–192 (2004)
14. Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R.: AutoTutor: a simulation of a human tutor. *Cognit. Syst. Res.* **1**, 35–51 (1999)
15. D'Mello, S., Graesser, A.C.: AutoTutor and affective AutoTutor: learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.* **2**, 23 (2012)
16. Wolfe, C.R., et al.: Efficacy of a web-based intelligent tutoring system for communicating genetic risk of breast cancer: a fuzzy-trace theory approach. *Med. Decis. Mak.* **35**, 46–59 (2015)
17. Widmer, C.L., Wolfe, C.R., Reyna, V.F., Cedillos-Whynott, E.M., Brust-Renck, P.G., Weil, A.M.: Tutorial dialogues and gist explanations of genetic breast cancer risk. *Behav. Res. Methods* **47**, 632–648 (2015)
18. Bickmore, T.W., et al.: Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *J. Health Commun.* **15**, 197–210 (2010)
19. Bickmore, T., Pfeifer, L., Jack, B.: Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), pp. 1265–1274. ACM (2009)
20. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Interact.* **12**, 293–327 (2005)
21. Wang, C., et al.: Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. *Genet. Med.* **17**, 822 (2015)
22. Bickmore, T., Utami, D., Zhou, S., Sidner, C., Quintiliani, L., Paasche-Orlow, M.K.: Automated explanation of research informed consent by virtual agents. In: Brinkman, W.-P., Broekens, J., Heylen, D. (eds.) IVA 2015. LNCS (LNAI), vol. 9238, pp. 260–269. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-21996-7\\_26](https://doi.org/10.1007/978-3-319-21996-7_26)
23. Zhou, S., et al.: A relational agent for alcohol misuse screening and intervention in primary care. In: CHI 2017 Workshop on Interactive Systems in Healthcare (WISH) (2017)
24. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* **86**, 42–78 (2016)
25. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477–486. ACM (2001)

# **Industry Papers**



# A Multimodal Alerting System for Online Class Quality Assurance

Jiahao Chen, Hang Li, Wenxin Wang, Wenbiao Ding, Gale Yan Huang,  
and Zitao Liu<sup>(✉)</sup>

TAL AI Lab, TAL Education Group,  
DanLing SOHO, No. 6 DanLing Street, Beijing, China  
{chenjiahao,lihang4,wangwenxin2,dingwenbiao,  
galehuang,liuzitao}@100tal.com

**Abstract.** Online 1 on 1 class is created for more personalized learning experience. It demands a large number of teaching resources, which are scarce in China. To alleviate this problem, we build a platform (marketplace), i.e., *Dahai* to allow college students from top Chinese universities to register as part-time instructors for the online 1 on 1 classes. To warn the unqualified instructors and ensure the overall education quality, we build a monitoring and alerting system by utilizing multimodal information from the online environment. Our system mainly consists of two key components: banned word detector and class quality predictor. The system performance is demonstrated both offline and online. By conducting experimental evaluation of real-world online courses, we are able to achieve 74.3% alerting accuracy in our production environment.

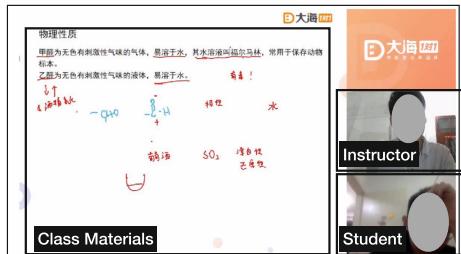
**Keywords:** Multimodal learning · Online class · Quality assurance

## 1 Introduction

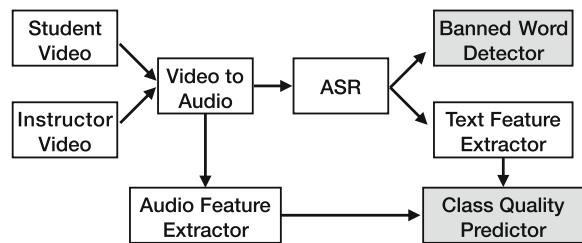
With the recent development of technology such as digital video processing and live streaming, there has been a steady increase in the number of students enrolling online courses worldwide [2]. Online 1 on 1 class is created to offer more personalized education experience. Both students and instructors are able to choose their out-class available time slots and have the class anywhere. To better allocate education resources in China, we create an online learning platform, i.e., *Dahai* (<http://www.dahai.com>) with two distinct types of participants representing supply (instructors) and demand (students). On Dahai platform, instructors are senior college students from top Chinese universities and students come to Dahai for online tutoring. Once the study plan agreement is reached, the matched student and instructor start online courses in Dahai's virtual classroom via live streaming. Dahai provides a wide range of online teaching tools to enable better teaching performance and interactions. Figure 1(a) shows the 1 on 1 learning environment provided by Dahai. Example industries include accommodation (Airbnb), ride sharing (Uber, Lyft, DiDi), online shops (Etsy,

Taobao), etc. Without a doubt, quality assurance for these types of marketplaces need to satisfy both supply and demand sides of the ecosystem in order to grow and prosper [1].

Allowing college students to be tutoring instructors<sup>1</sup> is a double-edged sword. On one hand, it greatly alleviates the problem of imbalanced teaching resources in China. However, on the other, part-time instructors may not have enough teaching experience. Some unprofessional behaviors may lead to low class quality and inferior learning performance. Being an online education platform, Dahai is responsible for its class quality. The most common way to alleviate this problem is to allow students to give ratings for the online classes and detect low-quality classes by utilizing ratings. However, such approaches usually fail in online K-12 education since K-12 students rarely give responsible ratings. For example, a student may give a 5-star rating to an instructor teaching video games. Therefore, we build a multimodal alerting system to automatically monitor the quality of each class in Dahai.



(a) Dahai online course scenario illustration. The virtual classroom consists of three panels: class materials, student and instructor.



(b) Overview of our multimodal alerting system. The gray boxes indicate the key components in our system. ASR is short for automatic speech recognition.

**Fig. 1.** Dahai online course scenario illustration and an overview of our multimodal alerting system. Both student's and instructor's faces are hided by gray circles due to the privacy issue.

## 2 The Multimodal Alerting System

Multimodal information of the entire course is stored in the online learning environment. When a class is finished in Dahai, both the student and the instructor videos are passed to the backend for alerting and monitoring analysis. First, we extract audio tracks from videos. We transcribe the teaching conversations by using an automatic speech recognition (ASR) system. After that, we apply the banned word detector to scan all the contents to ensure there is no misbehavior happened. Second, we extract both the linguistic and prosodic features and build a logistic regression predictor to automatically evaluate the overall online 1 on 1 course quality. The entire workflow of our alerting system is illustrated in Fig. 1(b).

<sup>1</sup> Tutoring instructors have to pass a series of interviews and training exercises before teaching the class.

## 2.1 Banned Word Detector

Instructors may thoughtlessly speak out swearing or insulting words or phrases. These words are referred to as *banned words* and are definitely not allowed to appear in the class. However, banned words may happen in many scenarios. For example, instructors may lose their patience when students couldn't response after given many hints. Another example, instructors may speak their casual mantras during the class accidentally, which contains banned words. Besides, there are also cases that the teaching environment is very noisy and the banned words appear from the background. As an education platform, we must assure a cyberbully-free and positive learning environment. Therefore, we develop the detector to take charge of banned word monitoring. We build the banned word detector by the following two steps:

Step 1. We construct a banned word bank that covers all possible banned words and their variants. In Chinese, the smallest semantic unit is character instead of word. A word is made up of several characters. This leads to more linguistic variants. To tackle this problem, we first pre-define a seed set of banned words and then expand the seed set by finding the nearest neighbor words from the gigantic Chinese Internet corpus. The nearest neighbor search is conducted in the pre-trained Chinese word embedding space. The word embeddings are learned by directional skip-gram, which explicitly distinguishes left and right context in Chinese [4]. After this expansion, we end up with a banned word bank with more than 3000 banned words.

Step 2. We detect banned words by applying several fuzzy matching rules and heuristics. The matching procedure is challenging because of the recognition errors in the ASR transcriptions. To address this issue, we first write fuzzy regular expressions to retrieve banned word candidates. Our fuzzy regular expressions match not only the Chinese words but the romanization of the Chinese characters based on their pronunciation, i.e., Pinyin [5]. After that, we conduct Chinese word segmentation on the candidate words and their corresponding contexts. The segmentation process takes account into the semantic meaning of each candidate and eliminate false positive candidates.

Here, we list a few classes caught by our detector in Table 1.

**Table 1.** Examples of classes caught by the banned word detector.

Examples	Instructor speech snippets with banned words
Class#1	Read it again. <b>Fuck</b> , can't you remember these two sentence?
Class#2	<b>Damn it</b> . I knew it. You didn't read the paper
Class#3	(Background noises) Come. Come. There is a group of <b>idiots</b>

## 2.2 Class Quality Predictor

Besides catching the class with banned words, we are responsible for the overall quality of the class. The 1 on 1 online class is more like a black box that only happens between the instructor and the student. First, majority of parents have no time to watch their kids during the class, which makes no pressure from the demand side in this online marketplace. Second, students wouldn't tell the truth about the class quality. For example, we caught one class that the instructor spent the entire class talking about a mobile game, which makes the student highly satisfied. Third, one of the largest advantages of 1 on 1 class is that instructors are able to frequently interact with students. Students have many chances to ask questions and talk about their own thoughts. However, due to the lack of teaching experience, some instructors may still keep using the traditional offline teaching paradigm. There are barely any interactions and instructors talk for 60 min without stops.

Therefore, we build an automated quality predictor to monitor all the online courses on Dahai. We extract linguistic features from the ASR transcriptions and prosodic features from audio tracks. The linguistic features include the number of characters, words, and sentences, the number of class subject related words, etc. The prosodic features include signal energy, loudness, mel-frequency cepstral coefficients (MFCC) [3], etc. We asked our teaching professionals to annotate 972 positive (good) courses and 219 negative (bad) courses. We use 80% of them for training our logistic regression classifier and use the rest for testing purpose. We evaluate the effectiveness of linguistic and prosodic features respectively. We report accuracy, precision, recall and F1 score of the quality prediction performance in Table 2. As we can see, both two types of features are very important to the quality prediction and the combination of both yields to the best results.

**Table 2.** Offline experimental results of class quality prediction.

Features	Accuracy	Precision	Recall	F1 score
Linguistic only	0.897	0.899	0.986	0.940
Prosodic only	0.949	0.944	0.997	0.970
Linguistic + Prosodic	0.954	0.949	0.997	0.972

## 2.3 Online System Performance

We deployed our monitoring and alerting system online. We set a few alerts based on the results of banned word detector and class quality predictor. Once the alarms are fired, we have operation staffs to watch the playback videos to conduct the final judgments. After comparing the staffs' ratings with our system's altering results, we achieve 74.3% accuracy in system alerting.

### 3 Conclusion and Future Work

In this paper, we presented our monitoring and alerting system for online 1 on 1 classes. By using the multimodal information, we are able to not only find misbehaviors in the online courses but measure the class quality. With the banned word detector and the class quality predictor, we are able to achieve 74.3% accuracy in our online production system. In the future, we plan to explore information from the class materials panel and improve the alerting performance as well.

### References

1. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: a survey. *J. Netw. Comput. Appl.* **68**, 90–113 (2016)
2. ChinaEducationResources: The largest education system in the world is going online (2012). <http://www.chinaeducationresources.com/s/OurMarket.asp>. Accessed 5 Feb 2019
3. Rabiner, L.R., Gold, B.: Theory and Application of Digital Signal Processing, 777 p. Prentice-Hall, Inc., Englewood Cliffs (1975)
4. Song, Y., Shi, S., Li, J., Zhang, H.: Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Short Papers, vol. 2, pp. 175–180 (2018)
5. Wikipedia: Pinyin (2019). <https://en.wikipedia.org/wiki/Pinyin/>. Accessed 7 Feb 2019



# Leveraging Cognitive Science and Artificial Intelligence to Save Lives

Matthew Jensen Hays<sup>1</sup>(✉), Aaron Richard Glick<sup>1</sup>, and H. Chad Lane<sup>2</sup>

<sup>1</sup> Amplifire, Boulder, CO 80301, USA

[mhays@amplifire.com](mailto:mhays@amplifire.com)

<sup>2</sup> University of Illinois Urbana-Champaign, Champaign, IL 61801, USA

**Abstract.** Medical error is the third-leading cause of death in the United States, just behind heart disease and cancer. We describe a software platform used to train healthcare workers to prevent their errors. The platform (Amplifire) harnesses artificial intelligence and principles of cognitive psychology. Amplifire's AI continuously decides whether and when to require additional learning events, provide corrective and metacognitive feedback, and/or deliver self-regulatory guidance for the learner (e.g., "slow down"). Amplifire was deployed to several thousand nurses at a large healthcare system in attempts to reduce the rate of two types of hospital-acquired infections. The result was a 48% reduction in central-line-associated bloodstream infections (CLABSI) and a 32% reduction in catheter-associated urinary-tract infections (CAUTI). These findings demonstrate the effectiveness of using cognitive science along with AI in an e-learning platform.

**Keywords:** Artificial intelligence · CAUTI · CLABSI · Cognitive science · Confidence · Feedback · Healthcare · Metacognition · Training

## 1 Introduction

Artificial intelligence (AI) is widely used in education because it can substantially enhance learning. Successful intelligent tutoring systems (ITSs) incorporate AI at various stages of the learning process in order to promote all facets of the Learner-Instructor-Knowledge triangle [1]. For example, the Andes Tutor leverages Bayesian-network solution maps to provide customized feedback as the student solves physics problems, while consequently improving the instruction provided by updating the probabilities of the Bayesian networks [2]. Other systems (e.g., ALEKS) determine a student's knowledge state and progress only to concepts for which the student has sufficient prerequisite knowledge [3]. Guru uses an animated tutor that integrates tutorials, collaborative dialogue, and direct instruction into a life-like user interface [4].

Although AI has permeated education, largely through ITSs, findings from cognitive psychology and other learning sciences have gained less traction in ITS research and the classroom. One reason may be that some cognitive phenomena are counter-intuitive; how learners, teachers, and even researchers think learning should work is not always how it actually works [5]. For example, the testing effect is the finding that retrieving information from memory is much more powerful than being re-exposed to

the information (e.g., by re-reading; [6]). But classrooms in 2019 still rely heavily on watching videos, sitting through lectures, and reading chapters. Even ITSs often use testing exclusively for assessment purposes (although there are exceptions, e.g., [7]).

We describe an e-learning platform—Amplifire—that uses AI and incorporates findings from cognitive science to optimize learning. Amplifire is designed to be content-agnostic. It has helped typical and non-traditional students perform better on exams, trained call-center employees to provide better customer service, and helped helicopter pilots earn recertification. Below, we review how Amplifire shapes the learner experience with AI and cognitive science, and we report on the reductions in CAUTI and CLABSI after nurses at a large healthcare system were trained in Amplifire.

## 2 AI-Directed Cognitive Science

Amplifire begins by asking questions in a variety of formats (multiple-choice, select-all, matching, interactive). This approach is beneficial even if the learner couldn't possibly provide the correct response to the question [8, 9]. Attempting to answer questions is perhaps the most powerful way to gain knowledge and skills [6], even if the generated answers are incorrect [10].

When responding to questions in Amplifire, learners indicate their confidence in their responses, making them consider the question more carefully [11] and improving their memory for the material [12]. This cognitive benefit only obtains when answers and confidence are considered simultaneously [13], a process Amplifire has patented. Learners in Amplifire click an answer once to indicate partial confidence or twice to indicate certainty. They can also click “I don’t know yet.”

After submitting a response, learners receive immediate feedback on whether their response was correct. Metacognitive feedback guides learners to understand whether they have been under- or overconfident [14]. Amplifire’s AI also determines whether and when to provide *self-regulatory feedback*, which is focused on correcting learner behavior in the platform. For example, a learner might be told to “make sure to read the question carefully” if they answer in less time than it would take to read the question.

Corrective feedback for a given item is provided after a delay, which enhances learning [15]. Amplifire’s AI optimizes this delay by considering information collected about the learner (e.g., their estimated ability), the content being learned (e.g., the item’s estimated difficulty), and the learner’s response to that particular item (e.g., how long the learner spent reading the prompt). The corrective feedback takes the form of elaborative explanation [16] and, when appropriate, worked examples [17]. The rationale behind the correct response is provided and the error the learner made is explained (e.g., miscalculation, buggy knowledge, etc.).

Amplifire does not provide corrective feedback after full-confidence correct responses because doing so does not improve retention [18]. Learners’ time is therefore better spent on more productive activities [19]. Corrective feedback is, however, provided after partial-confidence correct responses [20], and is especially powerful in cases of confidently held misinformation [21].

For problems or conceptual questions on which learners were not both fully confident and correct, Amplifire repeatedly tests the learner until its AI has determined that they have reached a mastery state. These repeated attempts profoundly improve the learner's long-term retention of the material [22]. Amplifire's AI considers learner, content, and response data in order to determine the optimal delay between successive attempts on a concept. This delay harnesses the spacing effect, which is the finding that distributing learning over time is more effective than massing it together [23]. Amplifire targets the point in the learner's forgetting curve where a retrieval attempt is difficult but not impossible [24, 25].

Altogether, Amplifire leverages AI and cognitive science to optimize the learner's time spent mastering the material, promote long-term retention and transfer to related tasks, and maintain learner engagement.

### 3 Application and Efficacy in Healthcare

Amplifire has partnered with career-focused online universities, GED providers, and other educational institutions that support non-traditional and underserved student populations. More recently, Amplifire has expanded into healthcare training. Medical errors are responsible for more than 250,000 fatalities in the United States annually, making them the third-leading cause of death [26]. More than half of all medical errors are attributed to the "cognitive failures" of healthcare professionals [27]. Amplifire was used at a large healthcare system to combat the cognitive failures that contribute to two hospital-acquired infections: CLABSI and CAUTI. The healthcare system made no other changes to policies, training, or available resources during this period; all effects were attributed to Amplifire.

#### 3.1 Central-Line-Associated Bloodstream Infections (CLABSI)

A central line is a thin tube (catheter) placed into a large vein. Central lines are used to administer nutrition or medication (e.g., drugs for chemotherapy), and to monitor central blood pressure during acute care. When a healthcare provider inadvertently contaminates the equipment or the insertion site, the patient can develop a central-line-associated bloodstream infection (CLABSI). The incidence of CLABSI is expressed in terms of the number of infections caused for every 1,000 days that patients had central lines ("CLABSI per 1,000 line-days").

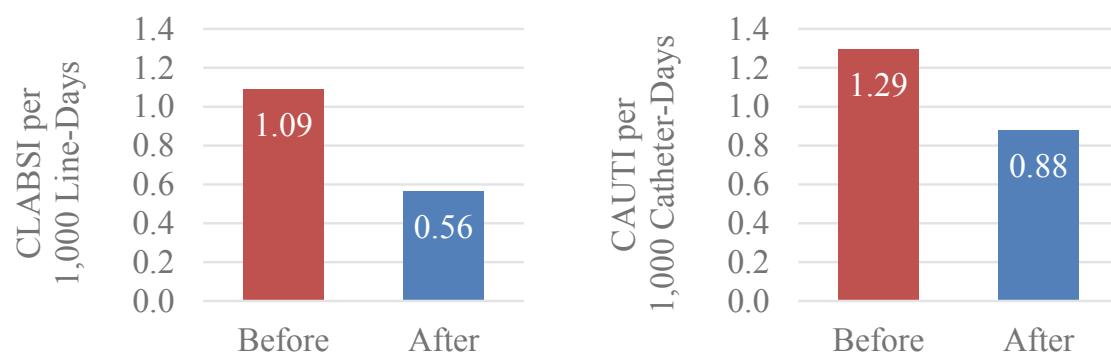
All central-line-attending nurses at a large healthcare system ( $N = 3,712$ ) were trained in Amplifire. The results are displayed in the left panel of Fig. 1. In the 28 months before training, there were 1.09 CLABSI per 1,000 line-days. In the seven months after training, there were 0.56 CLABSI per 1,000 line-days—a reduction of 48%. An exact Poisson test indicated a statistically significant reduction in the CLABSI rate after training:  $p = .00014$ . Given CLABSI's mortality rate of 25%, this reduction should save approximately 13 lives per year at this health system [28].

### 3.2 Catheter-Associated Urinary-Tract Infections (CAUTI)

A urinary catheter is a thin tube inserted into the bladder via the urethra. An indwelling catheter remains in the urethra and bladder for continuous drainage of urine and monitoring of urine output during acute care. As with central lines, healthcare workers' mistakes can contaminate the catheter and cause a catheter-associated urinary tract infection (CAUTI). Similar to CLABSI, the incidence of CAUTI is expressed in terms of the number of infections caused for every 1,000 days that patients were catheterized ("CAUTI per 1,000 catheter-days").

Urinary-catheter-attending nurses ( $N = 4,512$ ) at the same healthcare system were trained in Amplifire. The results are displayed in the right panel of Fig. 1. In the 28 months before training, there were 1.29 CAUTI per 1,000 catheter-days. In the seven months after training, there were 0.88 CAUTI per 1,000 catheter-days—a reduction of 32%. An exact Poisson test indicated a statistically significant reduction in the CAUTI rate after training:  $p = .01363$ .

Although both CLABSI and CAUTI were reliably reduced, the smaller magnitude of the CAUTI reduction may be attributable to two factors. First, only nurses interact with central lines, but both nurses and technicians interact with urinary catheters; part of the caregiver population was not trained on CAUTI. Second, the CAUTI course did not employ any multimedia [29]. A revised and improved CAUTI course will be distributed to both nurses and technicians in the coming months.



**Fig. 1.** Rates of CLABSI (left) and CAUTI (right) before and after Amplifire training.

## 4 Conclusion

Amplifire is an online learning platform that relies on principles of cognitive science. By allowing AI to determine how best to leverage many of those principles in real time, Amplifire delivers individually optimized learning in a wide variety of domains. Its test-focused approach improves learners' ability to retrieve information from memory. Its emphasis on confidence creates an additional dimension of learner introspection and understanding. Its multiple types of scaffolded feedback ensure that difficulty, engagement, and remediation are managed effectively, while also supporting metacognition and self-regulation. Amplifire's ability to substantially reduce medical error demonstrates the power of cognitive science working hand in hand with AI.

## References

1. Kapros, E., Koutsombogera, M. (eds.): Designing for the User Experience in Learning Systems. HCIS, pp. 1–11. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-94794-5>
2. Gertner, A., Conati, C., VanLehn, K.: Procedural help in Andes: generating hints using a Bayesian network student model. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, AAAI, vol. 98, pp. 106–111. The MIT Press, Cambridge (1998)
3. Falmagne, J.-C., Cosyn, E., Doignon, J.-P., Thiéry, N.: The assessment of knowledge, in theory and in practice. In: Missaoui, R., Schmidt, J. (eds.) ICFCA 2006. LNCS (LNAI), vol. 3874, pp. 61–79. Springer, Heidelberg (2006). [https://doi.org/10.1007/11671404\\_4](https://doi.org/10.1007/11671404_4)
4. Olney, A.M., et al.: Guru: a computer tutor that models expert human tutors. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 256–261. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-30950-2\\_32](https://doi.org/10.1007/978-3-642-30950-2_32)
5. Bjork, R.: Memory and metamemory considerations in the training of human beings. In: Metcalfe, J., Shimamura, A. (eds.) Metacognition: Knowing About Knowing, pp. 185–205. MIT Press, Cambridge (1994)
6. Roediger, H., Karpicke, J.: Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* **17**, 249–255 (2006)
7. Bhatnagar, S., Lasry, N., Desmarais, M., Charles, E.: DALITE: asynchronous peer instruction for MOOCs. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) EC-TEL 2016. LNCS, vol. 9891, pp. 505–508. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45153-4\\_50](https://doi.org/10.1007/978-3-319-45153-4_50)
8. Bransford, J., Schwartz, D.: Rethinking transfer: a simple proposal with multiple implications. In: Iran-Nejad, A., Pearson, P. (eds.) Review of Research in Education, vol. 24, pp. 61–100. American Educational Research Association, Washington, DC (1999)
9. Hays, M., Kornell, N., Bjork, R.: When and why a failed test potentiates the effectiveness of subsequent study. *J. Exp. Psychol. Learn. Mem. Cogn.* **39**, 290–296 (2013)
10. Pashler, H., Rohrer, D., Cepeda, N., Carpenter, S.: Enhancing learning and retarding forgetting: choices and consequences. *Psychon. Bull. Rev.* **14**, 187–193 (2007)
11. Bruno, J.: Using MCW-APM test scoring to evaluate economics curricula. *J. Econ. Educ.* **20** (1), 5–22 (1989)
12. Soderstrom, N., Clark, C., Halamish, V., Bjork, E.: Judgments of learning as memory modifiers. *J. Exp. Psychol. Learn. Mem. Cogn.* **41**, 553–558 (2015)
13. Sparck, E., Bjork, E., Bjork, R.: On the learning benefits of confidence-weighted testing. In: Cognitive Research: Principles and Implications, vol. 1 (2016)
14. Azevedo, R.: Computer environments as metacognitive tools for enhancing learning. *Educ. Psychol.* **40**, 193–197 (2010)
15. Butler, A., Karpicke, J., Roediger, H.: The effect of type and timing of feedback on learning from multiple-choice tests. *J. Exp. Psychol. Appl.* **13**, 273–281 (2007)
16. Shute, V., Hansen, E., Almond, R.: An Assessment for Learning System Called ACED: Designing for Learning Effectiveness and Accessibility. ETS Research Report Series, pp. 1–45 (2007)
17. McLaren, B.M., van Gog, T., Ganoe, C., Yaron, D., Karabinos, M.: Worked examples are more efficient for learning than high-assistance instructional software. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 710–713. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_98](https://doi.org/10.1007/978-3-319-19773-9_98)
18. Pashler, H., Cepeda, N., Wixted, J., Rohrer, D.: When does feedback facilitate learning of words? *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 3–8 (2005)

19. Hays, M., Kornell, N., Bjork, R.: The costs and benefits of providing feedback during learning. *Psychon. Bull. Rev.* **17**, 797–801 (2010)
20. Butler, A., Karpicke, J., Roediger, H.: Corrective a metacognitive error: feedback increases retention of low-confidence correct responses. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 918–928 (2008)
21. Butterfield, B., Metcalfe, J.: Errors committed with high confidence are hyper-corrected. *J. Exp. Psychol. Learn. Mem. Cogn.* **27**, 1491–1494 (2001)
22. Karpicke, J., Roediger, H.: Repeated retrieval during learning is the key to long-term retention. *J. Mem. Lang.* **57**, 151–162 (2007)
23. Cepeda, N., Pashler, H., Vul, E., Wixted, J., Rohrer, D.: Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol. Bull.* **132**, 354–380 (2006)
24. Landauer, T., Bjork, R.: Optimal rehearsal patterns and name learning. In: Gruneberg, M., Morris, P., Sykes, R. (eds.) *Practical Aspects of Memory*, pp. 625–632. Academic Press, London (1978)
25. Hays, M., Darrell, J., Smith, C.: The forgetting curve(s) of 710,870 real-world learners. Poster Presented at the American Psychological Association 124th Annual Convention, Denver, CO, USA (2016)
26. Makary, M., Daniel, M.: Medical error—the third leading cause of death in the US. *BMJ* **353**, 2139 (2016)
27. Joint Commission: Patient safety. Joint Commission Online (2015). [https://www.jointcommission.org/assets/1/23/jconline\\_April\\_29\\_15.pdf](https://www.jointcommission.org/assets/1/23/jconline_April_29_15.pdf). Accessed 8 Feb 2019
28. CDC: Vital Signs: Central Line–Associated Blood Stream Infections — United States, 2001, 2008, and 2009. *Morbidity and Mortality Weekly Report (MMWR)*, vol. 60, pp. 1–6 (2011)
29. Mayer, R.: Using multimedia for e-learning. *J. Comput. Assist. Learn.* **33**, 403–423 (2017)



# A Task-Oriented Dialogue System for Moral Education

Yan Peng, Penghe Chen<sup>(✉)</sup>, Yu Lu, Qinggang Meng, Qi Xu,  
and Shengquan Yu

Advanced Innovation Center for Future Education, School of Educational Technology, Beijing Normal University, Beijing 100875, China  
[chenpenghe@bnu.edu.cn](mailto:chenpenghe@bnu.edu.cn)

**Abstract.** We present a novel and practical dialogue system specifically designed for teachers and parents to solve students' problems in moral education. Guided by the case-based reasoning theory, we collect the high-quality cases and teaching strategies from heterogeneous sources, and then construct the dedicated knowledge graph to manage the large volume of information in this domain. By leveraging on the latest natural language processing techniques, we finally implement a task-oriented dialogue system to precisely understand user's problem and subsequently recommend possible solutions. We show the great promise of the system for K-12 education and demonstrate how the system solves the problem raised by the teacher for moral education.

**Keywords:** Moral education · Dialogue system · Knowledge graph

## 1 Introduction

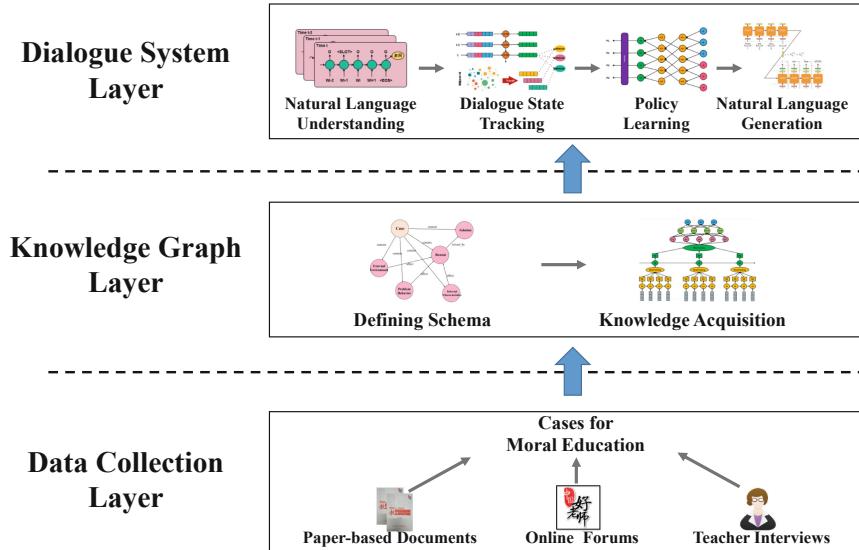
Moral education in general refers to guiding students to correct improper psychology and behavior (e.g., steal) and develop noble values (e.g., honesty), which is vital to the healthy growth of children. It is commonly seen that young students exhibit the improper behaviors in both school and home environment, like fighting with classmates, impolite with parents and egoism. Timely and properly correction of such behaviors imposes a challenging task for both teachers and parents. Specifically, moral education can be regarded as an interdisciplinary field that requires the knowledge from psychology, pedagogy and sociology, and obviously most teachers do not have the expertise in all such domains and thus cannot help their students solving such problems in practice. Moreover, it is difficult for teachers to learn the structured and systematic knowledge from this domain, which leads to fulfilling the moral education even harder.

To address the above issues, we design and implement a task-oriented intelligent dialogue system specifically for solving the problems raised by teachers or parents in the moral education domain. We mainly adopt the case-based reasoning (CBR) theory [5] to conduct the system design, which emphasizes on utilizing the previous similar cases and experiences to solve the current problem.

The CBR theory has been successfully used to guide the design of different intelligent systems for knowledge reasoning [8] and decision making [3]. Briefly speaking, we first collect the high-quality and heterogeneous data from moral education domain, including successful teaching and pedagogical cases from paper-based documents, online forums and teacher interviews. After that, we construct the dedicated knowledge graph by leveraging on the reasoning techniques in the CBR theory. With the built knowledge graph for moral education, we finally construct a task-oriented multi-round dialogue system that can effectively collect the desired information and consequently provide professional suggestions to teachers and parents for solving the problems in moral education.

## 2 System Design

As mentioned earlier, the main idea of the CBR theory is to utilize the previous similar cases and experiences to solve the current problem, which usually includes three key steps, namely *case collection* (gathering enough relevant cases), *case indexing* (properly organizing the collected cases for future reference), and *case processor* (understanding and recommending applicable cases and solutions). We adopt the similar design philosophy, and as shown in Fig. 1, our system mainly consists of three indispensable modules, namely data collection layer, knowledge graph layer and dialogue system layer. We will elaborate each one from the bottom to the top in this section.



**Fig. 1.** Three-layer architecture of the system

### 2.1 Data Collection Layer

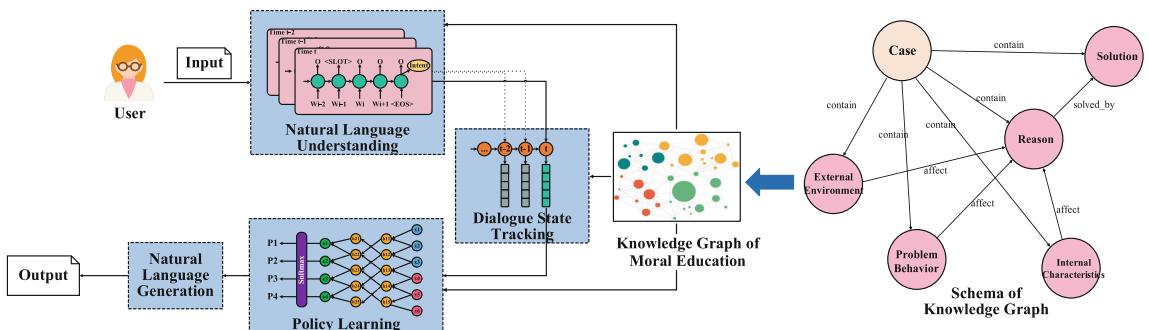
Data collection layer corresponds to the *case collection* in the CBR theory, as it mainly takes charge of collecting high-quality cases from heterogeneous sources

to form a rich case base about how to deal with students' improper behaviors. Due to the unique characteristics of moral education, many professional cases can be found in the paper-based documents that edited by the domain experts. Besides, discussions on the teacher's online forums also provide highly related cases for correcting students' improper behaviors. In addition, the interviews with the experienced teachers in this domain also supply valuable data for both experience collection and defining schema for the knowledge graph construction later. In order to obey ethic, all of data in this work will hide students' private information like name and only be allowed to use for research.

## 2.2 Knowledge Graph Layer

Knowledge graph layer constructs the domain knowledge graph of moral education, which not only implements the *case indexing* step of the CBR theory, but also establishes a systematic representation and knowledge structure for moral education. Simply speaking, it is responsible for building the dedicated knowledge graph for moral education, and We mainly complete the two main tasks: defining schema and knowledge acquisition. Defining schema requires revealing the key elements and their explicit relations in solving moral problems from the psychology, pedagogy and sociology perspectives. Drawn the experiences from the collected data and the related works [1, 6], we define three key affecting factors in the schema, including problem behavior [4], internal characteristics and external environment. Such three key factors can directly help to diagnose the potential reasons and accordingly suggest possible solutions with the reference cases. The defined schema is illustrated on the right side of Fig. 2, it has already been revised by experienced teachers and experts.

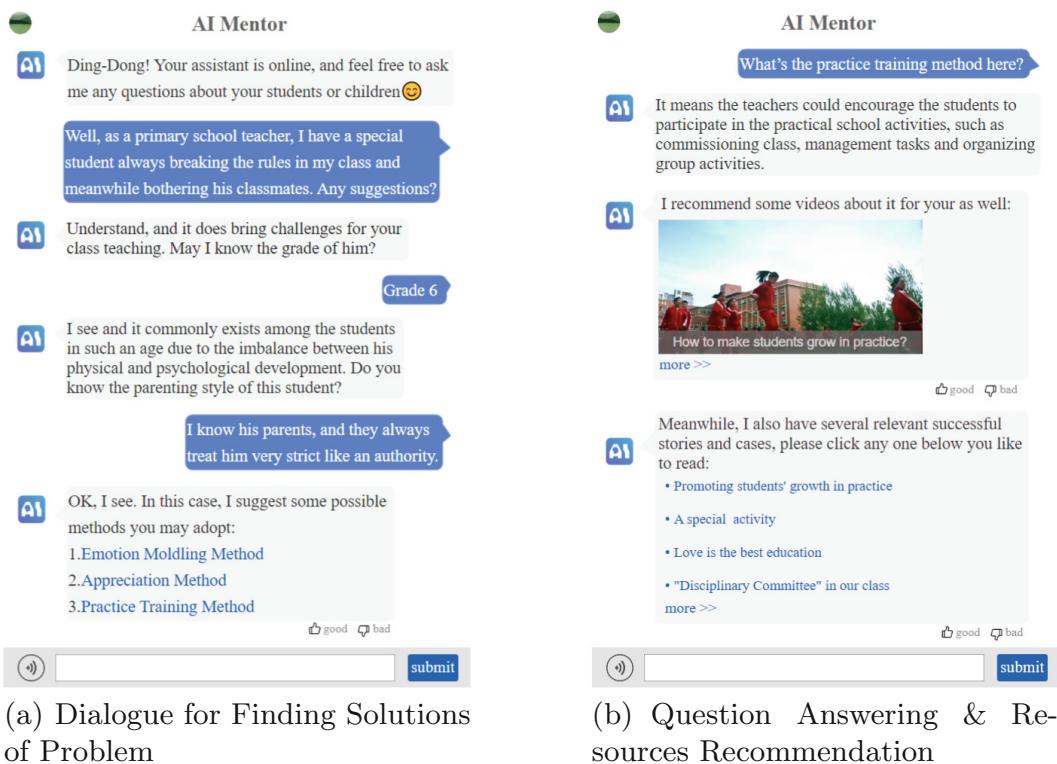
The knowledge acquisition, which converts new cases into a structured form defined by the above described schema to update the case base, is generally accomplished manually by the domain experts. Meanwhile, we are currently developing a multi-classification model to automate this process by leveraging on the natural language processing and deep learning techniques.



**Fig. 2.** Dialogue system with the knowledge graph schema

### 2.3 Dialogue System Layer

Dialogue system layer corresponds to the *case processor* in the CBR theory, as it works as the central processor to understand the key problem encountered by teachers or parents, and then recommending both possible solutions and closely relevant cases. Specifically, we adopt the pipeline approach to develop a task oriented dialogue system. As illustrated in Fig. 2, it consists of four modules: *Natural Language Understanding* module utilizes a hierarchical long short term memory (LSTM) network [10] to model the users' current and past utterances for properly understanding user's current intention and accomplishing the slot fitting task [7]; *Dialogue State Tracking* module adopts the LSTM network to derive the dialogue states and retrieve the similar cases from the embedded knowledge graph; *Policy Learning* module utilizes a multilayer perceptron (MLP) [9] network to decide the system next action (e.g., keep asking more questions or recommend solutions); *Natural Language Generation* module adopts a template-based approach [2] to generate system responses according to the determined system actions.



**Fig. 3.** A simplified demonstration of the system

### 3 System Demo

Figure 3 simply demonstrates how our system works with a teacher who is seeking for the help of handling his mischievous student. The entire conversation

essentially includes two parts: firstly, the system asks the teacher to acquire the necessary information, such as the problem behavior, grade and his parenting style, identifies potential reasons and then suggests several possible solutions (as shown in Fig. 3(a)); secondly, the system answers questions raised by the teacher and accordingly recommends the learning resources, including both the micro-lectures and relevant cases (as shown in Fig. 3(b)).

## 4 Conclusion and Implementation

In this work, we present a task-oriented dialogue system, which is specifically designed for teachers and parents to properly guide them in the moral education. By leveraging on the CBR theory, the system collects the relevant cases, constructs the dedicated knowledge graph, and eventually implements a dialogue system using the latest natural language processing techniques. Such a unique system solves a practical problem in the current K-12 education and has shown a high demand from different stakeholders. We are also working with a national-level teacher's platform to integrate our system as an online assistant for serving more than 160,000 teachers from 1100 local schools.

**Acknowledgment.** This research is partially supported by the National Natural Science Foundation of China (No. 61702039 and No. 61807003), the Humanities and Social Sciences Foundation of the Ministry of Education of China (No. 17YJCZH116) and the Fundamental Research Funds for the Central Universities.

## References

1. Bao, P., Jing, J., Jin, Y., Hu, X., Liu, B., Hu, M.: Trajectories and the influencing factors of behavior problems in preschool children: a longitudinal study in Guangzhou, China. *BMC Psychiatry* **16**(1), 178 (2016)
2. Cheyer, A., Guzzoni, D.: Method and apparatus for building an intelligent automated assistant. US Patent 9,501,741, 22 Nov 2016
3. Gómez-Vallejo, H., et al.: A case-based reasoning system for aiding detection and classification of nosocomial infections. *Decis. Support Syst.* **84**, 104–116 (2016)
4. Jessor, R., Jessor, S.L.: Problem behavior and psychosocial development: A longitudinal study of youth (1977)
5. Kolodner, J.L.: An introduction to case-based reasoning. *Artif. Intell. Rev.* **6**(1), 3–34 (1992)
6. Lee, J.R., Kim, G., Yi, Y., Song, S., Kim, J.: Classifying Korean children's behavioral problems and their influencing factors: a latent profile analysis. *Int. J. Child Care Educ. Policy* **11**(1), 6 (2017)
7. Li, X., Chen, Y.N., Li, L., Gao, J., Celikyilmaz, A.: End-to-end task-completion neural dialogue systems. arXiv preprint [arXiv:1703.01008](https://arxiv.org/abs/1703.01008) (2017)
8. Nikpour, H., Aamodt, A., Bach, K.: Bayesian-supported retrieval in BNCreek: a knowledge-intensive case-based reasoning system. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR 2018. LNCS (LNAI), vol. 11156, pp. 323–338. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01081-2\\_22](https://doi.org/10.1007/978-3-030-01081-2_22)

9. Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., Ettaouil, M.: Multilayer perceptron: architecture optimization and training. *IJIMAI* 4(1), 26–30 (2016)
10. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI, vol. 16, pp. 3776–3784 (2016)



# Leveraging Student Self-reports to Predict Learning Outcomes

Shaveen Singh<sup>(✉)</sup>

Faculty of Information Technology, Monash University, Melbourne, VIC, Australia  
[shaveen.singh@monash.edu](mailto:shaveen.singh@monash.edu)

**Abstract.** Academic performance is typically measured through assessments on standardised tests. However, considerably less is known about the relationship between students self-assessment (metacognition and affective states) captured during the reading process and their academic performance. This paper presents a preliminary analysis of data gathered during a blended course offering using student self-reports on learning material as predictor of their academic outcomes. The results point to the predictive potential of such self-reports and the potentially critical role of incorporating such student self-reports in learner modelling and for driving teaching interventions.

## 1 Introduction

In recent decades, researchers have stressed the need to develop assessment practices that not only measure what students have learned but also enhance learning [3,5]. This has led to a change in existing assessment practices, from summative assessment as the primary activity, towards emphasis on formative assessments. *Summative* assessment is cumulative, almost always graded, typically less frequent, and occur at the end of segments of instruction—whereas *formative* assessment is primarily aimed at educating and improving student performance and providing feedback to teachers or students to help students learn more effectively [4].

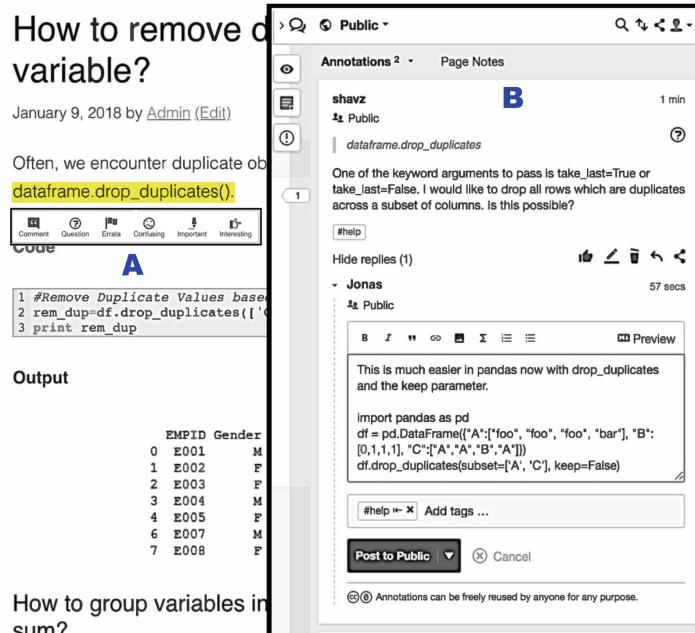
One way to effectively obtain formative feedback is by engaging students in their own assessments, for example through self-reporting on how they feel, what motivates them and what they have trouble learning [8,9]. Instructors in large courses, however, do not have the capacity to read or respond to every comment in discussion threads or feedback forms. This makes it difficult for them to identify students that need the most assistance. Prior studies have used crowd workers hired from systems such as Mechanical Turk [1,11] to manually tag students' posts with their affect. However, it is time-consuming and costly to have instructors or paid crowd workers annotate posts. Automatic identification or machine learning methods also struggle when the reflections are not rich, has misspellings or use poor vocabulary—making it difficult for the classifier to distinguish between more nuanced affective states, such as confusion, curiosity and so on [11].

To address this, we present a practical strategy for collecting nuanced affective states and experiences in learning material at scale. We expand on our previous work in [9], where we developed an annotation interface with options for students to include custom tags. Self-coding adds an additional step to the traditional annotation process of highlighting and commenting. It allows students to first engage with the context and then explicitly acknowledge their emotions or experiences before reflecting on it. The reason for choosing this method is because (1) student authors may be one of the best sources regarding their own affective state, (2) it provides researchers an easy and accurate way to acquire a labeled and contextual dataset and (3) the instructors are able to get formative learning feedback in a timely fashion. Determining these nuanced affective states and experiences have ramifications for instructors aiming to provide interventions for their students, as an instructor would likely have different responses for different types of reported encounters.

This paper presents a preliminary analysis of data generated using the annotation platform in a blended computer programming short-course conducted over 2 weeks at our university. We describe the self-reporting instrument used within the platform to facilitate contextual feedback from the students on key aspects of their experience with learning material. We then evaluate how this feedback relates to the potential risk of their attrition and/or poor performance in the course.

## 2 The Annotation Tool

Numerous studies [2, 10, 12] cite annotation as useful for learners but there is lack of investigation on how valuable this under-utilised data source is for predictive analytics. The annotation tool we adopted for this study has been extended from Hypothesis [6] and allows students to asynchronously annotate learning material such as readings, lecture notes and programming problem sets in a chat-like fashion. It captures the immediate student reactions to the educational documents while providing an avenue for learner-learner and learner-instructor communication and a space for voicing questions and concerns [7]. The



**Fig. 1.** Annotation tool showing (A) the popup dialog when a user selects the text and (B) the annotations in the righthand pane

student view of the course page is shown in Fig. 1. Figure 1 (*Panel A*) shows what the student sees after accessing the reading material and highlighting a specific passage on a page. Six different tagging options were made available to students to categorise their annotations. These options are *Comment*, *Important*, *Errata*, *Interesting*, *Confusing* and *Help*, as shown in Table 1. Upon selecting one of the six annotation categories a conversation window opens in the right hand pane where the student can pose a question or post a comment (Fig. 1 - *Panel B*).

Students can also add their own custom tags to their annotation and the corresponding text. These tags are searchable and can be used to connect with others working on the same topic or to follow a group's annotation activity across the material. Annotations are public by default but can be made private. If the annotation is public, other students can respond asynchronously to the conversation. Students can also upvote, share or flag annotations made by others.

### 3 Data Analysis and Results

A short-course on “Data Visualisation using Python” was conducted over 2 weeks. A total of 37 undergraduate and post-graduate students were enrolled in the course. The first week involved the preparatory phase where students were required to go over a 20-page study guide comprising of readings, coding instructions and quizzes made available via a content delivery platform as a HTML e-book. Students had to complete this task online on their own, at home, prior to attending a 3 h face-to-face session at the end of the second week. The choice to annotate was completely voluntary and no marks were awarded for making annotations, however, students were informed that the annotations made would be used to adapt the teaching focus in the face-to-face workshop session (at the end of Week 2). Fine-grained data regarding annotations, author, posting timestamp and counts was captured during the offering.

Over 307 annotations were made by 36 out of the 37 students during the course. In Table 1, we show the breakdown for each category of annotation. The average number of annotations made on a page was 15.3 with  $SD = 2.2$  and the average number of annotation posts per student was 8.57 ( $SD = 3.1$ ). The average number of interactions with an annotation by a student (upvote/view/read) stood at 24.6 ( $SD = 4.3$ ).

The Pearson correlations analysis reveal significant correlations between number of annotation posting with both their completion rate ( $r = 0.536, n = 36, p < 0.001$ ) and quiz performance ( $r = 0.503, n = 36, p < 0.001$ ). As part of regression analysis, we closely examine the potential of (i) the total number of annotations posted and (ii) the nature of annotation (percentage breakdown by annotation category) in predicting the completion rate and performance. The

**Table 1.** Breakdown of annotations

Category	Icon	Count	Percentage
Comment	“ ”	109	36%
Important	⚠	126	41%
Errata	🚩	23	7%
Interesting	↗	14	5%
Confusing	😊	23	7%
Help	?	12	4%

results show that total number of annotations per student alone explained 26.7% of the variance of the completion rate (Model C1). When we added the percentage breakdown of these annotations (categories), we find that we can predict a further 46% of the variability in student completion (Model C2). Model C3 shows that if we add the amount of time students spend reading we can predict an additional 4% (77%) of the variability in completion rate. An increase in the percentage of help annotations by one standard deviation (SD) decreases student completion rate by 0.487 of a SD ( $p < 0.01$ ). This statistically significant increase in variance suggests that categorization of annotations makes it a better predictor in this model. The results are summarised in Table 2.

**Table 2.** Regression coefficients and statistics

	Completion Rate (CR)			Quiz Performance (QP)		
	Model C1	Model C2	Model C3	Model P1	Model P2	Model P3
Num. of students (N)	36	36	36	36	36	36
$R^2$	0.267	0.727	0.767	0.253	0.561	0.569
Predictors	Standard coefficients			Standard coefficients		
Constant	-	-	-	-	-	-
Total annotations	0.536*	0.236	0.048	0.503*	0.371	0.311
% as Confusing		-0.475**	-0.345*		-0.318	-0.281
% as Errata		0.140	0.169		0.374	0.382
% as Help		-0.612**	-0.487**		-0.426**	-0.391
% as Important		-0.103	-0.144		-0.318	-0.328
% as Interesting		-0.337*	-0.240		-0.205	-1.780
Total Reading time			0.3944			0.116

\*\* $p < 0.01$ , \* $p < 0.05$ .

To study the relationship between annotations and quiz performance, we similarly built a series of linear regression models to predict students quiz performance (Table 2). We found that including the data regarding the proportion of each annotation category significantly improves the prediction of the model from 25.3% (Model P1) to 56.1% (Model P2). Adding total reading time (Model P3) to this model, however, only marginally improved the prediction (56.9%).

## 4 Discussion and Conclusion

This study demonstrates a streamlined process of students self-reporting their emotion, understanding and experience through annotations. The use of reading data to capture emotions, cognitive and meta-cognitive insights can be very promising, as by referring to these as feedback of students' learning, instructors can make decisions and intervene in a timely manner.

The proposed annotation platform provides a systematic and consistent way in which student's can self-report on their experience— something which can also

be used as an important learning and assessment tool. We showed that the categorization of annotations is a good predictor of the completion rate of students. The percentage (%) of confusing annotation is the dominant predictor in this model. Additionally, we found that the annotation categories also contributes towards explaining the variability in performance.

Although, this analysis uses a small number of students, the strong correlation between the annotation data with performance and completion rate is encouraging. Future work will involve using a larger cohort to conduct a rigorous investigation to possibly derive a *confusion threshold*, whereby if the ratio of confusion expressed is higher than the threshold, it may be the likely signal that the student would be at risk of attrition. Annotations also interestingly creates an atmosphere where students have the chance to provide contextual feedback to instructors for actioning improvements to the material written by the faculty.

Our next step us to run large-scale versions of our experiment in real-life university courses with enrolments in excess of 500 students for a whole semester to further test and validate the findings in this paper.

## References

1. Agrawal, A., Venkatraman, J., Leonard, S., Paepcke, A.: YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In: International Educational Data Mining Society (2015)
2. Behler, A.: E-readers in action. *Am. Libr.* **40**(10), 56–59 (2009)
3. Black, P., Wiliam, D.: Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* **92**(1), 81–90 (2010)
4. Dixson, D.D., Worrell, F.C.: Formative and summative assessment in the classroom. *Theory Pract.* **55**(2), 153–159 (2016)
5. Dolin, J., Black, P., Harlen, W., Tiberghien, A.: Exploring relations between formative and summative assessment. In: Dolin, J., Evans, R. (eds.) *Transforming Assessment*. CSER, vol. 4, pp. 53–80. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-63248-3\\_3](https://doi.org/10.1007/978-3-319-63248-3_3)
6. Hypothesis: Hypothesis - the internet, peer reviewed (2018). <https://web.hypothes.is/>
7. Pellet, J.-P., Parriaux, G., Overney, T.: A case study on the effect of using an anchored-discussion forum in a programming course. In: Pozdniakov, S.N., Dagienė, V. (eds.) *ISSEP 2018. LNCS*, vol. 11169, pp. 42–54. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-02750-6\\_4](https://doi.org/10.1007/978-3-030-02750-6_4)
8. Porayska-Pomsta, K., Mavrikis, M., Cukurova, M., Margeti, M., Samani, T.: Leveraging non-cognitive student self-reports to predict learning outcomes. In: Penstein Rosé, C. (ed.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 458–462. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_86](https://doi.org/10.1007/978-3-319-93846-2_86)
9. Singh, S., Meyer, B.: Using social annotations to augment the learning space and learner experience. In: *Proceedings of the 24th Annual ACM Conference on Innovation and Technology in Computer Science Education*. ACM (2019)
10. Thayer, A., Lee, C.P., Hwang, L.H., Sales, H., Sen, P., Dalal, N.: The imposition and superimposition of digital reading technology: the academic potential of e-readers. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2917–2926. ACM (2011)

11. Yang, D., Wen, M., Howley, I., Kraut, R., Rose, C.: Exploring the effect of confusion in discussion forums of massive open online courses. In: Proceedings of the Second (2015) ACM Conference on Learning@ scale, pp. 121–130. ACM (2015)
12. Young, J.R.: 6 lessons one campus learned about e-textbooks. Chronicle High. Educ. **55**(39) (2009)



# Toward a Scalable Learning Analytics Solution

Josine Verhagen<sup>(✉)</sup>, David Hatfield, and Dylan Arena

Kidaptive, Redwood City, USA

[josineverhagen@kidaptive.com](mailto:josineverhagen@kidaptive.com)

**Abstract.** Since its founding in 2011, Kidaptive has built customized models that provide adaptivity and/or personalization in online learning environments. We have supported adaptive game-based learning through rule-based and dynamic Bayesian psychometric models, and we have developed behavioral models for online learning and online test preparation environments based on learners' time management, answer behavior, and test scores. Our models are deployed on a scalable distributed-computing platform that has supported millions of learners, but the human expertise required to build custom models for every learning environment is not scalable. To address this limitation, we have recently been working toward an abstracted version of our psychometric and behavioral models, to be provided as an “out-of-the-box” product offering. This paper describes insights and challenges encountered in this process.

**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 powerful aspect of game-based learning environments is their potential to help 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). Proficiency-based adaptivity can support this type of learning by presenting challenges that align with learners' proficiency levels, which requires a valid assessment of each learner's proficiency.

Designing for proficiency-based adaptivity requires ensuring frequent measurements or other pieces of evidence to update a dynamic learner model. Learner proficiencies change over time, either because of interactions in the learning environment or because the learner is active in the world. To allow flexibility in our learner proficiency estimates, we have chosen 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 and Mislevy 1982). For this proficiency estimation to work, item characteristics such as the difficulty of the items presented to a learner must be known. This is difficult, 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. To accommodate the commercial infeasibility of calibrating (a random subset of) game challenges with a pilot sample of learners, we have developed calibration methods using a combination of initial “guesstimates” and creative empirical calibration and equation methods, many of which use an Elo-based algorithm (e.g. Pelánek 2016).

Our solution has provided scalable game-based assessment and personalization to millions of learners, but the process of building and validating the assessment models used in the 30+ games supported by our technology requires expert attention to the pedagogical goals, game mechanics, and interaction affordances of each supported game. The requirement of close attention by experts to the particularities of each context limits the scalability of this approach.

## 1.2 Supporting Productive Study Behavior

Learning curricula have moved online at a rapid and increasing pace. University courses have become MOOCs, publishers have converted their textbooks to interactive online courses, standardized-test preparation has shifted from workbooks to online programs, and tutors are now connecting with students virtually. Although replacing teacher instruction with videos is scalable, the motivational component of learner-teacher interaction may not be. Dropout rates are very high in MOOCs (e.g. Andres et al. 2018) and learners find creative ways to minimize effortful learning in nascent online learning environments (e.g. Baker et al. 2006).

For the past few years we have supported a Korean publisher that 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 developed dozens of models to support this partner: Bayesian models for working speed and learner ability relative to peers (considering relative item difficulty and question or study duration), which 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; cluster analyses on study behaviors to personalize recommendations for productive study habits; and score-prediction models to set realistic performance goals for each learner.

These models are delivering valuable insights to tens of thousands of tutors who support hundreds of thousands of learners, using a distributed-computing system that scales seamlessly to ingest millions or tens of millions of learning events daily. As with our game-based assessment learning, however, the development of these custom models has required considerable expert attention to this learning environment; this expert attention for focused analysis of a specific learning context is difficult to scale.

## 2 Towards Scalable Learning Analytics

We have briefly described the bespoke modeling work that has led to our current effort to provide a scalable (i.e., generalizable) learning-analytics solution. Elsewhere we have discussed the technical/evaluation details of that work (e.g. Verhagen and Arena 2018; Verhagen et al. 2015). Here we focus on our work to build an out-of-the-box offering that provides useful learning analytics for a variety of learning products.

### 2.1 Basic Learning Analytics

Although designers of learning products usually want their products to teach particular concepts or skills, they often have not thought explicitly about how to measure whether learners using their products are in fact learning. To do so means mapping 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 mapping process involves identifying both which questions measure which skills and which other activities might provide additional evidence of proficiency, mindset, and/or engagement. Once mapped, these events can be logged when they occur during learner activity.

Our new out-of-the-box effort provides a set of customer guidelines for mapping content to 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-, teacher-, or parent-facing reports about learner activity or to provide a dashboard giving the product owner an overview of what and how learners are doing in their product.

### 2.2 Insights from More Complex Models

Our next goal is to identify under which conditions the more complex psychometric and behavioral models we have developed for previous customers are feasible and valid, and to offer them to new customers when those conditions are met.

**Test Score Prediction:** In many learning environments, test scores are used as a proxy for learner proficiency. Machine learning (ML) 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 ML models of real-time test score prediction for students in test preparation environments that are scalable in the sense of working across different products. Although models customized to specific learning environments will necessarily result in better accuracy, we believe we can develop a standardized model that will do a good enough job to be useful in practice.

Recent studies suggest that a combination of performance- and behavior-related features are optimal for predicting standardized test scores based on interaction data from online learning environments (e.g. Ritter et al. 2013; Pardos et al. 2014; Feng and Roschelle 2016; Kostyuk et al. 2018; San Pedro et al. 2015). We are investigating several feature sets that will be available in most test-preparation learning environments: a set of features tracking performance on questions related to the various sub-domains 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).

**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 skills. 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.

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 which learners encounter which questions at what point in time. We aim to provide a set of recommendations for (approximate) linking of items across curricula and levels of adaptive learning environments. 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 et al. 2018; Klinkenberg et al. 2011; Pelánek 2016) is very efficient and provides mostly accurate item difficulties while accounting for learner abilities changing over time.

### 3 Conclusion

This paper has presented an overview of Kidaptive's effort to generalize the work it has done over the past seven years on learning analytics solutions for first-, second-, and third-party learning products to provide a learning analytics solution that can work out of the box with new learning environments: i.e., a scalable learning analytics solution.

### References

- Andres, J.M.L., Baker, R.S., Gašević, D., Siemens, G., Crossley, S.A., Joksimović, S.: 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 (2018)

- Baker, R.S.J.d., et al.: Adapting to when students game an intelligent tutoring system. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 392–401. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_39](https://doi.org/10.1007/11774303_39)
- Bock, R.D., Mislevy, R.J.: Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Measur.* **6**, 431–444 (1982)
- Brinkhuis, M.J., Savi, A.O., Hofman, A.D., Coomans, F., van der Maas, H.L., Maris, G.: Learning as it happens: a decade of analyzing and shaping a large-scale online learning system. *J. Learn. Anal.* **5**(2), 29–46 (2018)
- Feng, M., Roschelle, J.: Predicting students' standardized test scores using online homework. In: Proceedings of the Third ACM Conference on Learning@Scale, pp. 213–216. ACM (2016)
- Gee, J.P.: What Video Games Have to Teach Us About Learning and Literacy. Palgrave Macmillan, New York (2003)
- Klinkenberg, S., Straatemeier, M., van der Maas, H.L.: Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ.* **57**, 1813–1824 (2011)
- Kostyuk, V., Almeda, M.V., Baker, R.S.: 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 (2018)
- Pardos, Z.A., Baker, R.S., San Pedro, M., Gowda, S.M., Gowda, S.M.: Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. *J. Learn. Anal.* **1**, 107–128 (2014)
- Pelánek, R.: Applications of the Elo rating system in adaptive educational systems. *Comput. Educ.* **98**, 169–179 (2016)
- Rasch, G.: Probabilistic models for some intelligence and attainment tests. Danish Educational Research Institute, Copenhagen (1960)
- Ritter, S., Joshi, A., Fancsali, S., Nixon, T.: 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.: Exploring dynamical assessments of affect, behavior, and cognition and math state test achievement. In: International Educational Data Mining Society (2015)
- Shaffer, D.W.: How Computer Games Help Children Learn. Palgrave Macmillan, New York (2006)
- Verhagen, J., Arena, D.: A dynamic Bayesian network for a multidimensional longitudinal learner profile. In: Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, pp. 502–503, July 2018
- Verhagen, J., Hatfield, D., Watson, J., Liu, S., Arena, D.: Shapes and patterns of adaptive learning: an experiment. In: Proceedings of the Games, Learning, and Society Conference, vol. 11, pp. 248–254. ETC Press, Pittsburgh, June 2015



# Motivating Students to Ask More Questions

Yuan Wang<sup>1(✉)</sup>, Turner Bohlen<sup>2</sup>, Linda Elkins-Tanton<sup>1,2</sup>,  
and James Tanton<sup>2</sup>

<sup>1</sup> Arizona State University, Tempe, AZ 85287, USA

elle.wang@asu.edu

<sup>2</sup> Beagle Learning, Boston, MA 02210, USA

**Abstract.** Students don't ask enough questions in classrooms, in person or online. Asking questions has been found to be a critical skill toward developing critical thinking abilities, improve learning performance, as well as career development. However, 'how to ask productive questions' as a key skill, is not well studied. Therefore, the present paper introduces the question productivity index (QPI) and explores ways toward quantitatively and reliably measure student-generated questions.

**Keywords:** Question productivity · Critical thinking · Learning analytics

## 1 Background

### 1.1 Asking Questions Matters

It all starts with a good question! The popular saying is often heard in contexts of job interviews, formulating research questions, workplace discussions [1, 2], as well as classroom learning [3, 4]. Questioning, as a skill, has been found to have two major functions, one is learning, the other on liking. Recent research has shown that simply asking more questions can be an effective way to increase innovation, promote information exchanges, as well as building rapport and increase liking [1, 2]. This pattern was found in both in-person and online settings such as online chat rooms [11].

Despite various benefits of asking questions, research reveals that people do not ask enough questions or do not expect asking questions to be beneficial [2]. Not asking enough questions has been found to be a prevalent phenomenon in classrooms [3, 6–8]. Although question asking is generally encouraged and considered a sign of motivation, research showed that high-quality student-generated questions are scarce and many students rarely ask any questions [4].

### 1.2 Benefits of Asking Questions Needs to Be Explained

Challenges of asking questions overshadow its benefits [1, 7–10]. Students may have different reasons and concerns not to ask questions, such as to avoid being seen incompetent [2, 12], cultural origins [5], language proficiencies [6]. By contrast, benefits of asking questions are rarely stressed in classrooms [6]. Teachers may

welcome students to ask questions, yet skills and techniques of asking questions are rarely the center point of most classes. A vague understanding of why asking questions is beneficial, in itself, can prevent students from asking questions.

With the goal of helping students to ask more effective questions and increase learning performance, research is needed toward understanding (1) why asking more questions is good (2) what makes questions good.

Toward the above-mentioned goals, we proposed to develop a question productivity index and have expert teachers to rate students questions according to this index, with an aim to optimize neural nets that can automatically score student-generated questions.

## 2 Question Productivity Index (QPI)

As a pilot, 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 h (Table 1).

**Table 1.** Descriptions of the QPI dimension

Name	Descriptions
Relevance	How relevant is the question to the larger learning goal?
Scale	The question takes the class one reasonable step from their current knowledge
Articulation	The question is well-posed and uses good grammar
Overall	An overall score on the productivity of the question

## 3 Early Findings and Discussions

A multiple regression model was conducted to investigate if the three dimensions of question productivity while controlling word count and course types. The results (see Table 2) indicated that three dimensions of QPI, together with word count and whether the question was raised in a STEM course, explained 82% of the variance.

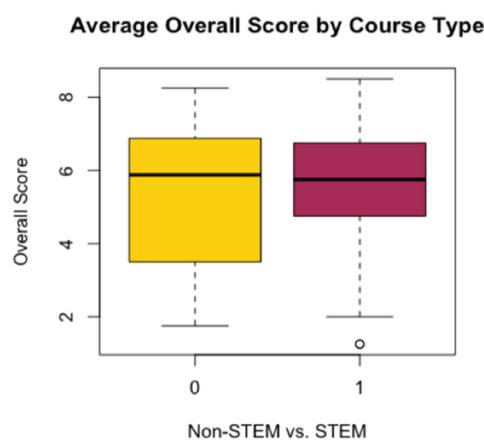
It was found that all three dimensions, each significantly predicted overall productivity scores, Scale ( $\beta = .15$ ,  $p < .001$ ), Articulation ( $\beta = .27$ ,  $p < .001$ ), Relevance ( $\beta = .62$ ,  $p < .001$ ). Relevance was found to be the strongest predictor among the three.

**Table 2.** Multiple regression to predict average overall score

Predictor	<i>b</i>	<i>b</i> 95% CI	$\beta$	<i>beta</i> 95% CI	$sr^2$	$sr^2$ 95% CI	<i>r</i>
(Intercept)	-0.41	[-1.13, 0.32]					
Scale	0.14	[0.06, 0.22]	0.15	[0.06, 0.24]	.02	[-.00, .04]	.45**
Articulation	0.25	[0.13, 0.38]	0.27	[0.13, 0.40]	.03	[-.00, .06]	.79**
Relevance	0.54	[0.43, 0.65]	0.62	[0.49, 0.75]	.16	[.09, .24]	.87**
Word Count	0.00	[-0.01, 0.01]	0.01	[-0.08, 0.10]	.00	[-.00, .00]	-.07
STEM	-0.03	[-0.44, 0.38]	-0.01	[-0.09, 0.08]	.00	[-.00, .00]	.01

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights.  $sr^2$  represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.  
 \*indicates  $p < .05$ . \*\*indicates  $p < .01$ .  $R^2 = .820$

Word count ( $\beta = .01$ ,  $p = .82$ ) and whether the question was raised in a STEM course ( $\beta = -.01$ ,  $p = .87$ ) did not show statistically significant relationship with the overall productivity score. This result is somewhat surprising since the content of question in a STEM subject can be different from those from a non-STEM subject, which may suggest that skills of asking questions may share more commonalities than differences. Further research is required to explain this null result (Fig. 1).

**Fig. 1.** Boxplot comparing average overall scores of questions between Non-STEM courses and STEM courses

#### 4 Creating Feasible Classroom Intervention and Future Directions

Following up, we plan to collect more student-generated questions across a diverse range of subject areas and increase the number of expert raters. Student demographic information such as gender, age, and language background will also be incorporated in

the model to investigate if students belonging to any specific demographic groups tend to have distinctive questioning patterns.

In the process of optimizing the QPI index, one of the instructors in the research team has used the QPI rubric as a classroom questioning guide to encourage students to ask questions while being mindful of the various dimensions of the questions. Research has been planned to also measure students' goal orientations [12] and grit [13], as well as student grades to explore their relationships with question productivity.

## References

1. Brooks, A.W., Gino, F., Schweitzer, M.E.: Smart people ask for (my) advice: seeking advice boosts perceptions of competence. *Manage. Sci.* **61**(6), 1421–1435 (2015)
2. Brooks, A.W. John, L.K.: The surprising power of questions: it goes far beyond exchanging information. *Harvard Business Review*, May/June 2018
3. Chin, C., Brown, D.E.: Student-generated questions: a meaningful aspect of learning in science. *Int. J. Sci. Educ.* **24**(5), 521–549 (2002)
4. Dillon, J.T.: The remedial status of student questioning. *J. Curriculum Stud.* **20**(3), 197–210 (1988)
5. Gresalfi, M., Martin, T., Hand, V., Greeno, J.: Constructing competence: an analysis of student participation in the activity systems of mathematics classrooms. *Educ. Stud. Math.* **70**(1), 49–70 (2009)
6. Good, T.L., Slavins, R.L., Harel, K.H., Emerson, H.: Student passivity: a study of question asking in K-12 classrooms. *Sociol. Educ.* **60**, 181–199 (1987)
7. Osborne, R., Wittrock, M.: The generative learning model and its implications for science education (1985)
8. Pizzini, E.L., Shepardson, D.P.: Student questioning in the presence of the teacher during problem solving in science. *School Sci. Math.* **91**(8), 348–352 (1991)
9. Ryan, A.M., Gheen, M.H., Midgley, C.: 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. *J. Educ. Psychol.* **90**(3), 528 (1998)
10. Watts, M., Alsop, S., Gould, G., Walsh, A.: Prompting teachers' constructive reflection: pupils' questions as critical incidents. *Int. J. Sci. Educ.* **19**(9), 1025–1037 (1997)
11. Huang, K., Yeomans, M., Brooks, A.W., Minson, J., Gino, F.: It doesn't hurt to ask: question-asking increases liking. *J. Pers. Soc. Psychol.* **113**(3), 430 (2017)
12. Pintrich, P.R.: The role of goal orientation in self-regulated learning. In: *Handbook of Self-regulation*, pp. 451–502 (2000)
13. Duckworth, A.L., Quinn, P.D.: Development and validation of the short grit scale (GRIT-S). *J. Pers. Assess.* **91**(2), 166–174 (2009)



# Towards Helping Teachers Select Optimal Content for Students

Xiaotian Zou<sup>1(✉)</sup>, Wei Ma<sup>2(✉)</sup>, Zhenjun Ma<sup>1(✉)</sup>,  
and Ryan S. Baker<sup>3(✉)</sup>

<sup>1</sup> Learnta Inc., 1460 Broadway, New York, NY 10036, USA

{zouxiaotian, will}@learnta.com

<sup>2</sup> Institute of Statistics and Big Data, Renmin University of China,  
59 Zhongguancun Street, Beijing 100872, China

mawei@ruc.edu.cn

<sup>3</sup> University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104, USA  
rybaker@upenn.edu

**Abstract.** In a personalized learning context, teachers decide which content to assign to students on the basis of data. However, it is not clear that simply providing teachers with data is sufficient to promote good instructional decisions. In this paper, we study data from an online learning platform that gives teachers data on student test performance and then allows them to decide which new skill students should work on. We then apply a knowledge graph algorithm to infer whether the content the teacher assigned the student is a skill that the student is ready to learn (i.e. the skill is within the student's Zone of Proximal Development), whether the student is not yet ready to learn the skill, or whether the student has already learned the skill. In this paper, we study how the teacher's decision of what skills or topics the student should work on correlate to the student's learning outcomes. We study this issue using logistic regression to compare whether students master more skills based on whether they are assigned ready-to-learn skills or unready-to-learn skills according to the knowledge graph. The results demonstrate that in both mathematics and English learning contexts, if the teacher selects skills which the student is assessed by the algorithm to be ready to learn, the student gains more mastery than if he or she is assigned skills he or she is not ready to learn. We conclude by proposing a visualization that more clearly surfaces the knowledge graph predictions to teachers.

**Keywords:** Instructional decision · Learning outcomes · Mastery · Knowledge graph · Ready-to-learn · Unready-to-learn · Zone of Proximal Development

## 1 Introduction

There has been considerable interest over the last decades in providing students with adaptable, personalized learning experiences and flexible content sequencing. However, there is more to the potential of AIED systems than just automated adaptivity. Increasingly, AIED systems also inform teacher decision-making [5], part of a broader trend to support data-driven decision-making by teachers.

However, data-driven decision making within a technologically rich medium will only be effective when the right data is clearly presented by teachers, and when teachers make the right teaching decisions. While there is increased interest in supporting teacher cognition and metacognition in the context of AIED systems [10] and creating better methods for informing teachers [5, 7], it remains unknown how effective teachers are at using the information they receive. As Earl and Katz [4] note, although many school districts have established large databases, teachers typically receive little guidance in terms of how to effectively use the data for differentiated instruction. In particular, how effective are teachers at selecting material to work from when given reports on student performance? In other words, even when the data are accessible to teachers, they still have difficulties in deciding what students need next, to phrase it in affective terms, to measure what contents fall in learner's zone of proximal development (ZPD).

We consider this in the specific context of selecting content in a learner's zone of proximal development (ZPD). A learner's ZPD represents the difference between what a learner can learn with assistance, and what he or she has already mastered without help [14]. As Vygotsky [14] notes, the term "proximal" means those skills a learner is "close" to mastering; a learning task assigned within this zone is likely to be learned effectively, and content outside the ZPD is likely to either be too difficult or too easy for the student. In its original formulation, ZPD is difficult to measure without intense one-on-one scaffolding, making it difficult for teachers to use it as a basis for instructional decisions, but Murray and Arroyo [13] propose that the ZPD can be measured by adaptive learning systems. As the first group of researchers who investigated how ZPD is measured in AIED systems, they proposed to categorize a learner's learning process into several states, using data such as task performance and the number of actions needed. ZPD was identified when learners' data demonstrated that they were appropriately challenged instead of being too bored or too confused. Inspired by their work, our current study proposes to use the knowledge graph as another potential tool for determining a student's ZPD in an adaptive learning system.

In this paper, we measure the ZPD using a knowledge graph and then use this measure to investigate whether teachers make good instructional decisions based on student performance data. Specifically, we investigate the impacts on student performance and mastery when they are assigned content inside or outside their assessed ZPD. Our hypothesis is that students will gain more mastery if assigned skills they are ready-to-learn (RtL) than if the teacher assigns unready-to-learn (UtL) skills. We conclude with ideas on how to communicate ZPD to teachers.

## 2 Method

### 2.1 Platform

In the current study, we use data from an online learning platform, Learnta, that gives teachers data on student test performance and then allows them to decide which content students should work on [1]. Learnta's knowledge graph maps content to a prerequisite structure, representing which content is necessary to know to learn content. A student's

mastery of each skill is assessed by Bayesian Knowledge Tracing (BKT) [3], determining whether the student has mastered a particular skill by predicting their latent knowledge. BKT has four parameters: the initial probability of knowing the skill -  $P(L_0)$ ; the probability of learning the skill each time it is encountered -  $P(T)$ ; the probability of making a mistake despite knowing the skill -  $P(S)$ , and the probability of guessing an unknown skill correctly -  $P(G)$ . Then the prerequisite structure is used to assess which content a student is ready-to-learn (RtL), defined as when the student has not yet mastered the skill but has mastered all of its prerequisites (i.e. the skill is within the student's ZPD), versus which content the student is unready-to-learn (UtL), i.e. not all prerequisites have been mastered. We investigate whether teachers given assessment data make effective instructional decisions, by seeing whether they assign materials that fall in the student's ZPD, and what the results are for learning.

## 2.2 Data Collection

In an English grammar learning, the topic titled “Pronoun and Noun”, used by 49 Learnta students, was randomly selected from the pool of topics. The math/Calculus topic “Integral Expression” was randomly selected from the pool of topics signed up for by the same group of students. During teaching, the teacher has access to student performance data and then makes decisions on the basis of performance data. When each action that determines what content to teach next – e.g. assigning a new skill – was made by the teacher, the system detected and collected it as an instructional decision. We collect data on whether that skill is assessed by the learning system as mastered or not, according to BKT, by the end of the learning period. A skill was considered mastered if BKT found probability of mastery greater than 95%, and as not mastered otherwise. The teacher then selects another skill for the student to work on. Learnta’s knowledge graph changes each time a new skill is encountered and assessed.

## 2.3 Statistical Analysis

We compare the degree to which students master skills, based on whether the teacher selects RtL skills, UtL skills, or already-mastered skills. The analyses are conducted separately for English and math. The outcome of interest is whether the student mastered the skill according to BKT. The number and percentage of skills that are mastered are tabulated for each type of teaching decision. We assess the association between instructional decisions and student mastery, looking at whether students are more likely to master RtL skills than UtL skills. The primary model is a logistic regression model with teaching decision as the single predictor variable. As a sensitivity analyses to assess the robustness of the primary model, mixed-effects logistic regressions are also conducted to adjust for the confounding effects of student and skill, either individually or both together. In these mixed-effects models, teaching decision is a ternary variable (RtL, UtL, already-mastered) and is considered as a fixed effect, while student-level and skill-level variables are treated as random effects. Odds ratios (OR) and corresponding P-values are calculated in R version 3.0.2 [5] using the `glm()` function for logistic regression and the `lme4` package [6] `glmer()` function for logistic regression with mixed effects.

### 3 Results

For mathematics learning (“Integral Expression”), the teacher made 619 instructional decisions. Among the decisions, the teacher taught RtL skills 238 times, and the students mastered them 63% of the time. The teacher taught UtL skills 208 times, and the mastery rate was only 46%. Already-mastered skills were taught 173 times, with a mastery rate of 80%. Note that the mastery rate of these already-mastered skills was well below 95% even though the algorithm had previously assessed the skill as mastered with over 95% confidence. This may be due to the probability of slipping, or forgetting the skill after it had been learned.

For English grammar learning (“Pronoun and Noun”), the teacher made 721 instructional decisions. Among the decisions, the teacher taught RtL skills 86 times, and the students mastered them 64% of the time. The teacher taught UtL skills 497 times, and the mastery rate was only 39%. Already-mastered skills were taught 138 times, with a mastery rate of 79%. As mentioned above, forgetting or slipping may lead to an actual mastery rate that is lower than 95%.

Logistic regression analyses and sensitivity analyses confirmed that instructional decisions were significantly associated with students’ learning outcomes,  $p < 0.001$  for both topics. Students who were taught a ready-to-learn skill were 4.34 times more likely to master an English grammar skill, and 2.78 times more likely for math.

### 4 Discussion and Conclusions

In a teacher-driven personalized learning environment, teachers decide which content to assign to students based on the student’s performance data. However, simply providing teachers with data is not always sufficient for good instructional decision making. This paper investigates whether knowledge graphs can be used to inform and improve teacher instructional decisions within an online learning platform, in terms of Vygotsky’s Zone of Proximal Development. A knowledge graph algorithm is applied to assess whether teachers assign content that a student is ready-to-learn (RtL), unready-to-learn (UtL), or already-mastered. We find that mastery is higher when teachers assign RtL skills than UtL skills, though not quite as high as if the teacher decides to re-teach a skill that the student already mastered. These findings, which generalize to both English and math, suggest steps we can take to optimize student learning outcomes and teacher decision making. Optimizing learning outcomes requires correct teaching decisions that lead students on the right path, based on the student’s ZPD. As such, it would be beneficial to create an interface to communicate what students are ready to learn to teachers, and what evidence this recommendation is based on. This can be accomplished by displaying knowledge graphs showing how the system’s recommendation of a skill is generated based on performance on prerequisite skills. Teacher training could emphasize the importance of using the ZPD to personalize learning and how to use knowledge graph recommendations in instructional design. While a knowledge graph may not provide a perfect operationalization of Vygotsky’s ZPD, it can offer teachers information that they can use to better support student learning.

## References

1. Baker, R., Wang, F., Ma, Z., Ma, W., Zheng, S.: Studying the effectiveness of an online language learning platform in China. *J. Interact. Learn. Res.* **29**(1), 5–24 (2018)
2. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief, pp. 1–57. Office of Educational Technology, U.S. Department of Education, Washington, DC (2012)
3. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* **4**(4), 253–278 (1995)
4. Earl, L., Katz, S.: Leading schools in a data-rich world. In: Leithwood, K., Hallinger, P. (eds.) *Second International Handbook of Educational Leadership and Administration*, pp. 1003–1022. Kluwer Academic, Dordrecht (2002)
5. Feng, M., Heffernan, N.T.: Informing teachers live about student learning: reporting in the assessment system. *Technol. Instr. Cognit. Learn.* **3**(1/2), 63 (2006)
6. Gilbert, S.W.: A widening gap: the support service crisis. *Syllabus* **14**(1), 18–57 (2000)
7. Holstein, K., McLaren, B.M., Aleven, V.: 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 (2017)
8. Jonassen, D.H.: Thinking technology: context is everything. *Educ. Technol.* **31**(6), 35–37 (1993)
9. Oliver, R., Omari, A.: Using online technologies to support problem-based learning: learners responses and perceptions. *Aust. J. Educ. Technol.* **15**(1), 58–79 (1999)
10. Porayska-Pomsta, K.: AI as a methodology for supporting educational praxis and teacher metacognition. *Int. J. Artif. Intell. Educ.* **26**(2), 679–700 (2016)
11. Mandinach, E.B., Jackson, S.S.: *Transforming Teaching and Learning Through Data-Driven Decision Making*. Corwin Press, Thousand Oaks (2012)
12. McLaren, B.M., Aleven, V.: 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 (2017)
13. Murray, T., Arroyo, I.: Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In: Cerri, S.A., Gouardères, G., Paraguaçu, F. (eds.) *ITS 2002. LNCS*, vol. 2363, pp. 749–758. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-47987-2\\_75](https://doi.org/10.1007/3-540-47987-2_75)
14. Vygotsky, L.S.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge (1978)

# **Workshop Papers**

# Supporting Lifelong Learning

Oluwabunmi (Adewoyin) Olakanmi<sup>1(✉)</sup>,  
Oluwabukola Mayowa Ishola<sup>2(✉)</sup>, Gord McCalla<sup>2</sup>, Ifeoma Adaji<sup>2</sup>,  
and Francisco J. Gutierrez<sup>3</sup>

<sup>1</sup> Department of Computing Science, University of Alberta, Edmonton, Canada  
[olakanmi@ualberta.ca](mailto:olakanmi@ualberta.ca)

<sup>2</sup> Department of Computer Science, University of Saskatchewan,  
Saskatoon, Canada  
[{bukola.ishola,gordon.mccalla,ifeoma.adaji}@usask.ca](mailto:{bukola.ishola,gordon.mccalla,ifeoma.adaji}@usask.ca)

<sup>3</sup> Department of Computer Science, University of Chile, Santiago, Chile  
[frgutier@dcc.uchile.cl](mailto:frgutier@dcc.uchile.cl)

## Workshop Description

Traditionally, lifelong learning has been accomplished through job training, short courses, and self-directed learning [1]. However, these approaches do not scale beyond the immediate learning needs of the learners. With the proliferation of social media, several knowledge resources and computer-mediated technologies exist that support lifelong learners in their day-to-day activities. Therefore, millions of these lifelong learners turn to online learning communities (OLCs) to help them overcome problems they may encounter in their day-to-day activities. In an OLC the challenge of supporting the evolving learning needs of learners is acute.

The goal of this workshop is to provide a forum for researchers to critically discuss ways to advance research in supporting lifelong learning beyond the walls of traditional educational systems. The workshop will provide an opportunity to discuss areas like social recommendation, adaptive technologies, collaborative tools, persuasive strategies, learning analytics and educational data mining to support lifelong learners; to look at enhancing lifelong learning through collaboration, educational games, personalized recommendation and educational diagnosis of lifelong learners; and to review literature addressing lifelong learning. Time will be allotted for presentation and questions; and at the end of the workshop, there will be a brainstorming session for overall discussion of the workshop presentations, challenges and the ways forward.

At the end of the workshop, we will develop a co-authored reference document, which summarizes the state-of-the-art, challenges and ways forward in supporting lifelong learners in online learning environments. Also, the workshop will provide an opportunity for researchers, both in industry and academia, to establish long term collaborations that can help expand on studies that support lifelong learners.

## Reference

1. Bruce, C.S.: Workplace experiences of information literacy. *Int. J. Inf. Manag.* **19**(1), 33–47 (1999)

# Educational Data Mining in Computer Science Education (CSEDM)

David Azcona<sup>1</sup>(✉), Yancy Vance Paredes<sup>2</sup>, Thomas W. Price<sup>3</sup>,  
and Sharon I-Han Hsiao<sup>2</sup>

<sup>1</sup> Dublin City University, Dublin, Ireland  
[david.azcona@insight-centre.org](mailto:david.azcona@insight-centre.org)

<sup>2</sup> Arizona State University, Tempe, AZ, USA  
[{yvmparedes,Sharon.Hsiao}@asu.edu](mailto:{yvmparedes,Sharon.Hsiao}@asu.edu)

<sup>3</sup> North Carolina State University, Raleigh, NC, USA  
[twprice@ncsu.edu](mailto:twprice@ncsu.edu)

There is a growing community of researchers at the intersection of AI, data mining and computing education research. The objective of this workshop is to facilitate a discussion among this research community, with a focus on how AI can uniquely impact Computer Science Education. The workshop is meant to be an interdisciplinary event at the intersection of AIED and Computing Education Research. Researchers, faculty and students are encouraged to share their AI- and data-driven approaches, methodologies and experiences where AI is transforming the way students learn Computer Science (CS) skills.

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 behavioural 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 challenges and issues in the domain-specific field, Computer Science Education.

# **Measuring, Analyzing, and Modeling Multimodal Multichannel Data for Supporting Self-regulated Learning by Making Systems More Intelligent for All in the 21st Century**

Roger Azevedo<sup>1</sup> and Gautam Biswas<sup>2(✉)</sup>

<sup>1</sup> University of Central Florida, Orlando, FL 32816, USA

<sup>2</sup> Vanderbilt University, Nashville, TN 37235, USA

[gautam.biswas@vanderbilt.edu](mailto:gautam.biswas@vanderbilt.edu)

**Abstract.** Learning with advanced learning technologies (ALTs) such as intelligent tutors, serious games, simulations, and immersive virtual environments, involves intricate and complex interactions among cognitive, metacognitive, motivational, affective, and social processes. Current psychological and educational research on learning with ALTs provides a wealth of empirical data indicating that learners of all ages have difficulty learning about complex topics in areas such as STEM. Learning with ALTs requires students to analyze the learning situation, set meaningful learning goals, determine which strategies to use, assess whether the strategies are effective in meeting the learning goal, and evaluate their emerging understanding of the topic. They also need to monitor and reflect on their understanding and modify their plans, goals, strategies, and effort in relation to contextual conditions (e.g., cognitive, motivational, and task conditions). We argue that understanding these processes necessitates the measurement, analyses, and modeling of multimodal multichannel data (e.g., log files, eye tracking, and physiological sensors) during learning and problem solving with ALTs.

Understanding the complex nature of the temporally unfolding SRL processes is being addressed by emerging interdisciplinary research using online trace methods (e.g., log-files, eye-tracking, think-aloud protocols, physiological sensors, screen recording of human-machine interactions, classroom discourse). The use of these methods has been widely applauded by the research community. Despite these benefits of multimodal multichannel data, analyzing these data come with their own set of challenges that will be addressed by the participants of this workshop. They include the following: (1) temporal alignment of data sources based on different sampling rates; (2) complexity in dealing with noisy and messy data (e.g., missing data) with traditional and contemporary data mining and machine learning techniques; (3) accurate classification and tracking of the underlying cognitive, metacognitive, and affective processes; (4) assessment of the levels of accuracy in modeling complex underlying processes, and confidence in inferences based on current analytical methods; and (5) implications of multimodal analyses on instruction and learning (e.g., providing timely scaffolding needed to facilitate emotion regulation).

# Ethics in AIED: Who Cares?

Wayne Holmes<sup>1</sup>  , Duygu Bektik<sup>1</sup>  , Maria Di Gennaro<sup>1</sup>,  
Beverly Park Woolf<sup>2</sup>  , and Rose Luckin<sup>3,4</sup> 

<sup>1</sup> Institute of Educational Technology, The Open University, Milton Keynes, UK  
[wayne.holmes@open.ac.uk](mailto:wayne.holmes@open.ac.uk)

<sup>2</sup> College of Information and Computer Sciences, University of Massachusetts,  
Amherst, MA, USA

<sup>3</sup> University College London, London, UK

<sup>4</sup> The Institute of Ethical AIED, London, UK

**Abstract.** Building on the outcomes of the first ‘ETHICS in AIED: Who Cares?’ workshop, held at the 2018 AIED conference, this year’s workshop recognized that, although there are encouraging signs, most AIED research, development and deployment continues to take place in what is essentially a moral vacuum. Still today, little research has been undertaken, no guidelines have been provided, no policies have been developed, and no regulations have been enacted to address the specific ethical issues raised by the application of AI in educational contexts. This year’s workshop was an opportunity for researchers to identify key ethical issues, to map out how to address the multiple challenges, and to help establish a basis for meaningful ethical reflection necessary for innovation in AIED.

**Keywords:** Artificial intelligence in education · Ethics · Ethical practice

## Ethics in AIED. Who Cares?

While the range of AI techniques researched in classrooms continues to grow, the ethical consequences are rarely fully considered—at least, there is very little published work considering the ethics of AIED. As a field (while we apply university research regulations), we continue to work without any fully-developed moral groundings specific to AIED. In fact, AIED raises an indeterminate number of as yet unanswered ethical questions. Concerns exist about the large volumes of data collected to support AIED (such as the recording of student competencies, emotions, strategies and misconceptions). Who owns and who is able to access this data, and what are the privacy concerns? Other major ethical concerns centre on AIED computational approaches. How should the data be analysed, interpreted and shared? However, the ethics of AIED cannot be reduced to questions about data or computation. AIED research also needs to account for the ethics of education—for example, the fact that many of its educational assumptions are contested by the learning sciences community. Further, the ethics of data, computational approaches, and education are the ‘known unknowns’. But what about the ‘unknown unknowns’, the ethical issues raised by AIED—at the intersection of data, computation and education—that have yet to be even identified?

The workshop helped develop a shared understanding of the multiple challenges and points of contention around the ethics of AIED, which will help inform policy for the International AIED Society and future AIED conferences.

# Adaptive and Intelligent Technologies for Informal Learning

H. Chad Lane<sup>1(✉)</sup>, Jonathan Rowe<sup>2</sup>, Stephen Blessing<sup>3</sup>,  
and Nesra Yannier<sup>4</sup>

<sup>1</sup> University of Illinois, Urbana-Champaign, Champaign, IL 61820, USA  
[hclane@illinois.edu](mailto:hclane@illinois.edu)

<sup>2</sup> North Carolina State University, Raleigh, NC 27695, USA  
[jprowe@ncsu.edu](mailto:jprowe@ncsu.edu)

<sup>3</sup> University of Tampa, Tampa, FL 33606, USA  
[sblessing@ut.edu](mailto:sblessing@ut.edu)

<sup>4</sup> Carnegie Mellon University, Pittsburgh, PA 15213, USA  
[nyannier@cs.cmu.edu](mailto:nyannier@cs.cmu.edu)

## Workshop Description

Early work in AIED often focused on formal learning environments (e.g., homework support tools, classroom technologies). As the field has evolved, this focus has expanded in a variety of ways, including considering how AIED technologies can be used in a wider range of learning contexts, such as informal settings. This workshop stems from this evolution and is a result of growing recognition of the importance for AIED technologies to (1) be more flexible in terms of the pedagogical tactics they employ and the ways they support effective learning, (2) be sensitive to noncognitive aspects of learning, such as motivation and affect, (3) be usable in physically and socially complex learning contexts, and (4) support higher levels of learner agency and self-direction (i.e., free-choice learning). Increasingly, AIED research embraces these challenges by creating technologies that are specifically designed to engage learners and be used in complex, informal learning environments.

Informal learning is often characterized as occurring in museums, science centers, zoos, aquariums, gardens, after-school programs, at-home learning, and with citizen science efforts, among other settings. Such settings raise a broad range of novel questions and challenges relevant to AIED. This workshop will seek to identify distinguishing characteristics of informal learning contexts, scrutinize common assumptions of AIED systems, revisit the questions related to the evaluation of learning, and discuss design implications for AIED outside of the classroom. Further, we will explore the suitability of classic AIED techniques (e.g., student modeling, automated feedback) for informal learning, including when they work and what modifications are necessary. The purpose of this workshop is to bring clarity to emerging AIED work that is situated within informal contexts, and to establish a community to discuss the design,

deployment, and evaluation of educational technologies for informal learning. Further, as the line between formal and informal learning is increasingly blurred, this workshop will enable progress toward developing a common research agenda on adaptive and intelligent technologies in informal settings.

# Designing Human-Centered AI Products

Kristen Olson, Maysam Moussalem, Di Dang<sup>(✉)</sup>, Kristie J. Fisher<sup>(✉)</sup>,  
Jess Holbrook<sup>(✉)</sup>, and Rebecca Salois<sup>(✉)</sup>

Google LLC, Mountain View, CA 94043, USA

{kristenolson,maysam,didang,kjfisher,jessh,rsalois}@google.com

**Abstract.** Enthusiasm for how AI can benefit students and educators continues to grow, but what are the unique design considerations for building products that use AI? This workshop will equip participants with exercises to use with their teams to build AI products that are grounded in human needs and avoid common AI design pitfalls. The Google People + AI Research and Engineering Education team will share insights from the People + AI Research Guidebook, a new resource that has steered design and development at Google over the past year. Through recommendations, frameworks, and examples, attendees will learn how to assess whether a problem is a good fit for machine learning, collect representative training data, and help users understand how to interact with AI systems.

**Keywords:** Human-centered design · Artificial intelligence · Mental models

# Standardization Opportunities for AI in Education

Robby Robson<sup>1</sup>, Richard Tong<sup>2(✉)</sup>, Robert Sottilare<sup>3</sup>, and K. P. Thai<sup>2</sup>

<sup>1</sup> Eduworks, Inc., Corvallis, OR 97333, USA  
[robby@computer.org](mailto:robby@computer.org)

<sup>2</sup> Squirrel AI Learning, Yixue Education Group, Highland Park, NJ 08904, USA  
[{richard.tong,kp.thai}@yixue.us](mailto:{richard.tong,kp.thai}@yixue.us)

<sup>3</sup> Soar Technology, Inc., Orlando, FL 32817, USA  
[bob.sottilare@soartech.com](mailto:bob.sottilare@soartech.com)

**Abstract.** This workshop explores opportunities to standardize conceptualization, components, best practices and processes used in educational systems that apply artificial intelligence (AI), including adaptive learning technologies, AI-based recommendation engines, and systems that use machine learning to model student interactions and preferences to improve learning outcomes.

**Keywords:** AI · Adaptive instructional systems · IEEE 2247.x

## Summary and Agenda

Educational tools enabled by AI have recently attracted attention for their potential to improve education quality and enhance traditional teaching and learning methods and are now being rolled out at scale in commercial and non-commercial products. Having achieved this level of maturity, standards for common interfaces, components and processes can serve as a foundation for new research and innovation while reducing the risk of adopting AI-based educational products and helping to avoid wasteful duplication of effort. Interoperability makes it possible to reuse existing technologies and content and to plug into existing educational ecosystems. This reduces costs and will accelerate advances the field of AI in Education by enabling researchers and innovators to more easily test and evaluate new approaches and technologies in real-world environments with large data sets. The goal of the proposed workshop is to explore opportunities to standardize components and processes used in educational systems that apply AI. These opportunities include: (1) standardizing conceptual models and taxonomies, to enable the meaningful description and comparison of AIS across a wide range of functionality (2) interoperability standards, to exchange and properly interpret the semantically rich data that AI-based learning technologies produce, and (3) recommended practices for AIS evaluation, to develop informed, consensus-driven best practices for evaluating AI systems.

The agenda includes: introductions, standards and standardization process for AI in education, conceptual models of AI in the practice of education, interoperability requirements and approaches, best practices for evaluation, critique and comments.

# Approaches and Challenges in Team Tutoring Workshop

Anne M. Sinatra<sup>1(✉)</sup> and Jeanine A. DeFalco<sup>1,2</sup>

<sup>1</sup> U.S. Army Combat Capabilities Command Soldier Center – Simulation and Training Technology Center, Orlando, FL, USA

{anne.m.sinatra.civ,jeanine.a.defalco.ctr}@mail.mil

<sup>2</sup> Oak Ridge Associated Universities, Oak Ridge, TN, USA

## Workshop Description

The “Approaches and Challenges in Team Tutoring” workshop is a follow up to a successful AIED 2018 conference workshop titled “Assessment and Intervention during Team Tutoring” [1]. During that workshop it was determined that while there were many approaches being used for team tutoring, there were still many challenges that need to be addressed. The current workshop covers the topic areas of approaches and challenges to team tutoring and collaborative learning in intelligent tutoring systems (ITSs).

The development of ITSs for teams and collaborative learning are time-intensive and difficult tasks that include technological, instructional and design based challenges. The goals of the current workshop include providing an opportunity for researchers to discuss the progress that they have made in team and collaborative adaptive tutoring, discuss the approaches that they have taken, and the challenges that they have encountered.

The workshop is broken down in three topic areas/themes: (1) Approaches to creating ITSs for teams, (2) Challenges and lessons learning in creating ITSs for teams, and (3) Collaborative learning/problem solving in ITSs. Each topic area will contain presentations of empirical and theoretical work, and after the presentations, there will be open discussion to identify commonalities in approaches. Through the open discussion, the current challenges and gaps in team tutoring research will be identified.

The expected outcomes of the workshop include determining approaches that have been successful or unsuccessful in meeting the challenges associated with team tutoring, identification of team tutoring gaps in varying learning domains, and determining the next steps in team tutoring research. The workshop is expected to be of interest to researchers in academia, government, and industry. The workshop will provide a forum for researchers who are working in this up and coming area to discuss both empirical and theoretical work that can contribute to furthering their future research.

## Reference

1. Sinatra, A.M., DeFalco, J.A. (eds.): Proceedings of the Assessment and Intervention During Team Tutoring Workshop, London, England, UK, 30 June, 2018. CEUR-WS.org. <http://ceur-ws.org/Vol-2153>

# Intelligent Textbooks

Sergey Sosnovsky<sup>1(✉)</sup>, Peter Brusilovsky<sup>2</sup>, Rakesh Agrawal<sup>3</sup>,  
Richard G. Baraniuk<sup>4</sup>, and Andrew S. Lan<sup>5</sup>

<sup>1</sup> Utrecht University, Princetonplein 5, 3584 CC Utrecht, the Netherlands  
s.a.sosnovsky@uu.nl

<sup>2</sup> University of Pittsburgh, 135 North Bellefield Avenue,  
Pittsburgh, PA 15260, USA  
peterb@pitt.edu

<sup>3</sup> Data Insights Laboratories, P.O. Box 41231, San Jose, CA 95160, USA  
rakesha.prof@gmail.com

<sup>4</sup> Rice University, 6100 Main Street, Houston, TX 77005, USA  
richb@rice.edu

<sup>5</sup> University of Massachusetts Amherst, 140 Governors Dr.,  
Amherst, MA 01003, USA  
andrewlan@cs.umass.edu

Textbooks have evolved over the last several decades in many aspects (how they are created, published, formatted, and maintained). Most textbooks these days can be accessed online, many of them freely. Commercial textbooks often come with libraries of supplementary educational resources or online educational services built on top of them. As a result, new research challenges and opportunities emerge that call for the application of artificial intelligence methods to enhance digital textbooks and learners' interaction with them. There are many exciting avenues for research in this new area; examples include, but not limited to:

- Modeling and representation of textbooks: examining the prerequisite and the semantic structure of textbooks to enhance their readability;
- Analysis of textbook usage logs: datamining patterns of learners' use of textbooks to examine learning and the pedagogical value of textbook content;
- Generation, manipulation, and presentation: exploring different formats and forms of textbook content to optimize presentation and comprehension of knowledge;
- Assessment and personalization: developing methods for generating assessment and enhancing textbooks with adaptive support to meet the needs of individual learners;
- Knowledge visualization: augmenting textbooks with concept maps, open learner models and other knowledge-rich extensions;
- Collaborative technologies: building and deploying social components of digital textbooks that enable learners to interact with not only content but other learners;
- Content enrichment: extending online textbooks with relevant external resources to improve learning, engagement, learner modeling, and personalization;
- Intelligent information retrieval and filtering: implementation of semantically-enhanced question-answering and browsing interfaces for digital textbooks

This workshop focuses on these and other research questions related to the idea of intelligent textbooks. It brings together researchers working in AI, human-computer interaction, information retrieval, intelligent tutoring systems, and user modeling to establish intelligent textbooks as a new, interdisciplinary research field.

# K12 Artificial Intelligence Education

Ning Wang<sup>1(✉)</sup> and James Lester<sup>2</sup>

<sup>1</sup> University of Southern California, Los Angeles, CA 90045, USA

nwang@ict.usc.edu

<sup>2</sup> North Carolina State University, Raleigh, NC 27695, USA

lester@ncsu.edu

## Summary

The workshop aims to provide a forum for researchers to discuss how to bring Artificial Intelligence (AI) education to K12 and how to apply state-of-the-art research on K12 computational thinking (CT) education, computer science (CS) education, data science, and the learning sciences to AI education in K12.

Over the past few decades, much research has been devoted to using AI to personalize and optimize the learning experience. Advances in AI have made educational technology more effective and efficient, and more prevalent in classroom and out-of-classroom learning settings. As AI has been gaining popularity in educational settings, it has also become ubiquitous in many other aspects of our everyday life and is redefining the future of work through human-machine alliances. Proficiency in the language of AI is key to a data-capable workforce that will continue to innovate and to support the AI-powered technology infrastructure and eco-system. Today's students will live a life heavily influenced by AI, and many will work in fields that involve or are influenced by AI. It is no longer sufficient to wait until students are in college to introduce AI concepts. They must begin to work with AI algorithmic problem solving and computational methods and tools in K-12. Now AI is no longer just part of the tools to educate students but also becomes front and center as a topic of education itself.

Much research in AI is needed to teach AI. For example, how do we make AI accessible to learners in K12? While AI algorithms are powerful tools for analyzing massive amounts of data, most of them use a decision-making process that is a “black box” to non-AI experts (and even to some AI experts). Recent advances in explainable AI are beginning to crack open this black box. Additionally, much research is needed on how to bring learning science to K12 AI education. For example, how do we design intelligent tutoring to teach AI, which incorporates most of the STEM subjects, while being a subject of problem-solving at its core? What lessons learned from STEM, CS, and CT education for K12 can be applied to teaching AI?

In summary, as AI becomes a critical skill for the future workforce, it will also become an integral part of K12 education. The workshop organizes the discussions on what are the challenges in teaching AI for K12 and how we should design AI-powered

technologies to teach AI for K12. A range of topics will be discussed at the workshop, including AI education and CS, CT education in K12, learning science in K12 AI education, K12 AI curriculum design, technological solutions for AI education, ethics in AI education, and teacher preparation.

# Author Index

- Abbas, Fakhri II-201  
Adaji, Ifeoma II-421  
Adeniran, Adetunji II-3  
Adorni, Giovanni I-1  
Afrin, Tazin II-9  
Agrawal, Rakesh II-431  
Agrawal, Sweety II-158  
Aguiar, Gene II-67  
Ahmed, Ishrat I-14  
Alamri, Ahmed II-67  
Albacete, Patricia I-37  
Aleven, Vincent I-157  
Alkaoud, Mohamed II-14  
Allen, Laura I-84  
Al-Luhaybi, Mashael I-26  
Alshehri, Mohammad II-67  
Alzetta, Chiara I-1  
Amadi, Chukwudi E. II-365  
An, Marshall II-252  
Andersen, Erik II-298  
Andres-Bray, Miguel I-445  
Arena, Dylan II-404  
Ausin, Markel Sanz I-544  
Ayedoun, Emmanuel II-19  
Azcona, David II-422  
Azevedo, Roger I-110, I-121, II-36, II-58, II-423  
Azizsoltani, Hamoon I-544
- Bailey, James I-96  
Baker, Ryan S. I-172, II-413  
Baker, Ryan I-144, I-445  
Baraniuk, Richard G. II-431  
Barnes, Tiffany I-544  
Basu, Satabdi II-116  
Beacham, Nigel II-3, II-206  
Bektik, Duygu II-424  
Bergner, Yoav I-14  
Beuget, Maël II-24  
Bickmore, Timothy II-374  
Biswas, Gautam I-532, II-116, II-263, II-354, II-423  
Blessing, Stephen II-426  
Bohlen, Turner II-409
- Bosch, Nigel I-296  
Boyer, Anne II-24  
Boyer, Kristy Elizabeth II-195, II-314  
Brand, Charleen II-30  
Brawner, Keith I-144  
Brenner, Daniel II-79  
Bretl, Timothy I-71  
Britt, M. Anne II-110  
Brooks, Christopher I-207  
Brusilovsky, Peter I-308, II-431  
Bryksin, Timofey II-174  
Burstyn, Judith I-432  
Butler, Eric I-383  
Buttery, Paula I-333
- Caines, Andrew I-333  
Castagnos, Sylvain II-24  
Čechák, Jaroslav I-48  
Chad Lane, H. II-426  
Chambel, Teresa II-89  
Cheema, Salman II-292  
Chen, Binglin I-71  
Chen, Guanliang I-59  
Chen, Jiahao II-381  
Chen, Penghe II-392  
Chi, Min I-544  
Chiu, Jennifer L. I-532  
Chounta, Irene-Angelica I-37  
Cloude, Elizabeth B. II-36  
Converse, Geoffrey II-41  
Cosyn, Eric I-258, II-179  
Coulson, Andrew II-79  
Counsell, Steve I-26  
Cristea, Alexandra II-67  
Crossley, Scott A. I-84, I-458  
Cukurova, Mutlu II-46, II-185  
Culver, Juliette II-105  
Curi, Mariana II-41
- D'Mello, Sidney K. I-296  
Daniel, Ben K. II-365  
Dang, Di II-428  
Dascălu, Mihai I-358, I-458, II-242  
Dashti, Cameron II-252

- Davaris, Myles [I-96](#)  
 DeFalco, Jeanine A. [II-52](#), [II-430](#)  
 Dehaene, Olivier [II-226](#)  
 Dehaene, Stanislas [II-226](#)  
 Deisadze, David [I-396](#)  
 Dever, Daryn A. [I-110](#), [I-121](#), [II-58](#)  
 Dhamecha, Tejas Indulal [I-469](#)  
 Di Eugenio, Barbara [II-94](#)  
 Di Gennaro, Maria [II-424](#)  
 Diana, Nicholas [II-62](#)  
 Díaz Tolentino, Armando [I-383](#)  
 Dickler, Rachel [II-163](#), [II-332](#)  
 Dimitrova, Vania [I-320](#)  
 Ding, Wenbiao [II-169](#), [II-381](#)  
 Dorodchi, Mohsen [I-370](#)  
 Dou, Wenwen [I-370](#)  
 Dowell, Nia [I-207](#)  
 Drijvers, Paul [II-281](#)  
 Duan, Bin [II-138](#)
- Effenberger, Tomáš [II-339](#)  
 Elkins-Tanton, Linda [II-409](#)  
 Eltayeby, Omar [II-201](#)
- Farzaneh, Amir Hossein [II-73](#)  
 Feng, Junchen [II-84](#)  
 Feng, Mingyu [II-79](#)  
 Feng, Xiaoqin [II-303](#)  
 Fernandes, David [II-67](#)  
 Ferrari, Deborah Viviane [II-220](#)  
 Ferreira, Rafael [II-257](#)  
 Fisher, Kristie J. [II-428](#)  
 Fonseca, Manuel J. [II-89](#)  
 Fossati, Davide [II-94](#)
- García Iruela, Miguel [II-89](#)  
 Gašević, Dragan [I-59](#), [II-257](#)  
 Gemmell, Laura [II-344](#)  
 Glick, Aaron Richard [II-386](#)  
 Gobert, Janice [II-163](#)  
 Godfrey, Andrew [I-207](#)  
 Goltz, Sean [I-194](#)  
 Gordon, Matthew [I-320](#)  
 Goswami, Mononito [I-283](#)  
 Goto, Mitsuhiro [II-128](#)  
 Grace, Kazjon [II-201](#)  
 Grande, Leo [I-396](#)  
 Gratch, Jonathan [II-122](#)  
 Green, Nicholas [II-94](#)
- Grover, Shuchi [II-116](#), [II-263](#)  
 Guerrero, Tricia A. [I-133](#)  
 Gurevych, Iryna [II-232](#)  
 Gutierrez, Francisco J. [II-421](#)  
 Guțu-Robu, Gabriel [II-242](#)  
 Gweon, Gahgene [II-153](#)
- Harpstead, Erik [I-518](#)  
 Hastings, Peter [II-110](#)  
 Hatfield, David [II-404](#)  
 Hauert, Sabine [II-344](#)  
 Hayashi, Yuki [II-19](#)  
 Hayashi, Yusuke [II-100](#), [II-319](#)  
 Hays, Matthew Jensen [II-386](#)  
 Heeren, Bastiaan [II-281](#)  
 Heffernan, Neil [I-396](#)  
 Henderson, Nathan L. [I-144](#)  
 Hijón Neira, Raquel [II-89](#)  
 Hilton, Michael [II-252](#)  
 Hirashima, Tsukasa [II-100](#), [II-319](#)  
 Holmes, Wayne [II-105](#), [II-133](#), [II-424](#)  
 Holbrook, Jess [II-428](#)  
 Holstein, Kenneth [I-157](#)  
 Hsiao, Sharon I-Han [II-422](#)  
 Huang, Gale Yan [II-169](#), [II-303](#), [II-381](#)  
 Hughes, Simon [II-110](#)  
 Hutchins, Nicole [II-116](#), [II-263](#)
- Inoue, Naoya [II-232](#)  
 Inui, Kentaro [II-232](#)  
 Ishino, Tatsuya [II-128](#)  
 Ishola, Oluwabukola Mayowa [II-421](#)  
 Isotani, Seiji [II-220](#)
- Jeuring, Johan [II-242](#), [II-281](#)  
 Johnson, Emmanuel [II-122](#)  
 Jordan, Pamela [I-37](#)
- Kamat, Venkatesh [II-359](#)  
 Kang, Seokmin [II-292](#)  
 Karkalas, Sokratis [II-185](#)  
 Karumbaiah, Shamya [I-172](#)  
 Kashihara, Akihiro [II-128](#)  
 Katz, Sandra [I-37](#)  
 Kennedy, Gregor [I-96](#)  
 Kent, Carmel [II-46](#)  
 Kim, Ji Hun [II-298](#)  
 Kim, Minkyung [I-84](#)  
 Kim, Peter [II-122](#)  
 Kim, Yanghee [II-73](#)

- Kizilkaya, Lawrence II-133  
 Koceva, Frosina I-1  
 Koedinger, Kenneth II-62  
 Kohama, Takeshi II-309  
 Krasich, Kristina I-296  
 Kuang, Yi II-138  
 Kulkarni, Mayank II-314  
 Kumar, Amruth N. II-143  
 Kurdi, M. Zakaria II-148  
 Kwak, Minji II-153
- Lala, Raja II-242  
 Lalwani, Amar II-158  
 Lan, Andrew S. II-431  
 Lane, H. Chad II-386  
 Latham, Annabel I-194  
 Lelei, David Edgar Kiprop I-182  
 Lester, James C. I-144  
 Lester, James II-36, II-195, II-314  
 Li, Haiying II-163  
 Li, Hang II-381  
 Li, Yuchen II-84  
 Likens, Aaron D. I-270  
 Lin, Yiwen I-207  
 Litman, Diane II-9  
 Liu, Ming I-220  
 Liu, Ruitao II-247  
 Liu, Tiaoqiao II-169  
 Liu, Zitao II-169, II-303, II-381  
 Lobanov, Artyom II-174  
 Lopez, Adam II-257  
 Lu, Yu II-392  
 Lubold, Nichola I-231  
 Lucas, Cherie I-220  
 Lucas, Christopher G. II-257  
 Lucas, Gale II-122  
 Luckin, Rose II-424  
 Luckin, Rosemary II-46  
 Luxembourger, Christophe II-24  
 Lv, Mengping II-138
- Ma, Wei II-413  
 Ma, Zhenjun II-413  
 Maharjan, Nabin I-244  
 Maher, Mary Lou I-370, II-201  
 Mahzoon, Mohammad Javad I-370  
 Mantzourani, Efi I-220  
 Massey-Allard, Jonathan II-30  
 Masthoff, Judith II-3, II-206
- Matayoshi, Jeffrey I-258, II-179  
 Matsuda, Noboru II-369  
 Mavrikis, Manolis II-185  
 Mawasi, Areej I-14  
 Mbipom, Blessing II-190  
 McCalla, Gord II-421  
 McCalla, Gordon I-182  
 McCarthy, Kathryn S. I-270  
 McElhaney, Kevin W. I-532  
 McKinniss, Tamera L. II-247  
 McLaren, Bruce M. I-37, I-157, I-445, I-518  
 McNamara, Danielle S. I-270, I-358  
 McNamara, Danielle I-84  
 Meng, Qinggang II-392  
 Mian, Shiven I-283  
 Mills, Caitlin I-296  
 Min, Wookhee II-195, II-314  
 Mirzaei, Mehrdad I-308, II-350  
 Misback, Edward I-432  
 Mitrovic, Antonija I-320  
 Miyazawa, Yoshimitsu I-482  
 Mogessie, Michael I-445  
 Mohseni, Maryam II-201  
 Moore, Russell I-333  
 Moore, Steven II-269  
 Morettin-Zupelari, Marina II-220  
 Mostow, Jack I-283  
 Mott, Bradford W. I-144  
 Mott, Bradford II-195, II-314  
 Moussalem, Maysam II-428  
 Mukhi, Nirmal I-469  
 Muldner, Kasia II-275, II-286  
 Munshi, Anabil II-354
- Naik, Vandana II-359  
 Najjar, Nadia II-201  
 Nam, SungJin I-345  
 Ndukwe, Ifeanyi G. II-365  
 Ngoh, Clarence II-252  
 Nguyen, Huy I-518  
 Nicula, Bogdan I-358  
 Nielsen, Rodney D. II-213  
 Niu, Xi I-370  
 Nomura, Toshihiro II-100  
 Nur, Nasheen I-370
- O'Leary, Stephen I-96  
 O'Rourke, Eleanor I-383  
 Ocumpaugh, Jaclyn I-172

- Odo, Chinasa II-206  
 Ogan, Amy I-231  
 Olakanmi, Oluwabunmi (Adewoyin) II-421  
 Oliveira, Elaine II-67  
 Oliveira, Suely II-41  
 Olson, Kristen II-428  
 Ozeki, Takashi II-309  
 Pailai, Jaruwat II-319  
 Paiva, Paula Maria Pereira II-220  
 Papapesiou, Emmanouela II-185  
 Parde, Natalie II-213  
 Pardos, Zachary A. II-14  
 Paredes, Yancy Vance II-422  
 Park, Kyungjin II-195  
 Park, Noseong I-370  
 Passalacqua, Samuele I-1  
 Patikorn, Thanaporn I-396  
 Pelánek, Radek I-48  
 Peng, Yan II-392  
 Penteado, Bruno Elias II-220  
 Pereira, Filipe D. II-67  
 Perez, Sarah II-30  
 Perret, Cecile A. I-358  
 Pinkwart, Niels II-237  
 Piotrkowicz, Alicja I-320  
 Piromchai, Patorn I-96  
 Pon-Barry, Heather I-231  
 Popović, Zoran I-383  
 Potier Watkins, Cassandra II-226  
 Price, Thomas W. II-422  
 Qi, Xiaojun II-73  
 Rau, Martina A. I-406, I-419, I-432  
 Reisert, Paul II-232  
 Rice, Andrew I-333  
 Richey, J. Elizabeth I-445  
 Robson, Robby II-429  
 Rodriguez, Elizabeth II-52  
 Roll, Ido II-30  
 Roscoe, Rod D. I-270  
 Rosé, Carolyn II-252  
 Rowe, Jonathan P. I-144, II-426  
 Rüdian, Sylvio II-237  
 Rummel, Nikol II-30  
 Rus, Vasile I-244  
 Ruşeti, Ştefan II-242  
 Sahebi, Shaghayegh I-308, II-350  
 Sakr, Majd II-252  
 Salois, Rebecca II-428  
 Samoilescu, Robert-Florian I-458  
 Samson, Perry I-345  
 San Pedro, Maria Ofelia Z. II-247  
 Sankaranarayanan, Sreecharan II-252  
 Schmidt, Tara A. I-419  
 Scruggs, Richard I-445  
 Sen, Ayon I-406  
 Seta, Kazuhisa II-19  
 Shimmei, Machi II-369  
 Shpilman, Alexey II-174  
 Shum, Simon Buckingham I-220  
 Silva, Luciano II-67  
 Sinatra, Anne M. II-52, II-430  
 Sinclair, Arabella J. II-257  
 Singh, Shaveen II-398  
 Sirbu, Maria-Dorinela I-458  
 Snyder, Caitlin II-116, II-263  
 Sosnovsky, Sergey II-431  
 Sottilare, Robert II-429  
 Stamper, John I-518, II-62, II-269  
 Stan Hum, R. II-52  
 Star, Jon I-445  
 Stranc, Samantha II-275  
 Sung, Chul I-469  
 Swift, Stephen I-26  
 Tacoma, Sietske II-281  
 Tang, Jiliang II-169, II-303  
 Tanton, James II-409  
 Taub, Michelle II-36  
 Thai, K. P. II-429  
 Theus, Anna-Lena II-286  
 Tong, Richard II-429  
 Torre, Ilaria I-1  
 Trausan-Matu, Stefan I-458  
 Tucker, Allan I-26  
 Ueno, Maomi I-482  
 Uto, Masaki I-494  
 Uzun, Hasan I-258, II-179  
 Vallejo, Gisela II-232  
 van Geest, Marcell II-242  
 VanLehn, Kurt I-507, II-292  
 Vanlehn, Kurt II-327  
 Verhagen, Josine II-404  
 Vince, David II-133  
 Viswanathan, Sree Aurovindh I-507, II-327

- Walker, Erin I-14, I-231  
Wang, Ning II-433  
Wang, Shang I-14  
Wang, Shuhan II-298  
Wang, Wenxin II-381  
Wang, Xu II-252  
Wang, Yeyu I-518  
Wang, Yuan II-409  
Wang, Zhiwei II-169, II-303  
Watanabe, Eiji II-309  
Wenham, Lucy II-344  
West, Matthew I-71  
Wetzel, Jon II-292  
Whitehurst, Amanda I-14  
Wiebe, Eric II-195, II-314  
Wiedbusch, Megan II-58  
Wiggins, Joseph B. II-314  
Wiggins, Joseph II-195  
Wijewickrema, Sudanthi I-96  
Wiley, Jennifer I-133  
Woolf, Beverly Park II-424  
Wu, Hao II-298  
Wunnasri, Warunya II-319  
Wylie, Ruth I-14  
Xu, Qi II-392  
Xu, Qiushi II-84  
Yang, Jie I-59  
Yannier, Nesra II-426  
Yousefi, Leila I-26  
Yu, Shengquan II-392  
Yu, Ziyang I-396  
Zahn, Miranda I-432  
Zhang, Bo II-84  
Zhang, Ningyu I-532  
Zhong, Shuyang II-138  
Zhou, Guojing I-544  
Zhou, Mengxi II-73  
Zhou, Shuo II-374  
Zhou, Yun I-96  
Zhu, Xiaojin I-406  
Zilles, Craig I-71  
Zou, Xiaotian II-413