

Viktoria Pammer-Schindler  
Mar Pérez-Sanagustín  
Hendrik Drachsler  
Raymond Elferink  
Maren Scheffel (Eds.)

LNCS 11082

# Lifelong Technology- Enhanced Learning

13th European Conference  
on Technology Enhanced Learning, EC-TEL 2018  
Leeds, UK, September 3–5, 2018, Proceedings

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, Lancaster, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Friedemann Mattern

*ETH Zurich, Zurich, Switzerland*

John C. Mitchell

*Stanford University, Stanford, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

C. Pandu Rangan

*Indian Institute of Technology Madras, Chennai, India*

Bernhard Steffen

*TU Dortmund University, Dortmund, Germany*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbrücken, Germany*


More information about this series at <http://www.springer.com/series/7409>

Viktoría Pammer-Schindler · Mar Pérez-Sanagustín  
Hendrik Drachsler · Raymond Elferink  
Maren Scheffel (Eds.)


# Lifelong Technology- Enhanced Learning

13th European Conference  
on Technology Enhanced Learning, EC-TEL 2018  
Leeds, UK, September 3–5, 2018  
Proceedings

*Editors*

Viktoria Pammer-Schindler   
Graz University of Technology  
Graz  
Austria

Mar Pérez-Sanagustín  
Pontificia Universidad Católica de Chile  
Providencia, Santiago de Chile  
Chile

Hendrik Drachslér   
DIPF | Leibniz Institute for Research and  
Information in Education  
Frankfurt  
Germany

Raymond Elferink  
RayCom BV  
Utrecht, Utrecht  
The Netherlands

Maren Scheffel  
Open University Netherlands  
Heerlen  
The Netherlands

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Computer Science  
ISBN 978-3-319-98571-8              ISBN 978-3-319-98572-5 (eBook)  
<https://doi.org/10.1007/978-3-319-98572-5>

Library of Congress Control Number: 2018950530

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2018, corrected publication 2018

Chapter “A Classification of Barriers that Influence Intention Achievement in MOOCs” is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see license information in the chapter.

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, express 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

Welcome to the proceedings of the 13th European Conference on Technology-Enhanced Learning (EC-TEL). This year, the conference was held in the city of Leeds, UK, September 3–5, 2018, and was hosted by the University of Leeds, which has a strong commitment to research-led and excellent technology-enhanced higher education at the university; In the framework of this endeavor, the conference a very active player in the European technology-enhanced learning community. In addition, and in order to promote interdisciplinary approaches to TEL, we embrace the opportunity to be co-located with the Medical Education Informatics conference.

We live in an increasingly digital and globalized world that offers great opportunities for information sharing and the generation of new knowledge. This reality has enabled us to move forward rapidly as a society in many respects, but has also led us to complex, diverse and interdisciplinary challenges that affect all areas of knowledge such as health, demographic change and well-being; food security and bioeconomy; secure and clean energy; smart and green energy; or climate action and environment.

In order to meet these major challenges, we need a society that enhances the development of 21st century skills for supporting lifelong learning citizens able to deal with the complexity and uncertainty that tomorrow's problems require. These 21st century skills encompass not only technical and domain-specific skills, but also domain-independent meta-skills such as the 4Cs: critical thinking, creativity, communication, and collaboration—all needed to manage the complexity of future problems. In this context, technology plays a key role in generating new learning environments that support learners across both formal and informal learning contexts, facilitating them in developing and practicing 21st century skills to face these future challenges.

To feed the debate on this topic, the 13th European Conference on Technology Enhanced Learning (EC-TEL) 2018 was organized around the theme “Lifelong Technology-Enhanced Learning: Dealing with the Complexity of 21st Century Challenges.”

This theme is visible in the following keynotes by outstanding speakers, all of whom are a reference in the TEL community. They are, in alphabetical order: Allison Littlejohn, from the Open University, UK, who spoke about “Professional Lifelong Learning”; Carolyn P. Rosé, from Carnegie Mellon University, who spoke about “Lifelong Learning in a Web-Scale Opportunity Space”; and David Wortley, from 360in360 Immersive Experiences, who spoke about “The impact of disruptive digital technologies on Education, Medicine, Health and Well-Being.”

We have accepted contributions covering the conference topic on many levels and encouraged participants to extend the debate around the role of and challenges for cutting-edge 21st century technologies such as artificial intelligence, robots, augmented reality, and ubiquitous computing technologies for learning. The theme and debate

were reflected throughout the conference through the workshops, papers, posters, and demos as well as the lively discussions.

Finally, the theme was also visible in the new format of practitioner papers, whereby we aim to step up our endeavor as a community to engage in active communication between research and practice, acknowledging that this is a two-ways communication in which research and practice inform each other, for the benefit of both.

EC-TEL 2018 received 142 full and short research papers, of which 42 were accepted for these proceedings (acceptance rate: 29.6%). We further accepted seven demos and 23 posters for this proceedings volume. Practitioner papers are published in adjunct CEUR WS proceedings volume.

As we do every year, we aimed at providing a high standard in our review process; there were at least three reviews for full and short research papers and at least two reviews by senior Program Committee members for full research papers. All reviews were checked for their content, not only their overall scores, by the program chairs; and in many cases the paper itself was checked – this was to ensure decisions were overall as fair as possible within the pool of all submitted papers and to balance individual differences in scoring/weighting different strengths and weaknesses of papers by reviewers. We thank all reviewers who provided constructive and informative reviews addressing both the authors and the decision-making chairs.

EC-TEL sees itself as a discussion venue for an interdisciplinary community interested in the pedagogical underpinnings for designing learning technologies—innovative, interactive, and intelligent technologies that have the potential to support learning; individual, social, and organizational learning processes; different learning communities and contexts; open learning arrangements—and seeks diversity in target user groups for technologies by being explicitly interested in TEL in developing countries and for users with special needs. We celebrate this interdisciplinarity. At the same time, this interdisciplinarity is challenging, as it requires of authors to at the same time make a novel contribution but also to connect to an interdisciplinary discourse; and of reviewers to appreciate contributions with a different angle than one's own. The Organizing Committee therefore continues to see – in line with last year – that the community needs to develop a shared vision of TEL, and of what constitutes valid research practice and methodology, understanding that at the intersection of disciplines, many methodologies may be valid without this leading to arbitrariness.

These challenges are also addressed within other activities of the European Association of Technology-Enhanced Learning (EATEL), of which EC-TEL is by now the most visible and prominent one:

- *Systematic training of early-stage researchers within the community*: Even before the EC-TEL itself was launched, EATEL launched the first Joint Summer School on Technology-Enhanced Learning (JTELSS - <http://ea-tel.eu/jtelss/>) as a training and networking event for early-stage TEL researchers in Europe. From the first EC-TEL in 2006 onward, EC-TEL and EATEL held a doctoral consortium at the EC-TEL to complement this summer school, with the overall goal of engaging the next generation of TEL researchers into the discourse of the community from early stages on.

- *Systematic methodological discourse within the community in order to increase shared methodological understanding:* This year, EATEL and EC-TEL further broadened the scope of TEL as a profession. This year’s focus on open science as part of the professional practice in TEL initiated a series of events addressing the ongoing professionalization of our field.
- *Systematic appreciation of practitioner perspectives into community:* This year, EC-TEL introduced the category of practitioner papers. We want to support the possibility of research to impact practice, and of practice to inform research.

Overall, this year’s EC-TEL showed its continued relevance for the TEL community in providing a world-class forum for academic and professional discourse with strong European grounding. The chairs aimed to create such a space for the attendees of this year’s EC-TEL, the authors, all contributors, and all readers of this proceedings volume. We are looking forward to a future in which this discourse will continue to be lively, innovative, and reflective.

We close by thanking all the authors who submitted their work to this year’s conference – you are the drivers of TEL research and practice in Europe. We also thank all Program Committee members and reviewers for their voluntary contributions – you are essential for sustaining the quality in our field. Finally, we thank the local organization team for their great work and their warm welcome in Leeds.

July 2018

Hendrik Drachsler  
Viktoria Pammer-Schindler  
Mar Pérez-Sanagustín  
Raymond Elferink  
Maren Scheffel  
Christian Glahn  
Mikhail Fominykh



# Organization

## Program Committee

Marie-Helene Abel	Heudiasyc, Université de Technologie de Compiègne, France
Andrea Adamoli	Università della Svizzera italiana, Switzerland
Carlos Alario-Hoyos	Universidad Carlos III de Madrid, Spain
Patricia Albacete	University of Pittsburgh, USA
Liaqat Ali	Simon Fraser University, Canada
Luis Anido Rifon	Universidade de Vigo, Spain
Alessandra Antonaci	Welten Institute, Open University, The Netherlands
Roberto Araya	Universidad de Chile, Chile
Inmaculada Arnedillo-Sánchez	Trinity College Dublin, Ireland
Juan I. Asensio-Pérez	Universidad de Valladolid, Spain
David Azcona	Dublin City University, Ireland
Zhen Bai	Carnegie Mellon University, USA
Antonio Balderas	Universidad de Cádiz, Spain
Merja Bauters	University of Helsinki and Aalto University, Finland
Jason Bernard	University of Saskatchewan, Canada
Anis Bey	University of La Rochelle, France
Sue Bickerdike	University of Leeds, UK
Miguel L. Bote-Lorenzo	Universidad de Valladolid, Spain
François Bouchet	Sorbonne Université, LIP6, France
Yolaine Bourda	LRI, CentraleSupélec, France
Bert Bredeweg	University of Amsterdam, The Netherlands
Julien Broisin	IRIT, University of Toulouse, France
Daniela Caballero	Universidad de Chile, Chile
Manuel Caeiro Rodríguez	University of Vigo, Spain
Lorenzo Cantoni	Università della Svizzera italiana, Switzerland
Teresa Cerratto-Pargman	Stockholm University, Sweden
Sven Charleer	Katholieke Universiteit Leuven, Belgium
Arunangsu Chatterjee	University of Plymouth, UK
Sunhea Choi	University of Southampton, UK
Irene-Angelica Chounta	Carnegie Mellon University, USA
Miguel Ángel Conde	University of León, Spain
Raquel M. Crespo García	Universidad Carlos III de Madrid, Spain
Ulrike Cress	Knowledge Media Research Center
Alexandra Cristea	Durham University, UK
Mutlu Cukurova	University College London, UK
Daniel Davis	Delft University of Technology, The Netherlands

Paul De Bra	Eindhoven University of Technology, The Netherlands
Maria De Marsico	Sapienza University of Rome, Italy
Carlos Delgado	Universidad Carlos III de Madrid, Spain
Stavros Demetriadis	Aristotle University of Thessaloniki, Greece
Christian Depover	Université de Mons, France
Michael Derntl	University of Tübingen, Germany
Philippe Dessus	University of Grenoble Alpes, LaRAC, France
Daniele Di Mitri	Open Universiteit, The Netherlands
Darina Dicheva	Winston-Salem State University, USA
Stefan Dietze	GESIS, Leibniz Institute for the Social Sciences, Germany
Yannis Dimitriadis	University of Valladolid, Spain
Vania Dimitrova	University of Leeds, UK
Lone Dirckinck-Holmfeld	Aalborg University, Denmark
Monica Divitini	Norwegian University of Science and Technology, Sweden
Juan Manuel Dodero	Universidad de Cádiz, Spain
Peter Dolog	Aalborg University, Denmark
Hendrik Drachslers	The Open University
Martin Ebner	University of Graz, Austria
Raymond Elferink	Raycom BV
Maka Eradze	Tallinn University, Estonia
Louis Faucon	Ecole Polytechnique Fédérale de Lausanne, Switzerland
Carmen Fernández-Panadero	Universidad Carlos III de Madrid, Spain
Angela Fessl	Know-Center GmbH, Graz, Austria
Olga Firssova	Welten Institute, Open University, The Netherlands
Beatriz Florian-Gaviria	Universidad del Valle, Colombia
Mikhail Fominykh	Independent/Self-employed
Fernando Gamboa	Universidad Nacional Autónoma de México, Mexico
Catherine Garbay	CNRS, LIG, France
Jesús Miguel García-Gorrostieta	INAOE
Serge Garlatti	IMT Atlantique, France
Dragan Gasevic	Monash University, Australia
Sebastien George	LIUM, Le Mans University, France
Panagiotis Germanakos	SAP SE and University of Cyprus, Cyprus
Michail Giannakos	Norwegian University of Science and Technology, Norway
Denis Gillet	Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland
Christian Glahn	University of Applied Sciences HTW Chur, Switzerland
Samuel González-López	Technological University of Nogales, Sonora, Mexico
Monique Grandbastien	LORIA, Université de Lorraine, France

Wolfgang Greller	Vienna University of Education, Austria
David Griffiths	Institute for Educational Cybernetics, University of Bolton, UK
Christian Guetl	Graz University of Technology, Austria
Gabriel Gutu-Robu	University Politehnica of Bucharest
Ashley Haberman-Lawson	Ludwig Maximilian University of Munich, Germany
Thanasis Hadzilacos	Open University of Cyprus and the Cyprus Institute
Cecilie Johanne Hansen	uniRes, University of Bergen, Norway
Andreas Harrer	University of Applied Sciences and Arts Dortmund, Germany
Matthias Hauswirth	University of Lugano, Switzerland
Josefina Hernandez	Pontificia Universidad Católica de Chile, Chile
Davinia Hernandez-Leo	Universitat Pompeu Fabra, Spain
Ángel Hernández-García	Universidad Politécnica de Madrid, Spain
Tore Hoel	Høgskolen i Oslo og Akershus, Norway
Sharon Hsiao	Arizona State University, USA
Stian Håklev	Ecole Polytechnique Fédérale de Lausanne, Switzerland
Petri Ihantola	University of Helsinki, Finland
Andri Ioannou	Cyprus University of Technology, Cyprus
Francis Jambon	Laboratoire d'Informatique de Grenoble, France
Patrick Jermann	Ecole Polytechnique Fédérale de Lausanne, Switzerland
Ioana Jivet	Open University, The Netherlands
Srecko Joksimovic	University of Edinburgh, UK
Pamela Jordan	University of Pittsburgh, USA
Ken Kahn	University of Oxford, UK
Marco Kalz	Heidelberg University of Education, Germany
Mihkel Kangur	Tallinn University, Estonia
Nikos Karacapilidis	University of Patras, Greece
Julia Kasch	Open University, The Netherlands
Susan Kennedy	National Health Services
Michael Kickmeier-Rust	Graz University of Technology, Austria
Andrea Kienle	University of Applied Sciences Dortmund, Germany
Barbara Kieslinger	Centre for Social Innovation
Ralf Klamma	RWTH Aachen University, Germany
Styliani Kleanthous	University of Cyprus
Roland Klemke	Open University, The Netherlands
Tomaž Klobučar	Jozef Stefan Institute, Slovenia
Carolien Knoop-Van Campen	Radboud University, The Netherlands
Ola Knutsson	Stockholm University, Sweden
Panagiotis Kosmas	Cyprus University of Technology, Cyprus
Vitomir Kovanovic	The University of Edinburgh, UK
Milos Kravcik	DFKI GmbH, Germany
Mart Laanpere	Tallinn University, Estonia

Chad Lane	University of Illinois at Urbana-Champaign, USA
Peter Lange	RWTH Aachen University, Germany
Lydia Lau	University of Leeds, UK
Élise Lavoué	Université Jean Moulin Lyon 3, France
Effie Lai-Chong Law	University of Leicester, UK
Dominique Lenne	Heudiasyc, Université de Technologie de Compiègne, France
Marina Lepp	University of Tartu, Estonia
Leo Leppänen	University of Helsinki, Finland
Tobias Ley	Tallinn University, Estonia
Andreas Lingnau	Humboldt-Universität zu Berlin, Germany
Martin Llamas-Nistal	University of Vigo, Spain
Christoph Lofi	Delft University of Technology, The Netherlands
Mathieu Loiseau	University of Grenoble Alpes, Grenoble, France
Aurelio Lopez	Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
Domitile Lourdeaux	Heudiasyc UMR7253, France
Vanda Luengo	Laboratoire d'informatique de Paris, LIP6, Sorbonne Université, France
Piret Luik	University of Tartu, Estonia
George Magoulas	Birkbeck College, Knowledge Lab, University of London, UK
Katherine Maillet	Institut Mines-Télécom, Télécom Ecole de Management, France
Jorge Maldonado	Universidad de Cuenca, Pontificia Universidad Católica de Chile, Chile
Nils Malzahn	Rhine-Ruhr Institute for Applied System Innovation e.V., Germany
Katerina Mangaroska	Norwegian University of Science and Technology, Norway
Estefania Martin	Universidad Rey Juan Carlos, Spain
Jean-Charles Marty	LIRIS - equipe SICAL
Alejandra Martínez	Universidad de Valladolid, Spain
Carlos Martínez Gaitero	Escola Universitària d'Infermeria Gimbernat, Spain
M. Antonia Martínez-Carreras	University of Murcia, Spain
Anna Mavroudi	KTH Royal Institute of Technology, Sweden
Rafael Mello	Federal Rural University of Pernambuco, Brazil
Agathe Merceron	Beuth University of Applied Sciences Berlin, Germany
Christine Michel	LIRIS, Université de Lyon, Insa Lyon, France
Konstantinos Michos	Universitat Pompeu Fabra
Alexander Mikroyannidis	The Open University
Eva Millan	Universidad de Málaga, Spain
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology
Inge Molenaar	Radboud University, The Netherlands
Baptiste Monterrat	INSA de Lyon, France

Pablo Moreno-Ger	Universidad Complutense de Madrid, Spain
Neil Morris	University of Leeds, UK
Pedro Munoz-Merino	Universidad Carlos III de Madrid, Spain
Mathieu Muratet	Sorbonne University, France
Juan A. Muñoz-Cristóbal	GSIC/EMIC, University of Valladolid, Spain
Mandran Nadine	Laboratoire d'Informatique de Grenoble (LIG), France
Rob Nadolski	Welten Institute, Open University, The Netherlands
Petru Nicolaescu	RWTH Aachen University, Germany
Jalal Nouri	Stockholm University, Sweden
Mikko Nurminen	Tampere University of Technology, Finland
Alejandro Ortega-Arranz	Universidad de Valladolid, Spain
Carmen L. Padrón-Nápoles	ATOS
Viktoria Pammer-Schindler	Know-Center GmbH and Graz University of Technology, Austria
Lucia Pannese	Imaginary
Pantelis Papadopoulos	Aarhus University, Denmark
Ionut Cristian Paraschiv	Politehnica University of Bucharest, Romania
Abelardo Pardo	University of South Australia
Kai Pata	Tallinn University
Mar Perez Sanagustin	Pontificia Universidad Católica de Chile, Chile
Zinayida Petrushyna	LOT Internet GmbH
Laëtitia Pierrot	Université de Poitiers, France
Niels Pinkwart	Humboldt-Universität zu Berlin, Germany
Gerti Pishtari	Tallinn University, Estonia
Elvira Popescu	University of Craiova, Romania
Richard Price	Health Education England, UK
Luis P. Prieto	Tallinn University, Estonia
Michael Prilla	Ruhr University of Bochum, Germany
Ronald Pérez Álvarez	Pontificia Universidad Católica de Chile, Chile
Eric Ras	Luxembourg Institute of Science and Technology, Luxembourg
Christoph Rensing	TU Darmstadt, Germany
Marc Rittberger	DIPF, Leibniz Institute for Research and Information in Education, Germany
Luz Stella Robles Pedrozo	UNED
María Jesús Rodríguez-Triana	Tallinn University, Estonia
Elisabeth Rolf	DSV, Stockholm University, Sweden
Adolfo Ruiz Calleja	Tallinn University, Estonia
Ellen Rusman	Open University, The Netherlands
Demetrios Sampson	Curtin University, Australia
Eric Sanchez	Université Fribourg, Switzerland
Olga C. Santos	aDeNu Research Group (UNED)
Luisa Sanz-Martínez	Universidad de Valladolid, Spain
Maren Scheffel	Open University, The Netherlands
Daniel Schiffner	Goethe Universität Frankfurt, Germany

Andreas Schmidt	Karlsruhe University of Applied Sciences, Germany
Marcel Schmitz	Zuyd Hogeschool
Jan Schneider	Deutsches Institut für Internationale Pädagogische Forschung, Germany
Ulrik Schroeder	RWTH Aachen University, Germany
Karim Sehaba	LIRIS, Université Lumière Lyon 2, France
Kshitij Sharma	Norwegian University of Science and Technology, Norway
Puneet Sharma	University of Tromsø, Norway
Mike Sharples	The Open University
Tanmay Sinha	Institute of Learning Sciences and Higher Education, ETH Zurich, Switzerland
Alan Smeaton	Dublin City University, Ireland
Sergey Sosnovsky	Utrecht University, The Netherlands
Marcus Specht	Open University, The Netherlands
Daniel Spikol	Malmö University, Sweden
Slavi Stoyanov	Open University, The Netherlands
Christian M. Stracke	Open University, The Netherlands
Alexander Streicher	Fraunhofer
Merilin Säde	University of Tartu, Estonia
Esther Tan	Open University, The Netherlands
Erika Tanhua-Piironen	University of Tampere, Finland
Stefano Tardini	Università della Svizzera italiana, Switzerland
Pierre Tchounikine	University of Grenoble, France
Marco Temperini	Sapienza University of Rome, Italy
Vladimir Tomberg	Tallinn University, Estonia
Richard Tortorella	University of North Texas, USA
Stefan Trausan-Matu	University Politehnica of Bucharest, Romania
Tamsin Treasure-Jones	University of Leeds, UK
Chrysanthi Tseloudi	Trinity College Dublin, Ireland
Eno Tõnisson	University of Tartu, Estonia
Carsten Ullrich	DFKI GmbH, Germany
Rémi Venant	IRIT
Katrien Verbert	Katholieke Universiteit Leuven, Belgium
Himanshu Verma	University of Fribourg, Switzerland
Julio Villena-Román	MeaningCloud
Massimo Vitiello	Graz University of Technology, Austria
Terje Väljataga	Tallinn University, Estonia
Wim Westera	CELSTEC-Centre for Learning Sciences and Technologies, Open University, The Netherlands
Denise Whitelock	The Open University, UK
Fridolin Wild	Oxford Brookes University, UK
Amel Yessad	LIP6, Sorbonne Université, France
Raphael Zender	University of Potsdam, Germany
Yue Zhao	Delft University of Technology, The Netherlands

**Additional Reviewers**

Balderas, Antonio  
Barreiros, Carla  
Dehler, Jessica  
Fernández, Camino  
Greven, Christoph  
Holtz, Peter  
Laforcade, Pierre  
Liaqat, Daniyal  
Lukarov, Vlatko  
Moreno-Marcos, Pedro Manuel

Oyelere, Solomon Sunday  
Perez De La Cruz, Jose-Luis  
Person Montero, Tatiana  
Pesonen, In Collab With Joonas  
Renner, Bettina  
Rubio, Aarón  
Ruiz-Rube, Iván  
Schulz, Sandra  
Wahid, Usman

# Contents

## Research Papers

A Classification of Barriers that Influence Intention Achievement in MOOCs . . . . .	3
<i>Maartje Henderikx, Karel Kreijns, and Marco Kalz</i>	
Tools to Support Self-Regulated Learning in Online Environments: Literature Review . . . . .	16
<i>Ronald Pérez-Álvarez, Jorge Maldonado-Mahauad, and Mar Pérez-Sanagustín</i>	
The Psychometric Properties of a Preliminary Social Presence Measure Using Rasch Analysis . . . . .	31
<i>Karel Kreijns, Joshua Weidlich, and Kamakshi Rajagopal</i>	
Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences. . . . .	45
<i>Jan Schneider, Daniele Di Mitri, Bibeg Limbu, and Hendrik Drachslér</i>	
How Teachers Prepare for the Unexpected Bright Spots and Breakdowns in Enacting Pedagogical Plans in Class . . . . .	59
<i>Ghita Jalal, Valentin Lachand, Aurélien Tabard, and Christine Michel</i>	
Evaluating the Robustness of Learning Analytics Results Against Fake Learners. . . . .	74
<i>Giora Alexandron, José A. Ruipérez-Valiente, Sunbok Lee, and David E. Pritchard</i>	
Where Is the Learning in Learning Analytics? A Systematic Literature Review to Identify Measures of Affected Learning . . . . .	88
<i>Justian Knobbout and Esther van der Stappen</i>	
Can I Have a Mooc2Go, Please? On the Viability of Mobile vs. Stationary Learning . . . . .	101
<i>Yue Zhao, Tarmo Robal, Christoph Lofi, and Claudia Hauff</i>	
Validation of the Revised Self-regulated Online Learning Questionnaire . . . .	116
<i>Renée S. Jansen, Anouschka van Leeuwen, Jeroen Janssen, and Liesbeth Kester</i>	
SRLx: A Personalized Learner Interface for MOOCs. . . . .	122
<i>Dan Davis, Vasileios Triglianios, Claudia Hauff, and Geert-Jan Houben</i>	



Motivating Students to Enhance Their Knowledge Levels Through Personalized and Scrutable Visual Narratives . . . . .	136
<i>Bilal Yousuf, Athanasios Staikopoulos, and Owen Conlan</i>	
Supporting the Adaptive Generation of Learning Game Scenarios with a Model-Driven Engineering Framework. . . . .	151
<i>Pierre Laforcade and Youness Laghouaouta</i>	
Student Drop-out Modelling Using Virtual Learning Environment Behaviour Data. . . . .	166
<i>Jakub Kuzilek, Jonas Vaclavek, Viktor Fuglik, and Zdenek Zdrahal</i>	
A Microservice Infrastructure for Distributed Communities of Practice. . . . .	172
<i>Peter de Lange, Bernhard Göschlberger, Tracie Farrell, and Ralf Klamma</i>	
Multimodal Analytics for Real-Time Feedback in Co-located Collaboration . . .	187
<i>Sambit Praharaaj, Maren Scheffel, Hendrik Drachslar, and Marcus Specht</i>	
Towards Personalized Learning Objectives in MOOCs. . . . .	202
<i>Tobias Rohloff and Christoph Meinel</i>	
PsychOut! a Mobile App to Support Mental Status Assessment Training . . . .	216
<i>Carrie DEMMANS EPP, Joe Horne, Britney B. Scolieri, Irene Kane, and Amy S. Bowser</i>	
Exploring Gamification to Prevent Gaming the System and Help Refusal in Tutoring Systems . . . . .	231
<i>Otávio Azevedo, Felipe de Moraes, and Patricia A. Jaques</i>	
Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese . . . . .	245
<i>Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins</i>	
Fine-Grained Cognitive Assessment Based on Free-Form Input for Math Story Problems . . . . .	262
<i>Bastiaan Heeren, Johan Jeuring, Sergey Sosnovsky, Paul Drijvers, Peter Boon, Sietske Tacoma, Jesse Koops, Armin Weinberger, Brigitte Grugeon-Allys, Françoise Chenevotot-Quentin, Jorn van Wijk, and Ferdinand van Walree</i>	
Extending the SIPS-Model: A Research Framework for Online Collaborative Learning. . . . .	277
<i>Karel Kreijns and Paul A. Kirschner</i>	

A Syllogism for Designing Collaborative Learning Technologies  
in the Age of AI and Multimodal Data . . . . . 291  
*Mutlu Cukurova*

“Make It Personal!” - Gathering Input from Stakeholders for a Learning  
Analytics-Supported Learning Design Tool. . . . . 297  
*Marcel Schmitz, Maren Scheffel, Evelien van Limbeek,  
Roger Bemelmans, and Hendrik Drachsler*

Investigating the Relationships Between Online Activity, Learning  
Strategies and Grades to Create Learning Analytics-Supported  
Learning Designs . . . . . 311  
*Marcel Schmitz, Maren Scheffel, Evelien van Limbeek,  
Nicolette van Halem, Ilja Cornelisz, Chris van Klaveren,  
Roger Bemelmans, and Hendrik Drachsler*

Evidence for Programming Strategies in University Coding Exercises . . . . . 326  
*Kshitij Sharma, Katerina Mangaroska, Halvard Trætteberg,  
Serena Lee-Cultura, and Michail Giannakos*

Incorporating Blended Learning Processes in K12 Mathematics Education  
Through BA-Khan Platform . . . . . 340  
*Valeria Henríquez, Eliana Scheihing, and Marta Silva*

Predicting Learners’ Success in a Self-paced MOOC Through Sequence  
Patterns of Self-regulated Learning . . . . . 355  
*Jorge Maldonado-Mahauad, Mar Pérez-Sanagustín,  
Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos,  
Pedro J. Muñoz-Merino, and Carlos Delgado-Kloos*

Semantically Meaningful Cohorts Enable Specialized Knowledge  
Sharing in a Collaborative MOOC . . . . . 370  
*Stian Håklev, Kshitij Sharma, Jim Slotta, and Pierre Dillenbourg*

Detecting Learning Strategies Through Process Mining . . . . . 385  
*John Saint, Dragan Gašević, and Abelardo Pardo*

Low-Investment, Realistic-Return Business Cases for Learning Analytics  
Dashboards: Leveraging Usage Data and Microinteractions . . . . . 399  
*Tom Broos, Katrien Verbert, Greet Langie, Carolien Van Soom,  
and Tinne De Laet*

Identifying Design Principles for Learning Design Tools: The Case  
of edCrumble . . . . . 406  
*Laia Albó and Davinia Hernández-Leo*

Exploring Causality Within Collaborative Problem Solving Using Eye-Tracking. . . . .	412
<i>Kshitij Sharma, Jennifer K. Olsen, Vincent Aleven, and Nikol Rummel</i>	
Towards an Automated Model of Comprehension (AMoC) . . . . .	427
<i>Mihai Dascalu, Ionut Cristian Paraschiv, Danielle S. McNamara, and Stefan Trausan-Matu</i>	
Course-Adaptive Content Recommender for Course Authoring . . . . .	437
<i>Hung Chau, Jordan Barria-Pineda, and Peter Brusilovsky</i>	
Assessing Leadership Competencies Through Social Network Analysis . . . . .	452
<i>Faisal Ghaffar, Neil Peirce, and Alec Serlie</i>	
Concept Focus: Semantic Meta-Data for Describing MOOC Content . . . . .	467
<i>Sepideh Mesbah, Guanliang Chen, Manuel Valle Torre, Alessandro Bozzon, Christoph Lofi, and Geert-Jan Houben</i>	
Help Me Understand This Conversation: Methods of Identifying Implicit Links Between CSCL Contributions . . . . .	482
<i>Mihai Masala, Stefan Ruseti, Gabriel Gutu-Robu, Traian Rebedea, Mihai Dascalu, and Stefan Trausan-Matu</i>	
The Effect of Personality and Course Attributes on Academic Performance in MOOCs. . . . .	497
<i>Mahdi Rahmani Hanzaki and Carrie Demmans Epp</i>	
Learning by Reviewing Paper-Based Programming Assessments . . . . .	510
<i>Yancy Vance Paredes, David Azcona, I-Han Hsiao, and Alan Smeaton</i>	
Which Learning Visualisations to Offer Students? . . . . .	524
<i>Susan Bull, Peter Brusilovsky, and Julio Guerra</i>	
Detection of Student Modelling Anomalies. . . . .	531
<i>Sergey Sosnovsky, Laurens Müter, Marc Valkenier, Matthieu Brinkhuis, and Abe Hofman</i>	
Expanding the Curricular Space with Educational Robotics: A Creative Course on Road Safety . . . . .	537
<i>Andri Ioannou, Chrysanthos Socratous, and Elena Nikolaedou</i>	
<b>Poster and Demo Papers</b>	
New Approaches to Training of Power Substation Operators Based on Interactive Virtual Reality . . . . .	551
<i>Rinat R. Nasyrov and Peter S. Excell</i>	

Enabling Systematic Adoption of Learning Analytics through a Policy Framework. . . . . 556  
*Yi-Shan Tsai, Maren Scheffel, and Dragan Gašević*

Diversity Profiling of Learners to Understand Their Domain Coverage While Watching Videos. . . . . 561  
*Entisar Abolkasim, Lydia Lau, Vania Dimitrova, and Antonija Mitrovic*

Using Digital Medical Collections to Support Radiology Training in E-learning Platforms . . . . . 566  
*Félix Buendía, Joaquín Gayoso-Cabada, and José-Luis Sierra*

Temporal Analytics of Workplace-Based Assessment Data to Support Self-regulated Learning . . . . . 570  
*Alicja Piotrkowicz, Vania Dimitrova, and Trudie E. Roberts*

Learning Analytics Dashboard Analysing First-Year Engineering Students . . . 575  
*Jonas Vaclavek, Jakub Kuzilek, Jan Skocilas, Zdenek Zdrahal, and Viktor Fuglik*

The Learning Analytics Indicator Repository . . . . . 579  
*Daniel Biedermann, Jan Schneider, and Hendrik Drachsler*

Eye-Tracking for User Attention Evaluation in Adaptive Serious Games . . . . 583  
*Alexander Streicher, Sebastian Leidig, and Wolfgang Roller*

Development of a Learning Economy Platform Based on Blockchain . . . . . 587  
*Masumi Hori, Seishi Ono, Toshihiro Kita, Hiroki Miyahara, Shiu Sakashita, Kensuke Miyashita, and Kazutuna Yamaji*

Enhancing Human Learning of Motions: An Approach Through Clustering . . . 591  
*Quentin Couland, Ludovic Hamon, and Sébastien George*

How to Help Teachers Adapt to Learners? Teachers’ Perspective on a Competency and Error-Type Centered Dashboard . . . . . 596  
*Iryna Nikolayeva, Bruno Martin, Amel Yessad, Françoise Chenevotot, Julia Pilet, Dominique Prévité, Brigitte Grugeon-Allyls, and Vanda Luengo*

MedSense: The Development of a Gamified Learning Platform for Undergraduate Medical Education . . . . . 600  
*Justin Choon Hwee Ng, Sarah Zhuling Tham, Chin Rui Chew, Amelia Jing Hua Lee, and Sook Muay Tay*

edCrumble: Designing for Learning with Data Analytics . . . . . 605  
*Laia Albó and Davinia Hernández-Leo*

Ensuring Novelty and Transparency in Learning Resource- Recommendation Based on Deep Learning Techniques . . . . .	609
<i>Wael Alkhatib, Eid Araache, Christoph Rensing, and Steffen Schnitzer</i>	
Exploring Math Achievement Through Gamified Virtual Reality . . . . .	613
<i>Espen Stranger-Johannessen</i>	
Observational Scaffolding for Learning Analytics: A Methodological Proposal . . . . .	617
<i>Jairo Rodríguez-Medina, María Jesús Rodríguez-Triana, Maka Eradze, and Sara García-Sastre</i>	
Cohesion-Centered Analysis of Sociograms for Online Communities and Courses Using <i>ReaderBench</i> . . . . .	622
<i>Mihai Dascalu, Maria-Dorinela Sirbu, Gabriel Gutu-Robu, Stefan Ruseti, Scott A. Crossley, and Stefan Trausan-Matu</i>	
A Digital Ecosystem for Digital Competences: The CRISS Project Demo . . .	627
<i>Manolis Mavrikis, Lourdes Guardia, Mutlu Cukurova, and Marcelo Maina</i>	
Towards Generation of Ambiguous Situations in Virtual Environments for Training . . . . .	631
<i>Azzeddine Benabbou, Domitile Lourdeaux, and Dominique Lenne</i>	
Digging for Gold: Motivating Users to Explore Alternative Search Interfaces. . . . .	636
<i>Angela Fessel, Alfred Wertner, and Viktoria Pammer-Schindler</i>	
The Role of Ubiquitous Computing and the Internet of Things for Developing 21st Century Skills Among Learners: Experts' Views. . . . .	640
<i>Olga Viberg and Anna Mavroudi</i>	
An Exploratory Study on Student Engagement with Adaptive Notifications in Programming Courses. . . . .	644
<i>David Azcona, I-Han Hsiao, and Alan Smeaton</i>	
Instrumentation of Classrooms Using Synchronous Speech Transcription . . . .	648
<i>Vincent Bettenfeld, Salima Mdhaffar, Christophe Choquet, and Claudine Piau-Toffolon</i>	
A Programming Language Independent Platform for Algorithm Learning. . . .	652
<i>Bruno Burke, Peter Weßeler, and Jürgen te Vrugt</i>	
Using Thematic Analysis to Understand Students' Learning of Soft Skills from Videos . . . . .	656
<i>Björn Sjöden, Vania Dimitrova, and Antonija Mitrovic</i>	

Formalizing CSCL Scripts with Logic and Constraints. . . . . 660  
*Andreas Papasalouros*

Correction to: Lifelong Technology-Enhanced Learning. . . . . E1  
*Viktoria Pammer-Schindler, Mar Pérez-Sanagustín, Hendrik Drachsler,  
Raymond Elferink, and Maren Scheffel*

**Author Index** . . . . . 665

# **Research Papers**



# A Classification of Barriers that Influence Intention Achievement in MOOCs

Maartje Henderikx<sup>1(✉)</sup>, Karel Kreijns<sup>1</sup>, and Marco Kalz<sup>2</sup>

<sup>1</sup> Welten Institute, Open University of the Netherlands,  
Heerlen, The Netherlands

{maartje.henderikx, karel.kreijns}@ou.nl

<sup>2</sup> Heidelberg University of Education, Heidelberg, Germany  
kalz@ph-heidelberg.de

**Abstract.** MOOC-learning can be challenging as barriers which prevent or hinder acting out MOOC-takers' individual learning intentions may be encountered. The aim of this research was to elicit and to empirically classify barriers that influence this intention achievement in MOOCs. The best fit model of our factor-analytical approach resulted in 4 distinctive components; 1. Technical and online-learning related skills, 2. Social context, 3. Course design/expectations management, 4. Time, support and motivation. The main finding of our study is that the experienced barriers by MOOC-takers are predominantly non-MOOC related. This knowledge can be of value for MOOC-designers and providers. It may guide them in finding suitable re-design solutions or interventions to support MOOC-takers in their learning, even if it concerns non-MOOC related issues. Furthermore, it makes a valuable contribution to the expanding empirical research on MOOCs.

**Keywords:** MOOCs · Online learning · Barriers · Factor analysis

## 1 Introduction

An often-heard concern regarding MOOCs is their high dropout rate [1]. These dropout rates—generally used to assess MOOC-success—are misleading, as often success measurements from traditional education are used [2–5]. Kalz, Kreijns, Walhout, Castaño-Munoz, Espasa, and Tovar [6] introduced a theoretical framework that combines distal and proximal variables and which takes into account individual intentions and barriers. Since different educational contexts deserve different educational measures [7], Henderikx, Kreijns and Kalz [2] further specified this theoretical framework into a model to take into account individual intentions of MOOC-takers as a starting point for measuring educational success in MOOCs. But, even when taking the individual intentions as a starting point, a study by Henderikx, Kreijns and Kalz [3] showed that there is still a substantial group of MOOC-takers who do not achieve what they intended to do. It seems that they encounter barriers preventing or hindering them from acting out their individual learning intentions.

These barriers can be either MOOC related or non-MOOC related and may cause MOOC-takers to change their individual intentions or even to stop [3]. While there are



related studies dealing with the empirical analysis of the effects of barriers to online learning and distance education using various statistical techniques [8–13], for the context of massive open online learning such analyses are limited. Current studies on barriers in MOOCs mainly focus on a restricted number of barriers in case studies, qualitative research setups, literature reviews and descriptive studies [14–17]. There are some studies which empirically investigate barriers to student retention, however these studies merely focus on the effect of specifically selected barriers [18, 19]. Furthermore, some studies in online learning or distance education context grouped types of barriers [9] or aimed to empirically identify barrier components [10]. But, apart from an exploratory study on barriers in MOOCs by Henderikx et al. [3], there is no synthesized overview of MOOC-specific barriers available.

In this study, an exploratory factor analysis was used to categorize these potential barriers and present a MOOC-specific barrier classification, that could contribute to purposefully improve MOOCs and enhance MOOC-taker experiences and intention achievement. First, a literature review will give a brief overview of the most relevant literature on barriers to online learning and MOOCs specifically. Second, the methodology of the study will be reported, followed by the results of the factor analysis. Lastly, the results will be discussed as well as the limitations, implications for practice and recommendations for future research.

## 2 Literature Review

Many different issues are perceived as possible barriers to online learning and distance education. An extensive literature review on barriers in distance education by Galusha [9] showed that students in a distance learning environment regard financial costs, disruption of family life, lack of support from the employer, lack of feedback, lack of instructor presence, lack of technical assistance, lack of planning assistance, lack of social contact, unfamiliarity with distance learning, lack of computer or writing skills as disablers to their learning. She grouped these barriers into five categories (1) costs and motivators, (2) feedback and teacher contact, (3) student support and services, (4) alienation and isolation and (5) lack of experience and training.

Peltier, Drago and Schibrowsky [12] chose to investigate which role six specific dimensions, drawn from literature, played in perceived effectiveness of online education. These dimensions were (1) instructor support and mentoring, (2) course content, (3) course structure, (4) student-to-student interaction, (5) information technology and (6) instructor-student. Their regression results showed that course content, instructor support and mentoring played a substantial role and can be regarded as the most important barriers - or success factors if positively experienced - to students' learning experiences.

Other reported challenging characteristics as perceived by students in online learning context are technical problems, perceived lack of community, time constraints and unclear course objectives as found by Song, Singleton, Hill and Koh [13] in their mixed-methods study.

Eom, Wen and Ashill [8] examined the determinants of students' satisfaction in the context of university online courses. They included the variables course structure,

instructor feedback, self-motivation, learning style, interaction, and instructor facilitation, quite similar to the study undertaken by Peltier et al. [12]. Results of the structural equation modelling analysis revealed that instructor feedback and learning style were significant predictors for student success, indicating that these issues are important for learning and could become barriers if students are not satisfied with these specific issues.

Qualitative research by Aragon and Johnson [20] uncovered that self-reported reasons for non-completion of community college online courses were time constraints, lack of instructor interaction, bad course content, lack of communication and technological issues. Furthermore, Park and Choi [11] found that lack of family- and work support are positively related to non-completion and can thus be regarded as barriers to online learning.

Research that sought to integrate perceived barriers students (expected to) face in an online distance education context was conducted by Muilenburg and Berge [10]. Their factor analytical study which used the principal component extraction method, revealed that these barriers could be assigned to eight distinctive components: (1) administrative/instructor issues, (2) social interactions, (3) academic skills, (4) technical skills, (5) learner motivation, (6) time and support for studies, (7) cost and access to the internet, (8) technical problems. A composite scores calculation per component identified social interactions as the most important barrier for students' online-learning. Academic skills have been identified as the least important barrier.

These studies, reporting on aforementioned barriers were all conducted in a general online learning or distance education context. Yet, with the still relatively new online learning environment of MOOCs, research on barriers in MOOC-specific context has caught on and is increasing.

In a study on student retention in MOOCs, Adamopoulos [18] used various text mining and predictive modelling techniques to analyse online student reviews and online available course characteristics. The analysis showed that the negative sentiment for the discussion forum, length of the course and workload had a significant negative effect on student retention. Belanger and Thornton [14] evaluated a MOOC on Bio-electricity by analysing pre- and post-questionnaires and log-data. The main barriers that were mentioned by students as reason for non-completion were time constraints and insufficient background knowledge. A literature review by Khalil and Ebner [15] found, in addition to the barriers mentioned in Belanger and Thornton's [14] study, that student motivation, feelings of isolation and hidden costs are also considered barriers to MOOC-learning. Further, a descriptive analysis of MOOC data to uncover reasons for dropout by Onah, Sinclair and Boyatt [16], showed that difficulty of the MOOC, timing, lack of digital skills and lack of in-MOOC support were often encountered barriers by MOOC-takers. In addition, Hone and El Said [18] explored factors which affect MOOC retention. Their factor analytic study focused on student experiences with the course instructor, experiences with other learners and experiences with the design features of the course and found that especially instructor interaction and course content are important features for students. If these features are not perceived positively by students, they have the potential to become barriers to their learning and ultimately retention.

Also, a very recent study by Shapiro, Lee, Roth, Li, Çetinkaya-Rundel and Canelas [17] on barriers to retention in MOOCs, sought to identify which antecedents, both inside and outside the course setting, had an impact on MOOC-learning. Their qualitative approach of conducting 36 online interviews identified, in order of severity, lack of time, bad previous experiences, online format and inadequate background as barriers to MOOC-learning.

Previous studies confirmed that research on barriers to learning in MOOCs is developing and has strong parallels with the research findings in online learning and distance education context. Still, a shortcoming of prior studies is that they merely examine several specific potential barriers to MOOC-learning and are limited in their empirical analysis. As it is important to continue to explore potential barriers to MOOC-learning to gain a richer understanding of these issues [17, 21], a next step is to generate a composite overview of potential MOOC-specific barriers or groupings of barriers based on literature and related studies as already available in online learning or distance education context [9, 10].

Henderikx et al. [3], composed an overview of potential barriers based on a limited literature review and made a first effort to categorize these barriers (see Fig. 1).

MOOC-related		Non-MOOC related	
<i>Design</i>	Lack of support	<i>General</i>	Lack of information literacy
Problems with the site	Content was not appropriate	Workplace issues	Insufficient academic background
Lack of interaction	<i>Expectations management</i>	Lack of time	Lack of motivation
Lack of instant feedback	Course was too easy	Family issues	Lack of personal commitment
Lack of instructor presence	Course did not meet expectations	Lack of workplace support	<i>Technical</i>
Lack of useful feedback	Course was too difficult	Lack of family support	Technological problems pc
		<i>Personal</i>	Bad internet connection
		Lack of technological skills	

Fig. 1. Overview of barriers arranged by type [3]

The choice for categorization was based on the rationale: which classification would be most useful to MOOC-designers and/or providers and MOOC-takers. The current study took this initial typology of barriers in MOOCs as a starting point. In addition, this overview was expanded by the (potential) barrier items based on findings in the previously discussed literature. An exploratory factor analysis was conducted to empirically summarize the data set and to categorize the barriers.

### 3 Method

#### 3.1 Participants

The participants were individuals who took part in one or more MOOCs in the Spanish language from different MOOC providers in the last 2 years and who indicated that we could contact them for further research, regardless of whether or not they successfully achieved their personal goals in these MOOCs. 1618 Potential respondents received an invitation to participate in the survey of whom 317 actually completed the survey

(163 women, 154 men,  $M_{\text{age}} = 47$ , age range: 20–83 years). Most of the participants hold a master (26.1%) or bachelor (32.9%) degree. 8.1% of the participants have a doctorate degree, while 24.8% have an associate or secondary education degree. The remaining 8.1% of the participants finished middle school or below. 66.1% Of the participants are employed for wages, while 13.9% are self-employed. A further 8.5% is currently looking for work and 1.7% is not looking for work. 3.4% of the participants are students, 0.3 military and 6.1% indicated that were retired or other. A majority of the participants participated in up to 5 MOOCs (45.2%). 27.9% participated in 6 to 10 MOOCs, 17% between 11 and 20 MOOCs and 9.9% between 21 and 100 MOOCs. Furthermore, 58.3% of the participants actually finished between 1 and 5 MOOCs, 23.7% finished between 6 and 10 MOOCs, 10.2% between 11 and 20 MOOCs and 7.8% indicated that they finished between 21 and 80 MOOCs. Lastly, 24.4% of the participants prefer the traditional face-to-face way of learning, 39.3% indicates that it makes no difference to them whether they learn face-to-face or online and 36.3% prefers to learn online. Overall, the sample is similar to samples reported in other research on MOOCs [22].

### 3.2 Materials

A ‘Barriers to MOOC-learning’ survey was developed, which contained items drawn from general online learning, distance education and MOOC-specific context literature on barriers and enablers to learning, as discussed in previous section. After answering several general questions on gender, age, educational background, employment status, MOOC-learning experience and preferred learning context, respondents were asked to indicate to what extent they considered the 44 listed items as barriers to learning in a MOOC on a 5-point Likert scale ranging from ‘to a very large extent’ to ‘not at all’. Examples of items are ‘lack of decent feedback’, ‘family issues’, ‘technical problems with the computer’ and ‘lack of instructor presence’.

### 3.3 Procedure

Over the course of several weeks potential respondents were invited via email batches using the open source online survey tool Limesurvey (visit <http://www.limesurvey.org>). Filling out the questionnaire took 5–10 min. After four and six weeks, a reminder was sent to those who did not yet completed the survey.

### 3.4 Data Screening

The Mahalanobis distance was calculated to identify possible outliers. Based on these calculations, 22 outliers were determined and removed, which resulted in a final sample of 295 cases, which is within the generally accepted item ratio to conduct a factor analysis of 5 to 10 respondents per item [23].

### 3.5 Analysis

The suitability of the data for factor analysis was assessed by first examining the correlation between items. It was observed that all items correlated with at least .3 with one other item, which is a positive indication of factorability. Additionally, the Kaiser-Meyer-Olkin measure showed a value of .95 which exceeded the recommended minimum value of .6 [24, 25] and the Bartlett's Test of Sphericity was statistically significant ( $p < .05$ ), which further supports the factorability of the data. Lastly, the communalities all exceeded .3 (see Fig. 2). Given these indicators, the factorability of the data could be considered positive.

Principal component analysis was selected as extraction method because this method allows for reducing the observed variables to a smaller set of independent composite variables. A cut-off of 0.4 was used for statistical significance of the component loadings and the component structure was examined using both Varimax and Oblimin rotation. After initial analysis, the Oblimin rotation was selected as this rotation method produced the simplest component structure. The Kaiser criterion [26], which retains components with an eigenvalue above 1, and inspection of the scree plot were used to determine the number of components. Yet, as these methods are not considered very accurate [27], parallel analysis was also performed. The first analysis showed the presence of 6 components with eigenvalues above 1, explaining respectively 48,2%, 9,2%, 5,8%, 4,5%, 2,6% and 2,3% of the variance, yet with very few or no loadings in the last two components. The screen plot indicated a break after the 4<sup>th</sup> component. This was further supported by the results of parallel analysis, which produced 4 random eigenvalues smaller than the first 4 eigenvalues of the PCA. Solutions for 4 and 5 components were then examined, also using Oblimin rotation. The 4-component solution, which explained 67,7% of the variability was preferred because of (a) the combined results of the scree plot and the parallel analyses and (b) the reasonably clear interpretable components.

A total of nine items were removed because they did not meet the criteria of no cross-loading of .4 and failed to have a primary component loading of more than .4, thus not contributing to a simple component structure. The items 'Procrastinate (delay), cannot get started', 'Lack of instructor presence', 'Insufficient training/experience to use the delivery system', 'Lack of adequate internet access', 'Lack of technical assistance', 'Technical problems with the site' and 'Lack of language skills' had cross-loadings of more than .4 on multiple components. The items 'Course content was too easy' and 'Course content was too hard' did not load above .4 on any component. Furthermore, two items which seem very similar: 'workplace issues' and 'workplace commitments' were not removed as their mutual correlation was low to medium.

For the final stage, a factor analysis of the remaining 35 items, using the principal component extraction method and oblimin rotation was conducted, forcing four components explaining 70,4% of the variance (see Table 1). All items in this analysis had primary loadings over .4 on one single component. The component loading matrix for this final solution is presented in Fig. 2.

Items	Pattern Matrix Components				Communalities
	1	2	3	4	
1. Lack of skills for using the delivery system	.883				.865
2. Lack of software skills	.882				.849
3. Shy or lack of confidence	.762				.661
4. Unfamiliar with online learning technical tools	.759				.751
5. Lack of information literacy skills	.758				.786
6. Lack of typing skills	.744				.821
7. Lack of reading skills	.661				.791
8. Lack of writing skills	.630				.734
9. Insufficient academic background (prior knowledge)	.610				.678
10. Technical problems with the computer	.505				.668
11. Feeling of isolation		.837			.742
12. Lack of social context cues		.818			.771
13. Learning feels impersonal		.792			.702
14. Lack of student collaboration		.760			.686
15. Lack of interaction/communication among students		.592			.573
16. Prefer to learn in person/face-to-face		.581			.398
17. Lack of clear expectations/instructions			-.840		.793
18. Low quality materials/instruction			-.808		.802
19. Unavailability of course materials			-.732		.560
20. Lack of in-course support			-.732		.698
21. Instructors do not know how to teach online			-.718		.666
22. Lack of interaction with instructor			-.715		.678
23. Lack of timely feedback from instructor			-.700		.699
24. Lack of decent feedback			-.674		.715
25. Course content was bad			-.619		.693
26. Workplace issues				.849	.797
27. Lack of support from employer				.828	.750
28. Too many interruptions during study time				.822	.657
29. Lack of time in general				.796	.650
30. Family issues				.753	.715
31. Lack of support from family, friends				.746	.699
32. Workplace commitments				.617	.570
33. The learning environment is not very motivating				.532	.697
34. Lack of personal motivation				.430	.713
35. Own responsibility for learning				.428	.609

**Fig. 2.** Component loadings and communalities based on a factor analysis with principal component extraction method and oblimin rotation for 35 items (N = 295)

**Table 1.** Total variance explained

Component	Initial Eigenvalues	% of Variance	Cumulative %
	Total		
1	16.72	47.76	47.76
2	3.63	10.37	58.13
3	2.43	6.93	65.06
4	1.86	5.32	70.38

## 4 Results

The data analysis indicated that four distinct components summarized the experienced barriers in MOOCs. Component labels were defined that fitted the extracted component/item-combinations. This resulted in the following labels:

- Component 1: Technical and online-learning related skills. MOOC-takers perceived lack of skills like information literacy, insufficient knowledge of the delivery systems, insufficient academic back ground as barrier to MOOC-learning
- Component 2: Social context. These issues are typically related to learning individually. In other words, not learning in a classical and/or physical learning environment. Issues like the impersonal feel of learning, lack of interaction, no collaboration, no interaction and feelings of isolation are included.
- Component 3: Course design/expectations management. This component concerns barriers related to the design and expectations management of the course like the low quality of the course materials, bad course instruction, no instructor interaction, bad course content and lack of feedback
- Component 4: Time, support and motivation. MOOC-takers experience time constraints due to workplace, family and general issues as well as support issues due to lack of family, peer and work support. Further, motivational issues like being responsible for your own learning and motivation are included in this component

As can be seen in Fig. 2, the majority of the commonalities are reasonably high, which indicates that the extracted components represent the variables well.

The internal consistency for each of the components was tested by calculating the Cronbach's alpha. The alphas were strong: .96 for component 1 (10 items), .882 for component 2 (6 items), .94 for component 3 (9 items) and .94 for component 4 (10 items). Removal of the item 'prefer to learn in person/face-to-face' in factor 2, would slightly improve that Cronbach alpha score to .90, yet as the initial score was already strong it was decided not to eliminate this item.

Furthermore, composite scores were calculated for each of the four components (see Table 2), based on the mean of the items that had their primary loadings on each component. Lower scores indicated that this component represented a more severe barrier to the respective MOOC-takers who completed the survey.

**Table 2.** Means and standard deviations per barrier component and the barrier perceived as most severe (N = 295)

Barrier components	Mean	SD
Technical and online learning skills	3.40	1.19
<i>Technical problems with the computer</i>	3.07	1.41
Social interactions	3.54	0.90
<i>Lack of interaction/communication among students</i>	3.35	1.09
Course design	2.93	1.09
<i>Course content was bad</i>	2.69	1.56
Time, support and motivation	2.95	1.09
<i>Lack of time in general</i>	2.45	1.32

Note: answers were rated on a 5-point Likert scale with 1 = too a very large extent and 5 = not at all

## 5 Discussion

This study has implemented a factor-analytical approach to identify the components that represent the barriers to intention achievement in MOOCs. The iterative process of determining the best fit model, resulted in 4 distinctive components; 1. Technical and online-learning related skills, 2. Social context, 3. Course design/expectations management, 4. Time, support and motivation. This result partly overlaps with a comparable study by Muilenburg and Berge [10], who combined barriers students (expected to) face in an online distance education context into a collective overview for factor analysis. Their analysis found eight components of which *administrative issues and costs* and *access to the internet* were not present in our analysis. The lack of barriers concerning administrative issues can be explained by the fact that we did not include administration related barriers in our questionnaire as the administrative issues in MOOCs as a non-formal learning context are not comparable to administrative issues in formal education. An explanation regarding internet issues can most likely be explained by the fact that the Muilenburg and Berge [10] study collected data in 2003. Internet was less available and affordable then compared to present time where access to the internet is inexpensive and available at practically all places and time using various devices.

Also, our study identified one component with technical related issues and online-learning related skill barriers whereas Muilenburg and Berge [10] found three separate components containing technical and academic skills and technical problems. Further, both studies found a *social interactions/social context* component but *time, support and motivation* barriers are part of one component in our study, while the Muilenburg and Berge [10] study found two components to cover these barriers. Lastly, our study found one distinct component containing MOOC-design related barriers, which is the largest difference compared to Muilenburg and Berge's [10] study that found instructor related issues combined with administrative issues in one component. However, this difference could be explained by the fact that, as stated before, we did not include any administrative related barriers in the questionnaire.



**Table 3.** Classification of barrier components

Component	Label	Type	Coping level
1	Technical and online related skills	Non-MOOC related	Can be dealt with on a personal level
2	Social context	Partly MOOC and partly non-MOOC related	Can be dealt with on both personal and MOOC-level
3	Course design	MOOC related	Can be dealt with on MOOC level
4	Time, support and motivation	Non-MOOC related	Can be dealt with on a personal level

The composite scores per barrier component (see Table 3) indicate that *course design* and *time, support and motivation* are near enough equally considered as most severe barrier components by the respondents of the barriers to MOOC learning questionnaire. *Social context* was rated as least severe barrier. In contrast, Muilenburg and Berge's [10] study found that the *social interactions* component was perceived as most severe. This is quite a big difference in perception, which might also be explained by the moment in time of the study. As online presence is part of everyday life nowadays, people are increasingly used to this phenomenon; in 2003, this was merely emerging.

Further, when looking at the *course design* barrier component, *bad course content* is rated as most severe barrier. Studies by Peltier et al. [12], and Aragon and Johnson [20], in online learning context, found similar results. In the MOOC-learning context, the study by Hone and El Said [18] also identified *course content* as an important feature for course retention. Additionally, the most severe barrier included in the *time, support and motivation* barrier component was *lack of time*. This is consistent with the findings of Song et al. [13] in online learning context and Belanger and Thornton [14] and Shapiro et al. [17] in MOOC-learning context.

When further assessing the literature review, it stands out that instructor related issues are consistently perceived as important for retention in online learning [8, 12, 20]. Yet, in MOOC-learning context this issue is only found by Hone and El Said [18] and in current study this issue was also not perceived as a severe barrier. This is an interesting observation, even though, with the exception of current study, all of these aforementioned studies merely focused on several specifically selected, mainly course related barriers in their research setup. Possibly, learners have higher expectations, or attach more value to, instructor related issues in a formal education context. As MOOCs are easily accessible and do not have a formal education status (yet), instructor issues, might not be perceived as important for a satisfying learning experience.

An assessment of the barrier components in light of the study by Henderikx et al. [3] resulted in Table 3. From Table 3, it can be inferred that the barrier components and thus the experienced barriers by MOOC-takers are predominantly non-MOOC related. This knowledge can be of value for MOOC-designers and providers. It may guide them in finding suitable re-design solutions or interventions to support MOOC-takers in their

learning, even if it concerns non-MOOC related issues. For instance, to support MOOC-takers regarding technical and online-learning related skills, it would be possible to, prior to the start of a MOOC, specifically draw attention to the minimum requirements regarding technical and online learning skills needed to be able to finish the MOOC. The barriers related to social context, that are considered MOOC-related like *lack of interaction* and *lack of collaboration* could be addressed in the design of the MOOC by for instance integrating assignments which demand or support interaction and collaboration with fellow MOOC-takers. Course design related barriers are addressable by re-design interventions depending on the specific issues at hand. Moreover, barriers concerning time, support and motivation could, even though not MOOC-related, be supported by MOOC providers and/or designers by for instance providing information on how to handle and cope with these kinds of barriers, as well as by providing supporting interventions.

There are some limitations that should be taken into account. Firstly, the sample is limited in the sense that it only considers MOOC-takers who took part in one or more MOOCs in the Spanish language. Future research should replicate this study finding respondents in other MOOC-taker populations. Also, we do not know to what extent the respondents who completed the survey were successful in achieving their personal goals when participating in their respective MOOCs. It would be interesting and potentially valuable to differentiate between these two groups to investigate if either group encounters different barriers. Furthermore, even though the item ratio of 6:1 is within the generally accepted limits for factor analysis (Comrey and Lee 1992), a bigger sample will add to the reliability of the analysis. Further research should be conducted using bigger samples to either confirm or contradict our results. Lastly, as this is the first study examining components influencing intention achievement in MOOCs, further refinement of the barrier overview is necessary. A possible next step is to expand this composed barrier overview into an assessment tool for MOOC-providers and/or designers that can support them in their effort to enhance the MOOC-learning experience, in identifying areas for improvement either MOOC related or not.

To conclude, the aim of this research was to empirically analyse barriers that influence intention achievement in MOOCs and translate this for practical purposes into MOOC or non-MOOC related barrier components. The findings identified 4 barrier components of which the majority contained non-MOOC related barriers, which is useful information for MOOC providers and designers and makes a valuable contribution to the expanding empirical MOOC-research.

**Acknowledgement.** This work is financed via a grant by the Dutch National Initiative for Education Research (NRO)/The Netherlands Organisation for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science under the grant nr. 405-15-705 (SOONER/<http://sooner.nu>).

## References

1. Jordan, K.: Initial trends in enrollment and completion of massive open online courses. *Int. Rev. Res. Open Distrib. Learn.* **15**(1), 133–160 (2014). <https://doi.org/10.19173/irrodl.v15i1.1651>
2. Henderikx, M., Kreijns, K., Kalz, M.: Refining success and dropout in MOOCs based on the intention-behavior gap. *Distance Educ.* **38**, 353–368 (2017). <https://doi.org/10.1080/01587919.2017.1369006>
3. Henderikx, M., Kreijns, K., Kalz, M.: To change or not to change? That’s the Question... On MOOC-success, barriers and their implications. In: Delgado Kloos, C., Jermann, P., Pérez-Sanagustín, M., Seaton, D.T., White, S. (eds.) *EMOOCs 2017*. LNCS, vol. 10254, pp. 210–216. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-59044-8\\_25](https://doi.org/10.1007/978-3-319-59044-8_25)
4. Huin, L., Bergheaud, Y., Caron, P.A., Codina, A., Disson, E.: Measuring completion and dropout in MOOCs: a learner-centered model. In: Khalil, M., Ebner, M., Koop, M., Lorenz, A., Kalz, M. (eds.) *Proceedings of the European MOOC Stakeholder Summit 2016*, pp. 55–68. Books on Demand GmbH, Nordstedt (2016)
5. Walji, S., Deacon, A., Small, J., Czerniewicz, L.: Learning through engagement: MOOCs as an emergent form of provision. *Distance Educ.* **37**(2), 208–223 (2016). <https://doi.org/10.1080/01587919.2016.1184400>
6. Kalz, M., Kreijns, K., Walhout, J., Castaño-Munoz, J., Espasa, A., Tovar, E.: Establishing a European cross-provider data collection about open online courses. *Int. Rev. Res. Open Distrib. Learn. (IRRODL)* **16**(6), 62–77 (2015)
7. DeBoer, J., Ho, A.D., Stump, G.S., Breslow, L.: Changing “course” reconceptualizing educational variables for massive open online courses. *Educ. Res.* **43**(2), 74–84 (2014). <https://doi.org/10.3102/0013189X14523038>
8. Eom, S.B., Wen, H.J., Ashill, N.: The determinants of students’ perceived learning outcomes and satisfaction in university online education: an empirical investigation. *Decis. Sci. J. Innov. Educ.* **4**(2), 215–235 (2006). <https://doi.org/10.1111/j.1540-4609.2006.00114.x>
9. Galusha, J.M.: Barriers to learning in distance education. *Interpers. Comput. Technol. Electron. J. 21st Century* **5**(3/4), 6–14 (1998)
10. Muilenburg, L.Y., Berge, Z.L.: Student barriers to online learning: a factor analytic study. *Distance Educ.* **26**(1), 29–48 (2005). <https://doi.org/10.1080/01587910500081269>
11. Park, J.H., Choi, H.J.: Factors influencing adult learners’ decision to drop out or persist in online learning. *Educ. Technol. Soc.* **12**(4), 207–217 (2009)
12. Peltier, J.W., Drago, W., Schibrowsky, J.A.: Virtual communities and the assessment of online marketing education. *J. Mark. Educ.* **25**(3), 260–276 (2003). <https://doi.org/10.1177/0273475303257762>
13. Song, L., Singleton, E.S., Hill, J.R., Koh, M.H.: Improving online learning: student perceptions of useful and challenging characteristics. *Internet High. Educ.* **7**(1), 59–70 (2004). <https://doi.org/10.1016/j.iheduc.2003.11.003>
14. Belanger, Y., Thornton, J.: Bioelectricity: a quantitative approach Duke University’s First MOOC (2013)
15. Khalil, H., Ebner, M.: MOOCs completion rates and possible methods to improve retention - a literature review. In: *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pp. 1236–1244. AACE, Chesapeake (2014)
16. Onah, D.F.O., Sinclair, J.E., Boyatt, R.: Dropout rates of massive open online courses: behavioural patterns. In: *6th International Conference on Education and New Learning Technologies, EDULEARN 2014, Barcelona*, pp. 5825–5834 (2014)

17. Shapiro, H.B., Lee, C.H., Roth, N.E.W., Li, K., Çetinkaya-Rundel, M., Canelas, D.A.: Understanding the massive open online course (MOOC) student experience: an examination of attitudes, motivations, and barriers. *Comput. Educ.* **110**, 35–50 (2017). <https://doi.org/10.1016/j.compedu.2017.03.003>
18. Adamopoulos, P.: What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In: *Proceedings of the Thirty Fourth International Conference on Information Systems, Milan, Italy* (2013)
19. Hone, K.S., El Said, G.R.: Exploring the factors affecting MOOC retention: a survey study. *Comput. Educ.* **98**, 157–168 (2016). <https://doi.org/10.1016/j.compedu.2016.03.016>
20. Aragon, S.R., Johnson, E.S.: Factors influencing completion and noncompletion of community college online courses. *Am. J. Distance Educ.* **22**(3), 146–158 (2008). <https://doi.org/10.1080/08923640802239962>
21. Hew, K.F.: Promoting engagement in online courses: what strategies can we learn from three highly rated MOOCs. *Br. J. Educ. Technol.* **47**(2), 320–341 (2016). <https://doi.org/10.1111/bjet.12235>
22. Ho, A.D., et al.: Harvardx and MITx: Two years of open online courses fall 2012-summer 2014 (2015). [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2586847](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2586847), <https://doi.org/10.2139/ssrn.2586847>
23. Comrey, A.L., Lee, H.B.: *A First Course in Factor Analysis*, 2nd edn. Lawrence Erlbaum, Hillsdale (1992)
24. Kaiser, H.F.: A second-generation little jiffy. *Psychometrika* **35**(4), 401–415 (1970)
25. Kaiser, H.F.: An index of factorial simplicity. *Psychometrika* **39**(1), 31–36 (1974)
26. Kaiser, H.F.: The application of electronic computers to factor analysis. *Educ. Psychol. Measur.* **20**(1), 141–151 (1960)
27. Velicer, W.F., Jackson, D.N.: Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivar. Behav. Res.* **25**(1), 1–28 (1990). [https://doi.org/10.1207/s15327906mbr2501\\_1](https://doi.org/10.1207/s15327906mbr2501_1)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Tools to Support Self-Regulated Learning in Online Environments: Literature Review

Ronald Pérez-Álvarez<sup>1,2(✉)</sup>, Jorge Maldonado-Mahauad<sup>1,3</sup>,  
and Mar Pérez-Sanagustín<sup>1,4(✉)</sup>

<sup>1</sup> Department of Computer Science,

Pontificia Universidad Católica de Chile, Santiago, Chile

{raperez13, jjmaldonado, mar.perez}@uc.cl

<sup>2</sup> University of Costa Rica, Sede Regional del Pacífico, Puntarenas, Costa Rica

<sup>3</sup> Department of Computer Science, University of Cuenca, Cuenca, Ecuador

<sup>4</sup> Université Toulouse III Paul Sabatier, Toulouse, France

mar.perez-sanagustin@irit.fr

**Abstract.** Self-regulated learning (SRL) skills are especially important in Massive Open Online Courses (MOOCs), where teacher guidance is scarce, and learners must engage in their learning process trying to succeed and achieve their learning goals. However, developing SRL strategies is difficult for learners given the autonomy that is required in this kind of courses. In order to support learners on this process, researchers have proposed a variety of tools designed to support certain aspects of self-regulation in online learning environments. Nevertheless, there is a lack of study to understand what the commonalities and differences in terms of design are, what the results in terms of the effect on learners' self-regulation are and which of them could be applied in MOOCs. Those are the questions that should be further explored. In this paper we present a systematic literature review where 22 tools designed to support SRL in online environments were analyzed. Our findings indicate that: (1) most of the studies do not evaluate the effect on learners' SRL strategies; (2) the use of interactive visualizations has a positive effect on learners' motivation; (3) the use of the social comparison component has a positive effect on engagement and time management; and (4) there is a lack of models to match learners' activity with the tools with SRL strategies. Finally, we present the lessons learned for guiding the community in the implementation of tools to support SRL strategies in MOOCs.

**Keywords:** Self-Regulated Learning · Tools · System · Online MOOC · Literature review · Massive Open Online Courses · Dashboard Learning analytics

## 1 Introduction

Recent research shows the importance of self-regulated learning (SRL) in traditional and online learning contexts [1]. Self-Regulated Learning refers to how students become masters of their own learning processes [2]. However, this definition can vary depending on the theoretical model used as a reference as well as the research context or focus of analysis (motivation, cognition, meta-cognition, feelings) [3]. In online contexts, the learners are required to have greater autonomy than in face-to-face classes

and they are expected to be able to deploy SRL strategies in order to achieve their objectives. That is, learners who are able to self-regulate their learning are more likely to succeed in completing courses [4, 5]. Self-regulation skills are even more relevant in a MOOC, which is characterized by the massiveness and heterogeneity of the participants; the lack of guidance from a tutor during the course; and the flexibility of schedules over time [6].

Recent research indicates that some SRL strategies are associated with the learners' performance and achievement of their goals. For example, strategies such as *goal setting* and *strategic planning*, as well as *time management* have been demonstrated to have an influence in performance and fulfillment of the learners' goals [6–8]. Likewise, [7, 8] showed that learners use strategies such as *organization*, *help seeking* and *effort regulation* to when working in a MOOC. However, current MOOC platforms do not offer adequate technological support for the deployment of learners' SRL strategies [9, 10]. For example, the Coursera platform offers the option of consulting the time spent on video lessons. In addition, it has a submission timetable that, together with email notifications, help learners to keep engaged with the course. Despite of this, researchers agreed that these mechanisms are not enough and it is necessary to develop new tools to support SRL in online platforms [11, 12]. Although tools have been developed to support learners' SRL in the context of traditional online learning [13–16], as well as in the MOOC context [17–20], there is a strong mismatch between the goal of the tool and its evaluation [21]. Furthermore, in the case of the MOOC context, the development of this type of tools is new, few tools are implemented, and more evaluations are required in these massive contexts to understand the impact on the learners' self-regulation [22]. The research points out a severe weakness regarding the evaluation of existing tools [22–25], as they focus their evaluation on usability and usefulness [23]; leaving a gap in the measurement of the tool's impact of the SRL strategies that they support.

In this light, the development of new tools aimed at supporting self-regulation in MOOC environments is a challenge that remains open. The lack of evaluations to measure the impact on SRL does not allow us to understand what characteristics should be considered in the design of new tools or how the self-regulation strategies that the learners use with the interactions they perform with the tool are related. In addition, there is no guide for the design, implementation and evaluation of this type of tools.

In this paper, and in order to understand the current state of the art in the development of tools designed to support learners' self-regulatory processes online, we present a systematic literature review that extends a previous work [22], but focusing on: (1) analyzing the relations between learning activities and self-regulation strategies defined in the design of the tools; (2) analyzing the characteristics and indicators used in the tools; and (3) presenting the lessons learned in each of the papers to understand what these tools should be design in a MOOC context.

## 2 Prior Work

In this section we analyze the results of the two literature reviews [21, 26] we found in the area of supporting learners SRL strategies online and summarize the results of our previous study of the literature [22].

Jivet et al. [21, 26] conducted two literature reviews on 26 tools to support learning processes in online environments. Of the 26 tools analyzed, 13 of these were designed for supporting self-regulation in online environments. The results show that SRL is supported through tools that provide learners' awareness and trigger reflection about their learning process. In addition, the authors point out that there is a separation between the purpose of the tool and its evaluation. Although these reviews shed some light on how SRL is addressed, they do not analyze in detail the characteristics of these tools in terms of design, nor the self-regulation strategies that they aim at supporting.

In the a previous literature review [22], we analyzed 21 tools aimed at supporting learners' self-regulation. In this review we analyzed their characteristics in terms of design, the SRL strategies supported, the methodology for their evaluation, and their impact of learners' self-regulation. The main findings are the following: (1) there is a lack of tools to support SRL in MOOC environments; (2) the evaluation of the existing tools is not aligned with the objectives of the research; (3) current research present proposals of tools but very few reach the implementation stage; and (4) current existing tools tend to support many SRL strategies at the same time.

The main gap identified in this prior work is the lack of alignment between the purpose of the tools in supporting self-regulation and the evaluations performed to assess their effectiveness. In this study, we propose to expand the previous literature review with the purpose of providing more insights about the relationship between the design of the tools, and how their functionalities relate with learners' self-regulated strategies in the course. Specifically, we defined 5 research questions to guide the literature review: **RQ1**. What is the context in which each tool has been applied, including the educational level and learning environment?, **RQ2**. What characteristics have been considered for the design of the tools to support the learners' SRL strategies?, **RQ3**. What SRL strategies are supported by these tools?, **RQ4**. How does the design of the tools relate with the learners' self-regulated learning activities? **RQ5**. How was the impact of the tool on learners' self-regulation measured?

### 3 Methodology

For the systematic literature review, we followed the phases proposed by Kitchenham [27]: planning, execution and reporting. However, for this review a process we did not carried out an analysis to determine the quality of the papers, given that the interest of the study is to include as many publications as possible. The search process was conducted in 5 databases were most of the papers in Technology Enhanced Learning can be found: Scopus, ACM Digital Library, IEEE Explorer, SpringerLink and Science Direct. The following keywords were used to formulate the search queries: *Self-Regulated Learning*, *Self-Directed Learning*, *Tools*, *System*, *Dashboard*, *Online*, *MOOCs*. This query is expressed symbolically as: (*Self-Regulated Learning*, *Self-Directed Learning*) AND (*Tools*, *System*, *Dashboard*) AND (*Online OR MOOCs*). The first part of the query focuses detecting articles related to self-regulation; the second part identifies tools proposed or implemented; and the third part identifies the context at which the research has been conducted. The review was conducted by 3 researchers.

Two investigators reviewed and selected the articles and the third investigator intervened in case the two investigators had doubts about the inclusion of an article.

1.829 articles were retrieved according to this search criteria. From these, we conducted a selection probes based on articles' the titles/abstracts and keywords. From this first pool of articles, we excluded those that did not match the following criteria: articles that do not describe a tool, articles that support self-regulation, but not through a tool; tools that support self-regulation, but not in an online environment; articles that addressed the use of tools such as social networks and e-portfolios to support self-regulation, but no development is proposed; and tools that support self-regulation, but are not designed for learners. At the end of this process, we ended up with 42 articles. Then, we eliminated duplicates (11) and conducted the analysis of the whole article. In order to broaden the range of tools analyzed, we also included in the analysis those references that were identified from the references of the articles analyzed (7).

A total of 38 articles was considered for this review. This selection considered articles that describe tools designed for supporting learners' self-regulation in both traditional online learning environments and MOOCs. The articles related to the same tool were counted, but for the analysis they were considered as a single tool. The analysis was performed on 22 tools described in the selected articles. Figure 1 depicts the process selection criteria conducted in this review. Although an important number of data sources were considered for the systematic search, there is a possibility that some publications that propose or implement tools have been left out of the study, which we assume as a limitation.

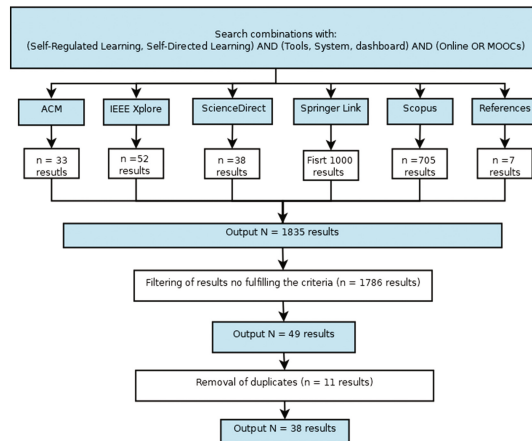


Fig. 1. Papers selections process

## 4 Results

The results are presented to answer each of the research questions posed. A total of 22 tools were analyzed in the literature review (see Table 1). From this pool, 19 tools are implemented and 3 propose only the design of a tool [20, 28, 29].



**Table 1.** Description of tools designed to support learners' SRL in online environments.

Name	Description
LET'S System [34]	It is a system aimed at improving the learners' performance through several theory-based such as real-time screen-sharing, synchronous demonstration, and learners' portfolio monitoring
ROLE [13]	It is a Framework that enables both widgets and learners in the same space to interact with each other. ROLE provide 15 SRL widgets to support learner to search information, planning activities, goal setting, etc.
Meta-Tutor [35]	Learning environment designed to detect, model, trace, and foster learners' SRL about human body system. Learners can generate several subgoals for the session, self-evaluation your knowledge and monitoring you learning process
Learning-B [36]	The <i>Learning-B</i> environment is a prototype aimed at supporting self-regulation in workplace learning. In this environment the learners choose the competences to learn and learning path to reach your learning goals
mCALs [37]	It is a framework, which uses learners' learning schedule to retrieve their location and available time contexts in order to suggest appropriate materials to them based on these, at the time of usage
INNOVRET [38]	Plugin for Moodle to support SRL online. This plugin recommends content according to the learners' current competence state
Video-Mapper [31]	It is a video annotation tool for MOOCs that allows collaborative annotation and supports self-organization
NoteMyProgress [19]	A plugin and a web app to support the learners' SRL in MOOC environments by setting interactive goals and visualizations of their own learning activity within the course resources
i-MySelf ePortafolio [39]	It is a goal-setting plugin to facilitate individuals' capacity for self-regulating their learning, strengthen their motivation and self-efficacy in a ePortafolio
Serious Game [4]	A tool designed to motivate learners' participation in MOOC, through interactive assessment for solving industrial problems
FORGE [30]	This project aims at promoting Self-Regulated Learning (SRL) through the use of a federation of high performance testbeds and at building unique learning paths based on the integration of a rich linked-data ontology
nStudy [14]	Supports learning with resources available on the Internet. Seeks to support SRL processes by tracking learner's searches, creating notes and terms about information in the web pages
Learning Tracker [10]	A widget for the edX MOOC platform that supports learners SRL by displaying indicators related to the learners' performance
Master Grids System [16]	It seeks to integrate SRL with motivation theories, as well as in social comparison. Uses a matrix to show the content of the learners' progress
eLDa [32]	It is a MOOC learning platform that encourages learners to define their learning goals and to establish learning routes

(continued)

**Table 1.** (continued)

Name	Description
Web2.0 SRL [29]	A tool that integrates web2.0 (RSS, Tag, Wiki, Blogs) services to support planning and management
MyLearningMentor [20]	Proposal of design of a mobile application to support planning through guidance and advice in MOOCs
LearnTracker [33]	A mobile application that tracks the time that learners invest on learning activities to support time management
SRL System [12]	A tool for supporting both learners and teachers in the development of their SRL learning skills by a conducive mobile learning environment for them. It tool support collaboration, self-monitoring, goal-setting, and strategic planning
WPAS [40]	A web-based portfolio for planning objectives or milestones and assess progress
Knowledge Visualization [41]	A tool that supports the development of SRL skills through interactive knowledge maps
Virtual Companion [28]	Proposal of widget for MOOCs platforms to support learners in the different phases of the self-regulation process through a combination of techniques of visualization and prompts

#### 4.1 RQ1. What Is the Context in Which Each Tool Has Been Applied, Including the Educational Level and Learning Environment?

9 of the tools were designed for supporting self-regulation in *higher education*; 2 for *high school*; 2 to *professional training*; 5 for *general education* (tools that do not focus on a specific level of education). 4 of the tools *do not specify* the educational level. 14 tools were designed for supporting SRL in *traditional online* learning environments and 8 in *MOOCs* [4, 10, 19, 20, 28, 30–32]. Two of the tools designed for MOOCs are only design proposals, but have not been implemented [20, 28]. 19 of the tools were designed only for the *web*, 3 for *mobile technologies* [12, 20, 33], and only 1 of supports both *web* and *mobile devices* [30].

#### 4.2 RQ2. What Characteristics Have Been Considered for the Design of the Tools to Support the Learners' SRL Strategies?

For analyzing the characteristics of the tools, we took as a references the categories defined by Bodily and Verbert [24]. These categories include: (1) *visualization*, if tool use any type of visualization to display data; (2) *class comparison*, if tool included a system that allowed learners to compare their data with other learners' data; (3) *recommendation*, if tool included a system that provided a recommendation to a learner; (4) *feedback*, if the tool offers *feedback* through text; and (5) *interactivity*, if it offers the possibility of clicking and exploring its data. In addition, two categories were included, (6) *collaboration*, if tool included a system that learners shared materials or knowledge (7) *input forms*, if the tool has forms for data entry. In Table 2, shows a summary of the categories identified in the analysis.

**Table 2.** Functionality and types of indicators identified in the tools (Link to the complete list of indicators identified in the tools <https://drive.google.com/open?id=1-U2xEnelilQPKZjL-OnZ7rHyA71bKxK-5W8XaLsGnmK>).

Functionality	Freq.	Papers	Type of indicator to support SRL	Freq.	Papers
Visualization	14	[10, 12–14, 16, 19, 20, 28, 31–33, 36, 38, 41]	action-related	13	[10, 12–14, 16, 19, 28, 32, 33, 35–37, 40]
Colaboration	11	[12–14, 29, 31, 32, 34–36, 40, 41]	content-related	13	[10, 12–14, 16, 19, 28, 30, 36–39]
Input forms	10	[12, 13, 19, 20, 33, 35–38, 40]	results-related	10	[4, 12, 13, 16, 19, 20, 28, 36, 38, 39]
Recomendation	9	[12, 13, 20, 28, 36–38, 41]	learner-related	1	[36]
Class comparison	5	[10, 16, 19, 33, 36]	social-related	1	[36]
Text feedback	4	[12, 16, 35, 36]	context-related	1	[30]
Interatectivity	4	[13, 16, 19, 36]	Others	1	[30]

*Visualization:* 13 tools use some type of visualization to support self-regulation strategies. The progress or interaction of the learner with the activities is displayed through using *graphs, tables, networks, calendars* or *progress bars* [10, 12–14, 16, 19, 28, 31–33, 36, 38, 41]. Visualizations such as *conceptual maps* are used to present the objectives produced by the learners [14, 31].

*Class comparison:* 5 of the tools report the use of social comparison components to support self-regulation. The tools offer mechanisms for the learners to compare their performance with the performance of their classmates [16, 33], or with the learners from previous editions [10, 19].

*Recommendation:* 9 of the tools use recommendation mechanisms. They recommend learning objectives or activities [12, 13, 28, 36–38], learning routes [36], strategies or tips for SRL [12, 20, 41], and the use of tools (widgets) [30].

*Feedback:* 4 tools offer textual feedback to the learners through motivational messages for performing an activity [36], presenting the correct answers to an exercise [16], time invested [35], or sending notifications [11].

*Interactivity:* 4 tools allow some kind of interactivity with the information presented to the learners. Learners can interact with the information and select the activity to analyze [13, 16, 19, 36], and activate or disable the social comparison [16, 19].

*Colaboration:* 11 tools integrate collaboration mechanisms that support learners' help seeking. Among these mechanisms are: the use of social networks, wikis or blogs [12, 13], discussion forums [13, 32], shared learning spaces [13, 14], and sharing of learning resources for getting feedback [12, 14, 31, 34, 36].

*Input forms:* 10 tools use some mechanism for allowing data entry by the learner. Learners can define and plan their goals [12, 13, 19, 33, 35–38, 40], record the time of an interruption in the study and the reason for the interruption [12], record the

beginning and the end of an activity [33], and record the level of completeness of the activities [40]. In addition, 5 tools propose the use of widgets or plugins to support learners' SRL [10, 13, 19, 28, 30].

To analyze the type of indicators proposed to support SRL, we categorized them according to the 6 groups proposed by Schwendimann et al. [25]: (1) action-related; (2) content-related; (3) results-related; (4) social-related; (5) context-related; and (6) learner-related. A total of 78 indicators were identified. Most of the indicators fall into two categories: action-related (30 indicators) and content-related (34 indicators). 13 tools use the action-related category and the same number of tools use the content-related category (Table 2). 10 of the tools used results-related indicators.

### 4.3 RQ3. What SRL Strategies Are Supported by These Tools?

For tools dedicated to traditional learning environments we identified 10 SRL strategies that are generally supported:

- *Goal setting*: present in 14 tools [4, 10, 30, 33, 40], those that implement mechanisms so that the learners can set their learning goals such as the selection of skills to develop [36] or the definition of activities to be developed on certain dates [13, 19, 20, 28, 34].
- *Self-evaluation*: present in 12 tools. The *self-evaluation* strategy is interpreted from two perspectives in the tools. First, to provide feedback when the learners complete the evaluation activities suggested in the course [4, 16, 31, 32, 34, 35, 41], and second, to provide learners' with information to evaluate their progress in their activities [12, 13, 19, 20, 33, 35–38, 40].
- *Help seeking* and *organization*: they are supported in 9 tools [13–15, 29, 31, 34, 41]. *Help seeking* is generally supported by enabling shared spaces, forums, chats or by integrating social networks. *Organization* is supported through the use of notebooks or supporting the generation of concept maps for content organization.
- *Self-efficacy* is supported in one tools [13] and *self-motivation* is supported in 2 tools [37, 39].

For tools dedicated specifically for supporting SRL in MOOCs, we identified 7 strategies as the most supported: (1) *goal setting* [19, 20, 28, 30, 32], which remains the most supported, (2) *time management* [10, 19, 20, 28], (3) *help seeking* [31] being the least supported strategy. The support of SRL strategies in MOOC is consistent with what the literature points out, as *goal setting*, *strategic planning* and *time management* are strategies shown as effective to achieve learners' objectives [6–8]. *Time management* is generally supported by displaying the time invested by the learners in the activities in study sessions [10, 19, 35], and *procrastination* [10, 19]. *Time management* is also supported through the scheduling and organization of activities [20, 35].

### 4.4 RQ4. How Was the Impact of the Tool on Learner' Self-Regulation Measured?

In 19 of the tools analyzed, it is not described how the design of the tool establishes a relationship between the activities of the learners and the SRL strategies that it tries to

support. Only 3 of the tools describe some type of relationship between the activities and SRL strategies. For example, in [12] there is a diagram with 7 different transition states that the learners can perform in the tool. In Fig. 2, an example of two states of the diagram are shown. The states are associated with the SRL phases of the Zimmerman model [3]. The transitions indicate specific activities that the learners perform interacting with the tool functionalities. In this way the transitions between one state and another allow to relate the activity with a self-regulation phase. However, the information about user transitions is not used for evaluating the effectiveness of the tool but for representing the learners' interaction with it.

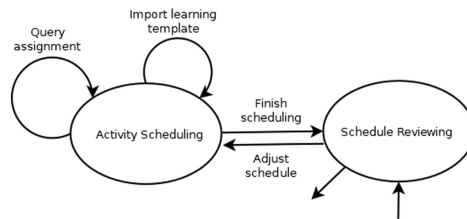


Fig. 2. State transition diagram to SRL, extracted from [12]

In another case such as [13] SRL activities are defined by learners. They define 7 groups or categories connected with the tool functionalities: (1) *Search & Get Recommendation*, (2) *Plan & Organize*, (3) *Communicate & Collaborate*, (4) *Create & Modify*, (5) *Train & Test*, (6) *Explore & View Content*, and (7) *Reflect & Evaluate*. Each group of features is associated with one of the phases of the SRL model that the tool is based on. The learners have the option of classifying the activity performed with a widget within one of these functional groups, thus trying to relate the activities performed by the learners to one of the phases of SRL.

In [36] an approximation is made relating the strategies of SRL with the tools' functionalities, in order to evaluate the usefulness perceived by the learners in the execution of self-regulation. The goal setting strategy is associated with the recommendation feature and the delivery of the information useful for the learner. The monitoring strategy is associated with the delivery of the information useful for the learner.

#### 4.5 RQ5. How Was the Impact of the Tool on Learner' Self-Regulation Measured?

The evaluations of the tools implemented focused on measuring aspects such as: usability (6), usefulness (4), satisfaction (4), and learning outcomes (4). However, this section presents the evaluations that proposed measures for analyzing the impact of the intervention with the tools on learners' behavior or performance. Three of the tools designed for MOOC assess the impact on the learners' behavior and completeness rate. In [10], the impact of the tool on the learners' behavior is measured with respect to evaluations. The results show a positive effect in the assignments delivery times, with

the learners sending evaluations in advance. However, the authors point out that no evidence of changes in the learners' behavior was found. In [4, 10], the impact of the tool is measured by the learners' completion rate. In both cases the results show an increase in the completion rate.

In [33] the authors measure the impact of the tool in learners' Time Management strategies using the Online Self-Regulated Learning Questionnaire (OSLQ). The results show a positive effect on the learners' time management ability with the use of the tool and the social comparison component. In [37] the monitoring of the schedule defined by the learners was analyzed as a measure of time management. As a result, it was observed that the learners who closely follow their schedule and prioritize their studies against other activities, usually work harder. In [39] the activities performed by the learners to manage their time and monitor their learning were analyzed as a measure of the impact on learners' performance. The results show that the execution of these activities minimizes the opportunities for interruption and loss of discipline at the time of studying.

In [40], a pre and post self-report test about self-regulation is used to measure the effect of the goal setting functionality included in the tool. In [16], the effect of the social comparison on the learners' engagement, performance, navigation and motivational profile is evaluated. The results indicate a positive effect of the social comparison component on engagement, efficiency, effectiveness and motivation.

In [13], authors analyze the interaction of the learners with the widgets (15 base widgets) developed to support SRL. The results show that few learners use SRL widgets. In the spaces where the learners add at least one SRL widget, the classification of Plan & Organize and Reflection & Evaluation is used, while in the other spaces, the Collaborate & Communicate classification is more frequent. Finally, the authors concluded that SRL is a new concept for the learners and the evaluation of the impact of the SRL on the learners requires long-term studies. In [35], the navigation of the learners was evaluated, and it was observed that the group that performed a non-linear navigation had a higher learning output. In addition, the time invested by the learners in the use of each strategy was evaluated and it was found that the learners usually spend more time on ineffective learning strategies used to select, organize and integrate multiple representations of the topics. Finally, in [34], the scores of the learners' evaluations were analyzed. The results show that the graphic and interactive visualization of the concepts of study contribute to improving the programming ability of the learners. In addition, a pre and post test was used to evaluate the impact on cognitive and meta-cognitive self-regulation strategies. The results show that learners improved their cognitive and meta-cognitive strategies.

## 5 Lessons Learned

In this study we have performed an analysis of tools that support learners' SRL in online contexts in order to understand how to develop tools that support these strategies in MOOCs. As a result of this analysis, we highlight three of the lessons learned that could help inform the development of future tools to support self-regulation strategies in MOOC-type of learning environments.

### **5.1 Visual Mechanisms, Interactive, and Social Comparison**

The tools use different mechanisms to support self-regulation of learners: visualizations, social comparison, recommendation, collaboration, and interfaces for data entry. The results show that tools that use visualization and allow some type of interactivity have a positive effect on learners' motivation. In the learning environment of MOOCs this can be an important mechanism to maintain learners' motivation. The social comparison component also has a positive effect on both the MOOCs environment and the traditional online environment. The effect is reflected in the time management and the commitment of the learners. This is a mechanism that must be explored in greater detail to measure its impact on learners' performance and behavior. In addition, in the context of MOOCs, it is necessary to analyze which comparison parameters have the greatest effect on learners, for example, comparing their performance with the learners from the previous editions or the same edition.

### **5.2 Design of the Tool Related to Self-Regulation Strategies**

The purpose of supporting the learners' SRL strategies is clear in all the tools analyzed. However, the design of the tools does not seem to have a clear connection to this purpose. The description of the tools focuses on explaining the features or mechanisms included in the tool, without offering enough detail about how the activities performed by the learners with these mechanisms support specific SRL strategies. The design stage of the tool should be more relevant than the implementation itself. In this stage it is necessary to establish clear relations between the activities performed by the learners, a specific SRL strategy and how the tool enhances support these activities. It is necessary designing the tool according to a theoretical-based model so as to define and integrate functionalities towards the strategies defined in the model. There is a lack of evaluations that relate learners' activities with the tool functionalities and SRL. For example, a tool aimed at supporting Time management evaluates its impact through the learners' self-report, without analyzing the planning and behavior changes of the learners regarding time spent on activities.

The report of the tools should detail the indicators used to measure the self-regulation activities of the learners. Characteristics are presented, but the indicators and how they relate to self-regulation strategies are not specified. The results show that the tools collect a lot of indicators about the learners' events on the platform and about the content. However, few of these indicators are used to evaluate the tool. Future work should consider the evaluation methods in advance and define the indicators carefully. The indicators must be defined during the design process of the tool and associated with learners' self-regulation strategies defined in the theoretical model taken as a reference.

### **5.3 Evaluations Aligned with the Purpose of the Tool**

Most of tools are evaluated in terms of usability and usefulness. However, there is little research on the impact of tools on learners' self-regulation behavior. In addition, few mechanisms that measure this impact are present in current studies. The self-report questionnaires are the instruments more frequently used to evaluate the impact of the

tool on the learners' SRL. However, new evaluation proposals are required to understand how the tool contributes to supporting self-regulation and learners' performance. For example, and since *goal setting* is one of the most common strategies supported in the tools analyzed, the evaluations could focus on analyzing the behavioral patterns from learners' traces, with respect to their goal setting, the fulfillment of the goals, the gap between the goals established and reached, or the percentage of the goals achieved. The learners' interaction with the SRL mechanisms implemented in the tools should be monitored in order to find correlations with performance. In addition, researchers should consider from the beginning what is the association between the activities performed by the learners with the tool, and the strategies of SRL so as to facilitate evaluation processes. Only few works propose this relationship, and most of the tools evaluations are poor. Finally, tools should be evaluated in actual learning environments, with actual users. Studies with controlled and small groups should be limited to test the tools, but not to evaluate its impact. This scenario is even more important on the tools that support self-regulation in MOOCs courses, given that the characteristics of the learners are more particularly heterogeneous.

## 6 Conclusion

In this literature review, we analyze the relation defined between the activities performed by the learners and the SRL strategies that the tools support. The results indicate that only few researchers define this relationship and, consequently, it difficult to evaluate what is the impact of the tool in learners' SRL strategies. Further, evaluating the impact of the tool should be based in both self-reported questionnaires and actual interaction patterns of learners' activity with the online environment, the specific tool and their learning outcomes or performance.

In the MOOC context, there are already some tools designed to support SRL. However, most of these tools have not been evaluated in terms of impact on learners' strategies. The design of the future tools should be based on a clear relationship between learners' activities and SRL strategies to facilitate measuring their impact. The great challenge in the MOOC context will be how to measure the impact in the short and medium term, since most of the courses are only from 5 to 10 weeks. As future work, the features identified in the different tools were analyzed could serve as a guideline to evaluate tools for supporting SRL in MOOCs or online learning environments.

**Acknowledgments.** This work was supported by FONDECYT (11150231), University of Costa Rica (UCR), MOOC-Maker (561533-EPP-1-2015-1-ESEPPKA2-CBHE-JP, LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), CONICYT Doctorado Nacional 2017/21170467, CONICYT Doctorado Nacional 2016/21160081.



## References

1. Adam, N.L., Alzahri, F.B., Cik Soh, S., Abu Bakar, N., Mohamad Kamal, N.A.: Self-regulated learning and online learning: a systematic review. In: Badioze, Z.H., et al. (eds.) *Advances in Visual Informatics, IVIC 2017. Lecture Notes in Computer Science*, pp. 143–154. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-70010-6\\_14](https://doi.org/10.1007/978-3-319-70010-6_14)
2. Zimmerman, B.J.: Self-regulated learning: theories, measures, and outcomes. *Int. Encycl. Soc. Behav. Sci.*, 541–546 (2015)
3. Panadero, E.: A review of self-regulated learning: six models and four directions for research. *Front. Psychol.* **8**, 422 (2017)
4. Thirouard, M., Bernaert, O., Dhorne, L., Bianchi, S., Pidol, L., Petit, Y.: Learning by doing: integrating a serious game in a MOOC to promote new skills. In: *Proceedings of the Second MOOC European Stakeholders Summit, EMOOCs*, pp. 92–6 (2015)
5. Siadaty, M., et al.: Self-regulated workplace learning: a pedagogical framework and semantic web-based environment. *J. Educ. Technol. Soc.* **15**(4), 75–88 (2012)
6. Kizilcec, R.F., Pérez-Sanagustín, M., Maldonado, J.J.: Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Comput. Educ.* **104**, 18–33 (2017)
7. Veletsianos, G., Reich, J., Pasquini, L.A.: The life between big data log events. *AERA Open* **2**(3), 1–10 (2016)
8. Lee, D., Watson, S.L., Watson, W.R.: Systematic literature review on self-regulated learning in massive open online courses. *Australas. J. Educ. Technol.* **35**(1), 1449–5554 (2019)
9. Park, T., Cha, H., Lee, G.: A study on design guidelines of learning analytics to facilitate self-regulated learning in MOOCs. *Educ. Technol. Int.* **17**(1), 117–150 (2016)
10. Davis, D., Chen, G., Jivet, I., Hauff, C., Houben, G.: Encouraging metacognition & self-regulation in MOOCs through increased learner feedback. In: *Proceedings of CEUR Workshop, LAK 2016*, pp. 17–22 (2016)
11. Müller, N., Faltin, N.: IT-support for self-regulated learning and reflection on the learning process. In: *Proceedings of 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW 2011*, pp. 1–6. ACM Press (2011)
12. Shih, K.-P., Chen, H.-C., Chang, C.-Y., Kao, T.-C.: The development and implementation of scaffolding-based self-regulated learning system for e/m-learning. *Educ. Technol. Soc.* **1**, 80–93 (2010)
13. Nussbaumer, A., Kravcik, M., Renzel, D., Klamma, R., Berthold, M., Albert, D.: A Framework for Facilitating Self-Regulation in Responsive Open Learning Environments. *arXiv Preprint* (2014)
14. Winne, P.H., Hadwin, A.F.: nStudy: tracing and supporting self-regulated learning in the internet. In: Azevedo, R., Aleven, V. (eds.) *International Handbook of Metacognition and Learning Technologies*. SIHE, vol. 28, pp. 293–308. Springer, New York (2013). [https://doi.org/10.1007/978-1-4419-5546-3\\_20](https://doi.org/10.1007/978-1-4419-5546-3_20)
15. Azevedo, R., et al.: MetaTutor: Analyzing Self-Regulated Learning in a Tutoring System for Biology. In: *Proceedings of AIED*, pp. 635–637 (2009)
16. Guerra, J., Hosseini, R., Somyurek, S., Brusilovsky, P.: An intelligent interface for learning content: combining an open learner model and social comparison to support self-regulated learning and engagement. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI 2016*, pp. 152–63 (2016)
17. Broadbent, J., Poon, W.L.: Self-regulated learning strategies & academic achievement in online higher education learning environments: a systematic review. *Internet High. Educ.* **27**, 1–13 (2015)

18. Sunar, A.S., Abdullah N.A., White, S., Davis, H.C.: Personalisation of MOOCs: the state of the art. In: Proceedings of the 7th International Conference on Computer Supported Education, pp. 88–97 (2015)
19. Pérez-Álvarez, R., Maldonado-Mahauad, J.J., Sapunar-Opazo, D., Pérez-Sanagustín, M.: NoteMyProgress: a tool to support learners' self-regulated learning strategies in MOOC environments. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 460–466. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_43](https://doi.org/10.1007/978-3-319-66610-5_43)
20. Alario-Hoyos, C., Estévez-Ayres, I., Sanagustín, M.P., Leony, D., Kloos, C.D.: MyLearningMentor: a mobile app to support learners participating in MOOCs. *J. Univ. Comput. Sci.* **21**(5), 735–753 (2015)
21. Jivet, I., Scheffel, M., Specht, M., Drachler, H.: License to evaluate: preparing learning analytics dashboards for educational practice. In: International Conference on LA and Knowledge LAK 2018, pp. 31–40 (2018)
22. Perez-Sanagustín, R., Maldonado, M.J.J.: How to design tools for supporting self-regulated learning in MOOCs? Lessons learned from a literature review from 2008 and 2016. In: 2016 XLII Latin American Computing Conference (CLEI), pp. 1–12 (2016)
23. Verbert, K., Govaerts, S., Duval, E., Santos, J.L., Van Assche, F., Parra, G., et al.: Learning dashboards: an overview and future research opportunities. *Pers. Ubiquitous Comput.* **18**, 1499–1514 (2014)
24. Bodily, R., Verbert, K.: Trends and issues in student-facing learning analytics reporting systems research. In: Proceedings of Seventh International Learning Analytics & Knowledge Conference - LAK 2017, pp. 309–18 (2017)
25. Schwendimann, B.A., et al.: Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Trans. Learn. Technol.* **10**, 30–41 (2017)
26. Jivet, I., Scheffel, M., Drachler, H., Specht, M.: Awareness is not enough: pitfalls of learning analytics dashboards in the educational practice. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 82–96. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_7](https://doi.org/10.1007/978-3-319-66610-5_7)
27. Kitchenham, B.: Procedures for Performing Systematic Reviews, pp. 1–26. Keele University 24(TR/SE-0401) (2004)
28. Sambe, G., Bouchet, F., Labat, J.-M.: Towards a conceptual framework to scaffold self-regulation in a MOOC. In: M. F. Kebe, C., Gueye, A., Ndiaye, A. (eds.) InterSol/CNRIA - 2017. LNICST, vol. 204, pp. 245–256. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-72965-7\\_23](https://doi.org/10.1007/978-3-319-72965-7_23)
29. Tang, Y., Fan, A.: An integrated approach to self-regulated learning platform enhanced with Web 2.0 technology. In: 2011 IEEE 3rd International Conference on Communication Software and Networks, pp. 236–239 (2011)
30. Marquez-Barja, J.M., et al.: FORGE: enhancing elearning and research in ICT through remote experimentation. In: Proceedings of Global Engineering Education Conference, pp. 1157–1163 (2014)
31. Mohamed, A., Yousef, F., Chatti, M.A., Danoyan, N., Thüs, H.: Video-mapper : a video annotation tool to support collaborative learning in MOOCs video-mapper design. In: Proceedings of the Third European MOOCs Stakehold. pp. 131–40 (2015, Summit)
32. Onah, D.F.O., Sinclair, J.E.: A multi-dimensional investigation of self-regulated learning in a blended classroom context: a case study on eLDa MOOC. In: Proceedings of the Advances in Intelligent Systems and Computing, pp. 63–85 (2017)
33. Tabuenca, B., Kalz, M., Drachler, H., Specht, M.: Time will tell: the role of mobile learning analytics in self-regulated learning. *Comput. Educ.* **89**, 53–74 (2015)

34. Huang, T.C., et al.: Developing a self-regulated oriented online programming teaching and learning system. In: Proceedings of IEEE International Conference Teaching, Assessment, and Learning for Engineering, Learning for the Future Now, TALE 2014, pp. 115–120 (2015)
35. Azevedo, R., Johnson, A., Chauncey, A., Burkett, C.: Self-regulated learning with metatutor: advancing the science of learning with metacognitive tools. In: Khine, M., Saleh, I. (eds.) *New Science of Learning*, pp. 225–247. Springer, New York (2010). [https://doi.org/10.1007/978-1-4419-5716-0\\_11](https://doi.org/10.1007/978-1-4419-5716-0_11)
36. Siadaty, M., et al.: Learn-B: a social analytics-enabled tool for self-regulated workplace learning. In: Proceedings of LAK 2012, pp. 115–119 (2012)
37. Yau, J.Y.-K., Joy, M.: A self-regulated learning approach: a mobile context-aware and adaptive learning schedule (mCALS) tool. *Int. J. Interact. Mob. Technol.* **2**, 52–57 (2008)
38. Kopeinik, S., Nussbaumer, A., Winter, L.C., Albert, D., Dimache, A., Roche, T.: Combining self-regulation and competence-based guidance to personalise the learning experience in moodle. In: Proceedings of the IEEE 14th International Conference on Advanced Learning Technologies, ICALT 2014, pp. 62–64 (2014)
39. Alexiou, A., Paraskeva, F.: Managing time through a self-regulated oriented eportfolio for undergraduate students. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) *EC-TEL 2015. LNCS*, vol. 9307, pp. 547–550. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24258-3\\_56](https://doi.org/10.1007/978-3-319-24258-3_56)
40. Chang, C.-C., Tseng, K.-H., Liang, C., Liao, Y.-M.: Constructing and evaluating online goal-setting mechanisms in web-based portfolio assessment system for facilitating self-regulated learning. *Comput. Educ.* **69**, 237–249 (2013)
41. Wang, M., Peng, J., Cheng, B., Zhou, H., Liu, J.: Knowledge visualization for self-regulated learning. *Educ. Technol. Soc.* **14**(3), 28–42 (2011)



# The Psychometric Properties of a Preliminary Social Presence Measure Using Rasch Analysis

Karel Kreijns<sup>1</sup>(✉), Joshua Weidlich<sup>2</sup>, and Kamakshi Rajagopal<sup>1</sup>

<sup>1</sup> Open Universiteit Nederland,

Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands

{karel.kreijns, kamakshi.rajagopal}@ou.nl

<sup>2</sup> FernUniversität Hagen, Universitätsstraße 33, 58084 Hagen, Germany

joshua.weidlich@fernuni-hagen.de

**Abstract.** Social presence is an important construct in computer mediated communication, such as found in online collaborative learning (OCL) settings. It is hypothesized that social presence influences the degree of perceived learning and learning outcomes of OCL group members. However, the construct social presence is contested as many incompatible definitions exist in the research community and so do the many measures of social presence. Also, none of the existing social presence measures has undergone a rigid construct validation process such as proposed by Rasch Measurement theory. As a result, hypothesis testing using these measures produced unreliable findings. To address this undesirable situation, we returned to the original definition of Short et al. [29] and redefined it as the degree to which the other person is perceived as physical ‘real’ in the communication. We present a social presence measure that assesses this perception of realness. Rasch analysis was used to validate the raw social presence measure. Our findings revealed that measuring the degree of realness was excellent for those who have high perceptions of realness of the other (i.e., they could be well differentiated), whereas this was moderate for those who have low perceptions (i.e., they could be less well differentiated). Our conclusion is that the social presence measure is already an improvement when compared to existing social presence measures that emphasize realness but it surely needs further improvement: those who have low perceptions of realness should equally well be differentiated as those with high perceptions of it.

**Keywords:** Online collaborative learning · Rash measurement model  
Social presence theory · Social presence measure

## 1 Introduction

Social presence was originally defined in 1976 by Short et al. [29; p. 65] as “the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationship.” Social presence as salience of the other person was the critical element to differentiate between the various communication media with respect to their potential for establishing interpersonal relationships. They found face-to-face meetings to have the highest degree of social presence, then closed-circuit video channels followed by audio channels; telephone had the lowest degree of social

presence. Even today, in a world full of technology, where all the communication, coordination, and collaboration are increasingly taking place through different kinds of advanced computer mediated communication tools including social media tools<sup>1</sup>, social presence still is an important construct as evidenced by the many publications that explored social presence in these tools (see, for example: [12, 13]). Educational researchers who investigated the impact of those advanced computer mediated communication tools on learning in online collaborative learning (OCL) settings were attracted to the social presence construct. This was because in OCL settings social presence was affecting the way how persons communicate with each other and for how long and, thus, in establishing interpersonal relationships. Interpersonal relationships among the OCL group members were found to be important for knowledge sharing and knowledge co-construction. They, thus, found that social presence affected OCL learning experiences and learning outcomes, especially when all communication and collaboration is taking place a-synchronously rather than synchronously [20, 27, 31].

However, since its conception, social presence has undergone many reformulations and interpretations of what the construct should be [16, 23, 36]. Lowenthal [22; p. 125] pointed out that “despite its intuitive appeal, researchers and practitioners alike often define and conceptualize this popular construct differently. In fact, it is often hard to distinguish between whether someone is talking about social interaction, immediacy, intimacy, emotion, and/or connectedness when they talk about social presence.” As a result, (1) the set of factors potentially affecting the degree of perceived social presence and (2) the measurement of the degree of social presence may vary from the one definition to the other. Kreijns et al. [16] have outlined these issues extensively as did Lowenthal and Snelson [23]. These issues, of course, makes it difficult to compare current findings in the social presence domain and future research is at risk if the confounding situation continues to exist. In particular, for Lowenthal and Snelson [23] it remains a question whether social presence is indeed influencing the degree of perceived learning and learning outcomes as stated by so many social presence researchers.

The aim of the current study was to address the undesirable situation of different definitions and different measures. It was not our aim to question or test relationships wherein social presence play an important role—such as providing evidence that social presence has effect on learning—but (1) to present an operationalizable definition of social presence whose semantic content matches the meaning of the original definition of social presence given by Short et al. [29], and (2) to present a solid social presence measure with good psychometric qualities. To reach our aim, we turned to Short, Williams, and Christie’s original definition as a starting point and concentrated us on the first part of their definition, namely ‘degree of salience of the other person in the interaction.’ With this first part, Short et al. [29] meant the degree to which the other person is perceived as physical ‘real.’ They held the physical attributes of the media as determinative for the degree of realness. In other words, they saw the objective qualities of the communicating medium to be responsible for reconstituting the other in the

---

<sup>1</sup> Examples of social media tools are Whatsapp (<http://www.whatsapp.com>), Yahoo! Groups (<http://groups.yahoo.com>), Skype (<http://www.skype.com>), Instagram (<http://www.instagram.com>), and Facebook (<http://www.facebook.com>).

communication; ideally, the other should be as real as in face-to-face settings. Their position about realness is even clearer when Short et al. [29; pv] expressed their expectation in 1976 that “[i]t is within the scope of foreseeable technology to reconstitute by electronic means a virtual three-dimensional representation of an individual who is hundreds of miles distant.” They, perhaps, saw such representation as the highest form of physical realness of the other. Nowadays, three-dimensional representations of others are reality, enabling holographic communication<sup>2</sup>. Based on all this, we redefined social presence as the degree to which the other person is perceived as physical ‘real’ in the communication<sup>3</sup>. This definition also made social presence operationalizable as items can be constructed to tap realness. By defining social presence as realness of the other, we joined to the stream of social presence researchers who have similar definitions of social presence. For example, Gunawardena and Zittle [11; p. 9] defined social presence as “the degree to which a person is perceived as a ‘real person’ in mediated communication” and Abdullah [1; p. 3] stated that social presence “can be understood as a sense that online users have of the communicators being ‘real’ interlocutors with personalities and physical presence [...]. In other words, an interlocutor’s [social presence] is like the impression one would have of him or her if that interlocutor were physically present in the communication.”

With regard to the second part of Short et al. [29] definition, namely ‘the consequent salience of the interpersonal relationship,’ Kreijns et al. [16] saw it as pointing to another construct, which they identified as ‘social space.’ Social space is the network of interpersonal relationships that exists among communicating persons (e.g., the OCL group members), which is embedded in group structures of norms and values, rules and roles, beliefs and ideals [16; p. 11]. A sound social space is manifest when it is characterized by sense of belonging, feeling of connectedness, mutual trust, open atmosphere, shared social identity, and sense of community. For many social presence researchers, these features were the reason to formulate alternative definitions of social presence and which has led to the contested situation of many incompatible social presence definitions. For example, Garrison [9; p. 352] defines social presence as “the ability of participants to identify with the community (e.g., course or study), communicate purposefully in a trusting environment, and develop interpersonal relationships by way of projecting their individual personalities.” Other examples of alternative definitions can be found in Lowenthal and Snelson [23].

Going back to social presence as realness of the other, we saw social presence not only determined by the physical attributes of the medium—as Short et al. [29] did—but also by many other factors, such as social context, subject of the conversation, the identity of the communicating partner, and online communication style [33, 34].

---

<sup>2</sup> See, for example, <http://research.microsoft.com/holoportation>.

<sup>3</sup> An additional benefit of adhering to social presence as realness of the other also enables us to investigate social presence in the context of virtual reality (VR) or augmented reality (AR) settings as it is compatible with the concept of telepresence advanced by telepresence researchers [e.g., 4]. Lombart and Ditton [21], for example, defined telepresence as “the perceptual illusion of non-mediation [of the other].” Rosakranse et al. [26] explored the role of social presence in VR settings and Kim et al. [14] in an AR-based telecommunication system.

Based on our definition of social presence, a measure was developed and validated using Rasch analysis techniques [7, 25, 40].

Kreijns et al. [15] already presented a social presence measure addressing the realness of the other in the communication and in which a distinction was made between a synchronous and an a-synchronous communication setting. We believe this distinction is superfluous: according to the Rasch Measurement Model, a measure should be invariant across settings and situations. Furthermore, its validation was accomplished through the use of principal component analysis which may be compromised because of the use of Likert scales; these scales—more often than not—are nonlinear as they are ordinal and not interval measures and principal component analysis depends on interval measures [7, 30, 38]. Rasch analysis will ultimately produce more robust measures than can be achieved by applying the usual statistical analyses such as principal component analysis or factor analysis on the item scores of a raw social presence measure. Finally, the small number of five items in this measure may point to a potential under represent of the social presence construct [24]. Other social presence measures may suffer similar and other issues which motivated us to present a new social presence measure. These other issues are described in the next section describing the construction of the social presence measure.

The structure of the paper is as follows. First, we present the raw social presence measure and explain how it aligns with the Rasch Measurement Model. After describing the sample and data collection, the raw social presence scale is validated using the Rasch analysis and results are reported. The paper concludes with a number of limitations to be tackled in future research regarding social presence measurement development.

## 2 Construction of the Social Presence Measure

As stated in the Introduction section, we defined social presence as the degree to which the other person is perceived as physical ‘real’ in the communication. Accordingly, items that assesses social presence should all tap this realness aspect. Furthermore, the wording of the items should be aligned with the Rasch Measurement Model [5, 6, 25, 40]. This latter aspect is important as the Rasch Measurement Model requires items that vary in their degree to be endorsed by respondents. In particular: the Rasch Measurement Model requires items that are easy, moderate, and hard to endorse by respondents so to differentiate respondents who have low, average, and high perceptions of the other in terms of realness [8; Chapter 4]. A second requirement of the Rasch Measurement Model is the requirement of the uni-dimensionality of a measure. Uni-dimensionality means that *the same items* should not assess other constructs at the same time as the target construct<sup>4</sup> [6, 8]. Thus, when items tap the realness aspect, then the same items

---

<sup>4</sup> Note that the requirement of uni-dimensionality does not mean that a construct cannot have more than one dimension. If a construct has more than one dimension, then *the different items* should assess all the sub-constructs underlying these dimensions; that is, one set of items will assess the first sub-construct, another set the second sub-construct and so on. However, for each set of items the requirement of uni-dimensionality would apply.

should not also tap other aspects such as whether the other in the communication is perceived as friendly or that the medium is useful for interpersonal communication.

With respect to the items of our raw social presence measure, we looked whether items of other social presence measures could be included in our social presence measure as long as they would fulfil the two above requirements (i.e., fit in a certain difficulty category and tap realness). However, none of the existing items did fulfill both these requirements. There were two reasons. First, almost all social presence definitions did not acknowledge physical realness as being the defining element of social presence. As explained in Kreijns et al. [16] and in Lowenthal and Snelson [23], depending on the definition of social presence, measures were constructed that operationalized these definitions. Thus, if social presence was seen as quality of the social climate, items of its associated measure will tap social climate. For example, Rourke and Anderson [28] used six, 5-point bipolar scale items that assessed the degree to which the social climate was perceived as trusting, warm, friendly, disinhibiting, close, and personal. However, if in contrast social presence was seen as an expression of immediacy, then items will tap immediacy behaviors. According to Short et al. [29; p. 72] (see also [35, 37]), immediacy is “a measure of the psychological distance which a communicator puts between himself and the object of his communication, his addressee or his communication.” Gunawardena and Zittle [11] developed 14, 5-point Likert scale items from which they contended tapped immediacy behaviors. Social presence can also be seen as an expression of intimacy which is according to Short et al. [29] (see also [2, 35]) an equilibrium theory postulating that communicators will reach an optimal level of ‘intimacy’ in which conflicting approaches and avoidance forces are in equilibrium. If so, items will tap intimacy behaviors. Gunawardena and Zittle [11] contended that Gunawardena’s [10] social presence measure consisting out of 15, 5-point bipolar scale items is tapping intimacy behaviors.

Second, our social presence definition addressed only the realness aspect; hence, it has only one dimension. In that respect, we followed Short et al. [29] who also regarded social presence as a single dimension. We realized that we deviated from many other social presence researchers who argued social presence to be a multi-dimensional construct. For example, Tu [33, 34] saw as dimensions of social presence (1) social context, (2) online communication, (3) interactivity, (4) system privacy, and (5) feelings of privacy. However, he actually identified these dimensions as variables affecting the degree of social presence. Wei et al. [35] saw as dimensions of social presence, (1) co-presence, (2) intimacy, and (3) immediacy. These two latter dimensions were also put forward by Short et al. [29] as we have seen above, but they saw social presence to be a factor contributing the level of intimacy and enabling immediacy.

As result, a 16-item raw social presence measure was constructed where all items were newly formulated. All items used a Likert scale with seven rating scale steps (1 = totally disagree, 2 = disagree, 3 = somewhat disagree, 4 neither disagree or agree, 4 = somewhat agree, 5 = agree, 6 = totally agree). Table 1 depicts the (raw) social presence measure. In this table, items that were found not fitting the Rasch Measurement Model are greyed, hence they are not part of our social presence measure. The mean  $M$  and standard deviation  $SD$  of each item were calculated by excluding (1) respondents who did not answer the item (intrinsic missing value), (2) respondents who had a misfitting answer on the item given their overall answer pattern on the rest of the



**Table 1.** Social presence measure

Nr Item	Item	7 rating scale steps			5 rating scale steps			N*
		<i>M</i>	<i>SD</i>	item measure	<i>M</i>	<i>SD</i>	item measure	
	In this learning environment ...							
SP01	... it feels as if we are a face to face group	2.93	1.47	39.30	2.22	.86	31.59	304
SP02	... it feels as if I cannot escape from the eyes of my fellow students	2.08	1.18		1.75	.70		299
SP03	... it feels as if I deal with 'real' persons and not with abstract anonymous persons	4.12	1.70	33.37	2.91	1.00	26.38	304
SP04	... I feel to be together with my fellow group students	3.22	1.51		2.38	.85		304
SP05	... I can form distinct impressions of some of my fellow students	3.66	1.51	35.90	2.59	.86	28.99	298
SP06	... it feels as if all my fellow students are 'real' physical persons	4.42	1.69	31.76	3.07	1.03	25.14	302
SP07	... I imagine that I really can 'see' my fellow students to be in front of me	2.98	1.56	39.04	2.23	.91	31.56	304
SP08	... my fellow students feel so 'real' that I almost believe that we are not virtual at all	3.01	1.62	38.91	2.24	.95	31.46	304
SP09	... I have my fellow group members in my minds' eye	3.09	1.70		2.30	.99		302
SP10	... all of my fellow students imagine that they really can 'see' me to be in front of them	2.82	1.46		2.17	.86		301
SP11	... all of my fellow students feel that I am a 'real' physical person	3.60	1.70	35.94	2.62	1.00	28.54	300
SP12	... it feels as if all my fellow students and I are in the same room	2.77	1.55	40.11	2.11	.89	32.49	304
SP13	... it feels as if I can really 'touch' my fellow students	1.86	1.12		1.61	.71		304
SP14	... it feels as if all my fellow students and I are in close proximity	2.45	1.48	41.80	1.96	.88	33.64	304
SP15	... I strongly feel the presence of my fellow students	3.09	1.68	38.54	2.29	.95	31.16	301
SP16	... all of my fellow students feel my presence	2.78	1.53		2.15	.90		302

\* N is calculated by taking the total number of respondents (=324) and subtracting 1) respondents who did not answer this item, 2) respondents who had a misfitting answer on this item—the misfitting answer was marked as missing, and 3) respondents who completely misfit the Rasch Measurement Model (Nmisfit = 20).

items—the misfitting answer was marked as missing (extrinsic missing value), and (3) respondents who completely misfit the Rasch Measurement Model (Nmisfit = 20). The mean *M* and standard deviation *SD* were also calculated when not seven rating scale steps were used but instead five rating scale steps. The latter was constructed by collapsing the steps 2 and 3, and also the steps 4 and 5. Collapsing turned out to be necessary in order to have proper probability distributions for the rating scale steps; the section Analysis will give all the details regarding misfit respondents and collapsing rating scale steps. That section will also explain the item measures<sup>5</sup> shown in the table; the items measures are only given for the remaining 10 items that fitted the Rasch Measurement Model.

<sup>5</sup> In Winsteps, the score assigned to a person (i.e., the respondent) is referred to as 'measure.' This is, thus, not to be confused with the meaning of measure as instrument to measure some trait or phenomenon such as social presence.

### 3 Method

#### 3.1 Respondents

Respondents were 324 students at the largest distance university in Germany, FernUniversität Hagen. This convenience sample consists of students enrolled in either B.Sc. Psychology or B.A. Educational Science. Table 2 shows the demographics for this sample. There was a total of 241 students enrolled in Educational Science, 71 in Psychology, spread over three semesters: winter semester of 2015/2016, winter semester of 2016/2017, and summer semester of 2017. Of these students, 260 were female, 55 were male. Mean age was 32.3 years. Note that due to missing values, numbers may not add up to total N.

**Table 2.** Demographics of sample.

	N	Educational science, B.A.	Psychology B.Sc.	Female	Male	M <sub>age</sub>
WS 15/16	134	99	34	100	26	32.2
WS 16/17	112	92	19	99	12	33
SS 17	78	50	18	61	17	31.8
Total	324	241	71	260	55	32.3

#### 3.2 Procedure

Students were recruited for the survey through the learning management system Moodle, in which most learning activities took place. They were asked to participate in the survey with no course credit or reward attached to participation. A link in the learning environment directed them to the survey, which was created via LimeSurvey<sup>6</sup>. The 16 items of the raw social presence were only a small part of a larger survey concerned with student's perceptions and experiences in the learning environment. The survey took them a total of about 15 min to complete. It was administered over the course of three semesters, mentioned earlier.

#### 3.3 Analysis

The raw social presence measure (see Table 1) contained 16 items that tapped the realness of the other in the communication. All items used Likert scales with seven rating scale steps for getting an item score. Winsteps version 3.90 was used as analyzing tool as it implements the Rasch Measurement Model [19]. With the Rasch Measurement Model [25, 40] scale validation can be conducted and—at the same time—item and person measures determined as the Rasch Measurement Model allows the separation of these measures. With scale validation is meant the verification whether the set of items assesses the same underlying latent 'trait' of social presence, namely realness of the other. As such, scale validation is also testing the uni-dimensionality of social presence.

<sup>6</sup> See <http://www.limesurvey.org>.

Item and person measures represent the more ‘true’ scores; in particular, person measures contrast total scores that are commonly used by many researchers [6, 7, 32]. Total scores are the summation of the items scores, but as explained by Boone [7], and already mentioned above in this paper, items scores are nonlinear because Likert scales are ordinal rather than interval measures. Therefore, total scores also may not be assumed to be linear [5, 6]. In contrast, item and person measures are linear and both are expressed on the same interval scale and denoted in *logits*<sup>7</sup> that can be either negative or positive. Measures are ‘better’ when going from the most negative measure to the most positive measure [6; Chapter 4]. However, we applied a linear transformation to these measures to obtain only positive measures exhibiting the same range as would total scores.

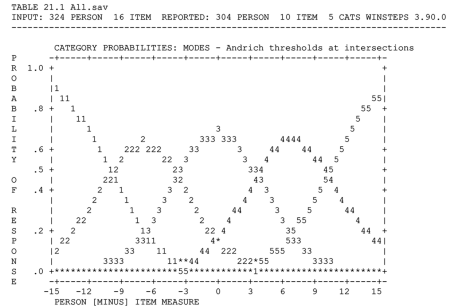
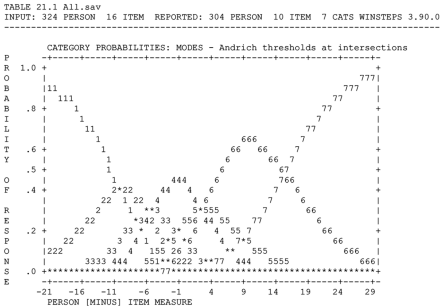
Conducting the Rasch analyses was an iterative process requiring many steps in which misfitting items and persons were identified, probability curves of the rating scales inspected, and item and person separation indices calculated as better alternatives for Cronbach’s alpha [17]. Item and person misfit means that these items and persons do not contribute to the construction of a valid measurement instrument. Item misfit can be detected when the index Outfit Mean Square (MNSQ) is above the value of 1.5 or below the value of .5 [39]. Person misfit follows the same criterion as item misfit; that is, if MNSQ is above the value of 1.5 or below the value of .5 then the person is misfitting. However, following Boone [6; p. 173], it was decided to use the Outfit Z-standardized (ZSTD) which absolute value must not exceed the value of 3.0 if the item is not to be considered as a misfit [6; p. 173].

The first step was detecting respondents who had misfitting answers to some of the items given the overall pattern of answers to the other items. These respondents may fit the Rasch Measurement Model when these misfitting are ‘repaired;’ that is, they are marked as missing. Misfitting answers were identified by inspecting the Z-residuals of each person answer on the all items [6; p. 177]. A total of 17 persons were ‘repaired;’ that is, their misfitting answers were marked as missing. The second step was detecting misfitting items by inspecting the MNSQ values. Four items were found to misfit, these items were item SP02, SP10, SP13, and SP16 and were, therefore, not included in the next iteration step. Because we were aware that each iteration step reshifts item and person measures, we inspected the changed MNSQ values again in the next iteration and found two more misfitting items. These two items were item SP04 and SP09 and they were excluded for further analyses. The third step was detecting misfitting persons (i.e., respondents). Two iteration steps revealed 20 misfitting persons and they were excluded for further analyses. The number of respondents eligible for analyses was reduced to 304. The fourth step considered whether (1) the observed ordering of the seven rating scale steps matched the theoretical ordering (1 = totally disagree, 2 = disagree, 3 = somewhat disagree, 4 neither disagree or agree, 4 = somewhat agree, 5 = agree, 6 = totally agree) and (2) whether all seven rating steps were used [6, Chapter 9]. The analyses revealed no problems with these two issues but in the fifth analysis step we saw that the probability curves were less than ideal, especially the rating steps 3 and 5 had a lower probability than ideally should be suggesting to collapse the rating steps 2 and 3 as well as to collapse the rating steps 4 and 5. The left

---

<sup>7</sup> The *logit* is the unit in which the person and item measures are expressed [5, 6].

and right graphs in Fig. 1 clearly show the difference in rating step probabilities: the left graph shows the probabilities of each of the seven rating scale steps; the right graph of the five rating scale steps. It was decided to continue the analyses with the collapsed rating scale steps.

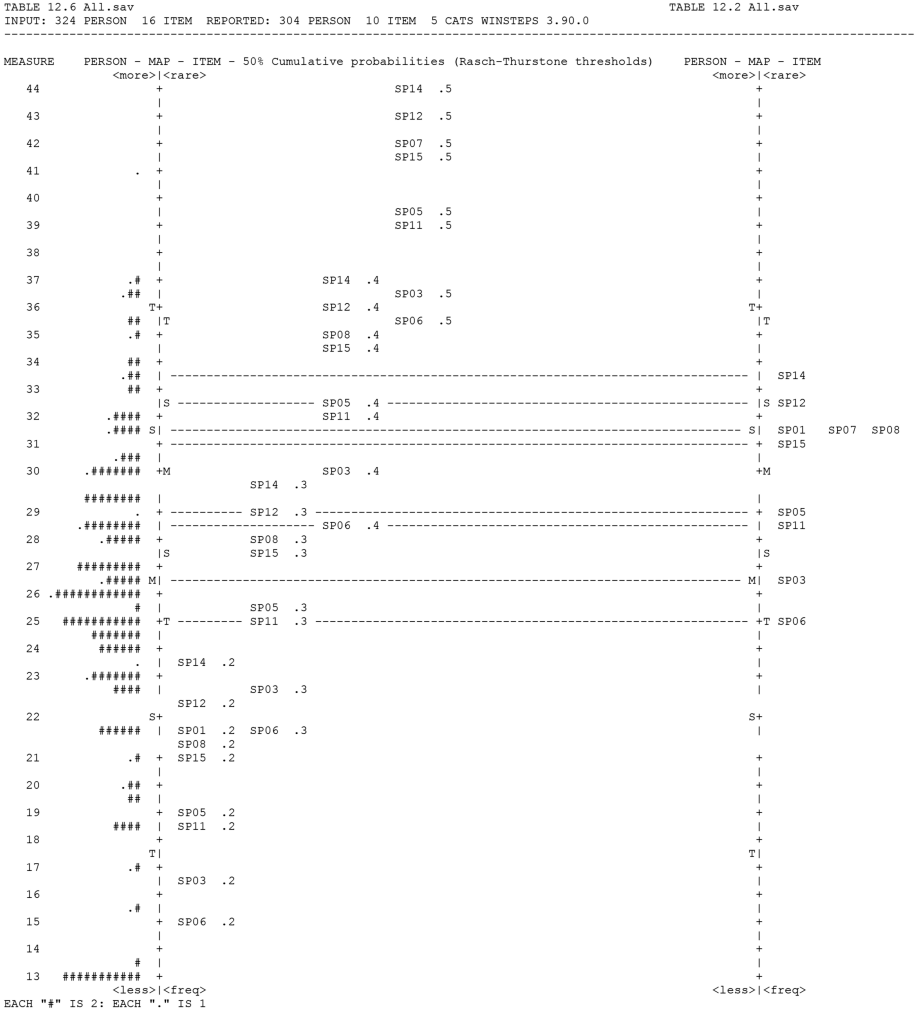


**Fig. 1.** The left graph shows the probability of the seven rating scale steps. The right graph shows the probability of the five rating scale steps.

### 4 Results

All the above analyses produced the two Wright maps [6; Chapter 6] in Fig. 2. The left Wright map shows the distribution of the item rating scale step numbers (right of the left axis) and, at the same time, the distribution of the person measures (left of the vertical left axis). The position of the item rating scale step number on the vertical axis indicates that the probability of a person, whose measure is at the same position on the vertical axis, is 50% to be in that rating scale step or those above, and 50% to be in the rating scale step which number is one less or those below. Thus, when the item rating scale step number, for example, is 2, it refers to the 50% threshold between the rating scale steps 1 and 2. The Wright map only shows the rating scale step numbers 2, 3, 4, and 5, which stands for the 50% thresholds between the rating scale steps 1 and 2, 2 and 3, 3 and 4, and 4 and 5 respectively. There were 22 respondents who answered ‘extreme;’ that is, these respondents answered all 10 items as ‘totally disagree’ (=1). The 22 minimum extreme persons are depicted at the lowest position on the vertical axis. The right Wright map shows the distribution of the item measures along the right vertical axis. The item measure is the position on the vertical axis at which the probability of a person, whose measure is at the same position on the vertical axis, is 50% to be in the higher categories and 50% to be in the lower categories.

The two Wright maps show four psychometric properties of the social presence measure. First, the mean person measure (including the minimum extreme persons) was 25.51 and the mean item measure was 30.09, a difference of 4.58. Because the mean person measure is less than the mean item measure, it means that the social presence items were a bit difficult to endorse by the respondents. In other words, respondents had difficulties perceiving the realness of the others, in particular those 22 minimum extreme respondents. Ideally, the mean item measure should be about 1 logit



**Fig. 2.** Left: Wright map showing the distribution of the item rating scale step numbers (right of the left axis) and at the same time showing the distribution of the person measures (left of the left vertical axis). Note that sometimes an item rating scale step number of SP01, SP07, and SP08 is not shown because their positions on the vertical axis were very close to each other, and therefore could not be printed all together on the same line but one is. Right: Wright map showing the distribution of the item measures (right of the right axis).

lower than the mean person measure [18; p. 27]; it is now >1 logit higher than the mean item measure.

Second, in the left Wright map, the item category numbers (after collapsing) are all in an ascending order (i.e., category 2, at the bottom, followed by category 3 and then category 4, and category 5 at the top), which positively adds to the construct validity of the measure [3].

Third, the left Wright map shows that at the lower end of the person measure distribution along the vertical axis is not covered by even the lowest item rating scale step (i.e., item rating scale step number 2). This is clearer seen in the right Wright map: there is no item whose measure is lower than 25 whereas 113 persons have measures lower than 25 (including minimum extreme persons). Consequently, the current social presence measure is moderate in differentiating respondents with low perceptions of the realness of the other whereas it can excellently differentiate persons with high perceptions of the realness of the other. This indicates that there is some underrepresentation of the construct [24] which would undermine the statistical validity (i.e., the reliability) of the social presence measure [3]. Nevertheless, item and person separation indices were very good [6; p. 231]; item separation index was 10.01 (should be at least 2.5 for the analysis of groups) and person separation index was 3.01 in case minimum extreme persons were included (should be at least 3.0 to represent an excellent level of separation). Classical test theory Cronbach's alpha was .93.

Fourth, the item measures of SP01, SP07, and SP08 are almost of the same difficulty level, the measures were 31.59, 31.56, and 31.46 respectively. This suggests that two of these three items are redundant and could be removed from the 10-items social presence measure.

## 5 Discussion and Conclusions

In this paper we shortly outlined the confounding issues surrounding the concept of social presence and which may cause future research at risk as none of the findings can be compared due to different reformulations and interpretations of it and the different measurement instruments to assess the degree of social presence. As for the different measures, they have issues pertaining to the (1) invariance of the measure across setting and situations, (2) sole use of principal component analysis for construct validation, (3) underrepresentation of the social presence construct, (4) items that may tap other latent 'traits' rather than the realness of the other alone, and (5) uni-dimensionality of the social presence measure.

We, therefore, were motivated to start from the original definition of social presence given by Short et al. [29] so to develop a social presence measure that assesses the realness of the other in the communication. It is hoped for that this measure will become a standard. Starting with a raw social presence measure containing 16 items using ordered categorical polytomous rating scales with seven rating scale steps, the Rasch Measurement Model was used to assess the psychometric properties of the measure. Our findings were that the resulting 10-items social presence measure has good psychometric properties but there are a number of issues that have to be taken care of in our future research to improve the 10-items social presence measure: First, overall the social presence measure was a bit difficult to endorse by all respondents (i.e., the mean of the person measures was lower than the mean of the item measures). Second, rather than seven rating scale steps, five rating scale steps showed better probability distributions for each step. Third, measuring the degree of realness was excellent for those who have high perceptions of realness of the other (i.e., they could be well differentiated), whereas this was moderate for those who have low perceptions

(i.e., they could be less well differentiated). Finally, three items were found to have almost the same difficulty (i.e., had almost the same item measure) and, therefore, two of them could be removed.

However, our study has also its limitations that may affect the usefulness of the social presence measure. First, we did administer the survey with the raw social presence measure only to students in collaborative learning settings that use a-synchronous communication media and not to students that use synchronous media. Consequently, we cannot say the social presence measure is invariant with respect to a-synchronous and synchronous media. Second, we did not differentiate between men and women, and between the study the students were enrolled in (i.e., B.Sc. Psychology and B.A. Educational Science) when performing the Rasch analysis. Consequently, we also cannot say that the social presence measure is invariant for men and women, or for the study the students were enrolled in. Third, all Rasch analyses were performed on one sample. Therefore, we cannot say whether the instrument is invariant across samples (see [8; p. 38–40]).

Taking all the issues together, we consider the current social presence measure as a preliminary social presence measure that surely needs further improvement. That is, we first to have to include new (easy) items so that those who have low perceptions of realness could equally well be differentiated as those with high perceptions of it. Second, we will administer the survey to students in collaborative learning settings that use either use a-synchronous or synchronous communication media. Third, we will perform different item functioning (DIF) analyses [6; Chapter 13] when performing the Rasch analyses so to study the issue of invariance more closely with respect to the use of a-synchronous communication media versus synchronous media, gender, study enrollment, and potentially other factors that may influence DIF. Fourth, more samples will be used to test sample invariance.

Nevertheless, our conclusion is that the preliminary 10-items social presence measure is already an improvement when compared to existing social presence measures that emphasize realness. This preliminary social presence measure already can be used to assess effects of social presence in computer mediated communication such as found in online collaborative learning settings (e.g., providing evidence that social presence has effect on learning) as long as its current limitations are taken into account.

## References

1. Abdullah, M.H.: Social presence in online conferences: What makes people ‘real’? *Malays. J. Dist. Educ.* **6**(2), 1–22 (2004)
2. Argyle, M., Dean, J.: Eye contact, distance and affiliation. *Sociometry* **28**, 289–304 (1965)
3. Baghaei, P.: The Rasch model as a construct validation tool. *Rasch Measur. Trans.* **22**(1), 1145–1146 (2008)
4. Biocca, F., Harms, C., Burgoon, J.K.: Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence: Teleoperators Virtual Environ.* **12**(5), 456–480 (2003)
5. Bond, T., Fox, C.M.: *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd edn. Routledge, New York, London (2015)

6. Boone, W.J., Staver, J.S., Yale, M.S.: *Rasch Analysis in the Human Sciences*. Springer, Dordrecht, The Netherlands (2014)
7. Boone, W.J.: Rasch analysis for instrument development: Why, when, and how? *CBE-Life Sci. Educ.* **15**(4), rm4 (2016)
8. Engelhard Jr., G.: *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*. Routledge, New York, London (2013)
9. Garrison, D.R.: Communities of inquiry in online learning. In: Rogers, P.L. (ed.) *Encyclopedia of distance learning*, 2nd edn, pp. 352–355. IGI Global, Hershey, PA (2009)
10. Gunawardena, C.N.: Social presence theory and implications for interaction and collaborative learning in computer conferences. *Int. J. Educ. Telecommun.* **1**(2&3), 147–166 (1995)
11. Gunawardena, C.N., Zittle, F.: Social presence as a predictor of satisfaction within a computer mediated conferencing environment. *Am. J. Dist. Educ.* **11**(3), 8–25 (1997)
12. Hollis, H.: The impact of social media on social presence and student satisfaction in nursing education. Unpublished dissertation. University of Alabama, Tuscaloosa, AL (2014)
13. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. *Bus. Horiz.* **54**, 59–68 (2010)
14. Kim, J.I., Ha, T., Woo, W., Shi, C.-K.: Enhancing Social Presence in Augmented Reality-Based Telecommunication System. In: Shumaker, R. (ed.) *VAMR 2013, Part I. LNCS*, vol. 8021, pp. 359–367. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39405-8\\_40](https://doi.org/10.1007/978-3-642-39405-8_40)
15. Kreijns, K., Kirschner, P.A., Jochems, W., Van Buuren, H.: Measuring perceived social presence in distributed learning groups. *Educ. Inf. Technol.* **16**(4), 365–381 (2011)
16. Kreijns, K., Van Acker, F., Vermeulen, M., van Buuren, H.: Community of inquiry: Social presence revisited [Special Issue: Inquiry into “Communities of Inquiry:” Knowledge, Communication, Presence, Community]. *E-Learn. Digit. Media* **11**(1), 5–18 (2014)
17. Linacre, J.M.: KR-20/Cronbach alpha or Rasch person reliability: which tells the “truth”? *Rasch Measur. Trans.* **11**(3), 580–581 (1997)
18. Linacre, J.M.: Computer adaptive testing: a methodology whose time has come. In: Chae, S., Kang, U., Jeon, E., Linacre, J.M. (eds.) *Development of Computerized Middle School Achievement Test*. Komesa Press, Seoul, South Korea (2000)
19. Linacre, J.M.: *Winsteps® Rasch measurement computer program user’s guide*. Winsteps.com, Beaverton, OR (2016)
20. Liu, S.Y., Gomez, J., Yen, C.-J.: Community college online course retention and final grade: predictability of social presence. *J. Interact. Online Learn.* **8**(2), 165–182 (2009)
21. Lombart, M., Ditton, T.: At the heart of it all: The concept of presence. *J. Comput. Med. Commun.* **3**(2), (1997). <https://academic.oup.com/jcmc/article/3/2/JCMC321/4080403>. Accessed 16 June 2018
22. Lowenthal, P.R.: The evolution and influence of social presence theory on online learning. In: Kidd, T.T. (ed.) *Online Education and Adult Learning: New Frontiers for Teaching Practices*, pp. 124–134. IGI Global, Hershey, PA (2010)
23. Lowenthal, P.R., Snelson, C.: In search of a better understanding of social presence: an investigation into how researchers define social presence. *Dist. Educ.* **38**(2), 1–19 (2017)
24. Messick, S.: Validity and washback in language testing. *Lang. Test.* **13**(3), 241–256 (1996)
25. Rasch, G.: *Probabilistic Models for Some Intelligence and Attainment Tests*. Paedagogiske Institut, Copenhagen (1960)
26. Rosakranse, C., Nass, C., Oh, S.: Social presence in CMC and VR. In: Burgoon, J., Magnenat-Thalmann, N., Pantic, M., Vinciarelli, A. (eds.) *Social Signal Processing*, pp. 110–120. Cambridge University Press, Cambridge (2017)



27. Richardson, J.C., Maeda, Y., Lv, J., Caskurlu, S.: Social presence in relation to students' satisfaction and learning in the online environment: a meta-analysis. *Comput. Hum. Behav.* **71**, 402–417 (2017)
28. Rourke, L., Anderson, T.: Exploring social interaction in computer conferencing. *J. Interact. Learn. Res.* **13**(3), 257–273 (2002)
29. Short, J., Williams, E., Christie, B.: *The Social Psychology of Telecommunications*. Wiley, London (1976)
30. Sick, J.: Rasch measurement and factor analysis. *SHIKEN: JALT Test. Eval. SIG Newsllett.* **15**(1), 15–17 (2011)
31. Swan, K., Matthews, D., Bogle, D., Boles, E., Day, S.: Linking online course design and implementation to learning outcomes: a design experiment. *Internet High. Educ.* **15**(2), 81–88 (2012)
32. Tennant, A., Conaghan, P.G.: The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied and what should one look for in a Rasch paper? *Arthritis Rheum.* **57**(8), 1358–1362 (2007)
33. Tu, C.H.: The measurement of social presence in an online learning environment. *Int. J. Educ. Telecommun.* **1**(2), 34–45 (2002)
34. Tu, C.H.: The relationship between social presence and online privacy. *Internet High. Educ.* **5**(2002), 293–318 (2002)
35. Wei, C.-W., Chen, N.-S., Kinshuk, : A model for social presence in online classrooms. *Educ. Technol. Res. Develop.* **60**(3), 529–545 (2012)
36. Weidlich, J., Bastiaens, T.: Explaining social presence and the quality of online learning with the SIPS model. *Comput. Hum. Behav.* **72**, 479–487 (2017)
37. Wiener, M., Mehrabian, A.: *Language Within Language: Immediacy, A Channel in Verbal Communication*. Apple-Century-Crofts, New York (1968)
38. Wright, B.D.: Comparing Rasch measurement and factor analysis. *Struct. Eqn. Model.* **3**(1), 3–24 (1996)
39. Wright, B.D., Linacre, J.M.: Reasonable mean-square fit values. *Rasch Measur. Trans.* **8**, 370–371 (1994)
40. Wright, B.D., Masters, G.N.: *Rating Scale Analysis*. MESA Press, Chicago, IL (1982)



# Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences

Jan Schneider<sup>1</sup>(✉), Daniele Di Mitri<sup>2</sup>, Bibeg Limbu<sup>2</sup>,  
and Hendrik Drachslers<sup>1</sup>

<sup>1</sup> DIPF, Frankfurt am Main, Germany

{schneider.jan, drachslers}@dipf.de

<sup>2</sup> Welten Institute, Open University of the Netherlands,  
Heerlen, The Netherlands

{daniele.dimitri, bibeg.limbu}@ou.nl

**Abstract.** Studies in Learning Analytics provide concrete examples of how the analysis of direct interactions with learning management systems can be used to optimize and understand the learning process. Learning, however, does not necessarily only occur when the learner is directly interacting with such systems. With the use of sensors, it is possible to collect data from learners and their environment ubiquitously, therefore expanding the use cases of Learning Analytics. For this reason, we developed the Multimodal Learning Hub (MLH), a system designed to enhance learning in ubiquitous learning scenarios, by collecting and integrating multimodal data from customizable configurations of ubiquitous data providers. In this paper, we describe the MLH and report on the results of tests where we explored its reliability to integrate multimodal data.

**Keywords:** Multimodal Learning Analytics · Sensor-based learning  
System design

## 1 Introduction

Imagine it is a Sunday morning and you are walking in the countryside. You can see different types of trees, herbs, and bushes all over the place. You can hear some birds singing, feel a cool breeze gently brushing your cheeks, and perceive the particular fragrance of wet grass. Each modality perceived through your senses helps to make the experience of walking through the countryside more comprehensive and meaningful. Imagine you are in the countryside again and suddenly you hear a thunderous sound that alerts you. In this case, your auditory sense through the modality of sound captured crucial information that was missed by other senses that make use of other modalities.

Learning Analytics (LA) is the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [1]. The LA approach for gathering data is usually unimodal, as its main focus is the analysis of log-files containing student's interaction with technology-mediated learning environments such as:

Learning Management Systems (LMS) [2], Intelligent Tutoring Systems (ITS) [3], Massive Open Online Courses (MOOC) [4], or other types of systems that use a computer as an active component in the learning process. Analyzing the unimodal log-files of students in order to understand the learning process is analogous to experiencing a walk through the countryside using only one of the senses. In both cases, the captured experience is limited and might exclude some crucial information.

The increasing popularity of sensors [5] has driven the development of smart technologies such as the Internet of Things (IoT) and wearable devices, which are able to measure and record different physical properties. A set of such devices can be used to collect multiple physical properties or modalities from a phenomenon, hence generating multimodal data. Bridging the use of multimodal data with learning theories is a goal of the field of study called Multimodal Learning Analytics (MMLA) [6]. MMLA can offer new insights into learning as it can study the learning process in scenarios that are not entirely restricted by the direct mouse and keyboard interaction with a computer [7]. MMLA applications have already shown their potential to support a vast number of learning activities [8]. Examples of these learning activities include 21st-century skills such as public speaking [9–11], job interviews [12], negotiation scenarios [13], and collaboration [6, 14].

MMLA as a field of research presents multiple challenges that range from the collection of raw data to the exploitation of analyzed data in order to support learning. In terms of the collection and integration of multimodal data, current developers of MMLA applications are required to implement tailored solutions from scratch. This process of implementing tailored applications from scratch is expensive and hinders the emergence of common methods, specifications, and standards for the analysis and exploitation of multimodal data for learning. To address this problem, we developed the Multimodal Learning Hub (MLH). The MLH is an application that handles the collection and integration of data from multiple sources, supporting, in turn, the use of customizable configurations for capturing learners' behavior and/or environment with the use of multimodal data. This capture and integration of multimodal data for learning is referred to in this paper as a Multimodal Learning Experience. In this paper, we describe the MLH and report on preliminary results regarding its capacity to record Multimodal Learning Experiences.

## 2 Multimodal Learning Hub

The main Task of the MLH<sup>1</sup> is to deal with the collection and integration of multimodal data from customizable data provider configurations with the purpose to generate Multimodal Learning Experiences out of Meaningful Learning Tasks (See Sect. 2.2).

---

<sup>1</sup> <https://github.com/janschneiderou/LearningHub>.

## 2.1 Multimodal Data for Learning

As the name implies, multimodal data for learning is the data that comes from multiple sources with the purpose to support the learning process. As an example, consider the scenario of an application designed to support the development of public speaking skills. For such an application, it is possible to use a depth camera and a microphone to track specific aspects of the communication of a learner. The depth camera and the microphone produce a different type of data and at different rates. A depth camera such as a Microsoft Kinect V2 is able to retrieve the relative coordinates of the learner's joints at an average rate of 25 frames per second. On the other hand, a microphone used to record music typically retrieves 44100 of volume values every second. Both devices produce completely different streams of data values. Integrating and making sense of these streams of data in order to support learning is not a straightforward task. One function of the MLH is to collect data from different data source providers and create a unified multimodal experience.

Another characteristic of multimodal data lies in the difficulty for it to be interpreted, as it is generally noisy and has low semantic value [15]. For humans, the interpretation of raw data (numerical digital values) streams is a very difficult task. Adding multimodality to these streams makes the task of interpretation even harder. In the case of video and audio data streams, it is possible to display them as video and audio respectively. This makes the tasks of interpretation simple for us since we have evolved to make sense out of this type of input streams in order to interact with the environment. Nonetheless, displaying other types of data streams such as relative coordinates, acceleration, heart rate, skin conductance, temperature, pressure, tension, etc. in a way that can be easily interpreted by humans is a challenge. In order to support learning, either by humans or machines, multimodal data needs to be interpreted. The MLH assists this interpretation in two ways. It creates multimodal recordings of meaningful learning tasks and forwards critical sensor data to immediate feedback applications.

The work in [16] proposes a method to interpret multimodal recordings of meaningful learning tasks. It argues that human experts can manually label relevant aspects of a multimodal recording by looking at the video portion of it. The labeled multimodal data can be used to learn statistical models, which in turn can generate predictions (interpretations) of multimodal data.

The introduction of this article provides the example of walking in the field and becoming alerted by a thunderous sound. Critical sensor data can be used similarly to alert the learner. For example in the case of a multimodal application designed to support the development of public speaking skills, if the application identifies that the learner is speaking too soft, it immediately can alert the learner about it. The MLH can receive data that has been identified as critical and then forward this data to generic feedback applications.

## 2.2 Meaningful Learning Task

Multimodal data for learning is only useful in the context of the corresponding learning task. Speaking too soft in a presentation is different than speaking too soft during a

collaborative project performed in the university library. When using a multimodal learning application, it is important to define the Meaningful Learning Tasks that learners will perform while using it. The definition of a Meaningful Learning Task is context dependent. A dancing choreography is composed of dancing figures, and dancing figures are composed of steps. Meaningful Learning Tasks in a dancing lecture might be: practicing a step, practicing a figure, or practicing choreography. Ideally, once the learner masters the practice of a step, she can then move on to the practice of a figure, and finally to the choreography. The types of learning interventions, such as feedback, given to a learner for each meaningful learning task are different. Therefore, prior to the use of any multimodal application for learning, including the ones that can be built using the MLH, it is important to clearly define the Meaningful Learning Task that learners will perform. A clear definition of a Meaningful Learning Task will facilitate the process of manually labeling multimodal data, the acquisition of more accurate interpretations of the multimodal data, and the provision of relevant feedback to the learner. In other words, the definition of a Meaningful Learning Task provides the context for the analysis and exploitation of multimodal data for learning.

### 2.3 System Description

The MLH is built as a .NETFramework V4.6 desktop application, which allows for high-level and low-level programming. The design of the architecture and operational mode of the MLH is based on the results of a series of test where we investigated how to reliably integrate multimodal data (See Sect. 3). A sketch of the MLH architecture is displayed in Fig. 1.

It is neither feasible nor desirable to have applications retrieving, recording and analyzing data from learners all the time. As explained in Sect. 2.2 multimodal data for learning is valuable in the context of a Meaningful Learning Task. Therefore, in its current state, in order to create valuable multimodal recordings, the user of the MLH needs to manually start and stop a recording. The data used for the recordings are retrieved by Data Provider Applications. These applications can run locally or in different computers connected to the network. Example of this type of applications can be:

- Applications controlling specific sensor devices such as depth cameras, accelerometers, physiological sensors, or any type of sensor that can be connected to a computer system.
- Applications controlling video or audio recordings.

In order to create recordings of Multimodal Learning Experiences, the user first needs to configure the setup that will be used to capture the learner's behavior and/or environment. This setup configuration includes the selection of Data Provider Applications that will be used for the recording and the configuration of the communication channels that will be used between the Data Provider Applications and the MLH. The definitions of these communication channels include the path or address of the application and a set of port numbers (see Fig. 2).

Data Provider Applications should use the same channels that were defined by the user. To make the definition of the communication channels simpler in the side of the Data Provider Applications, we propose the use of dynamic libraries that can handle all

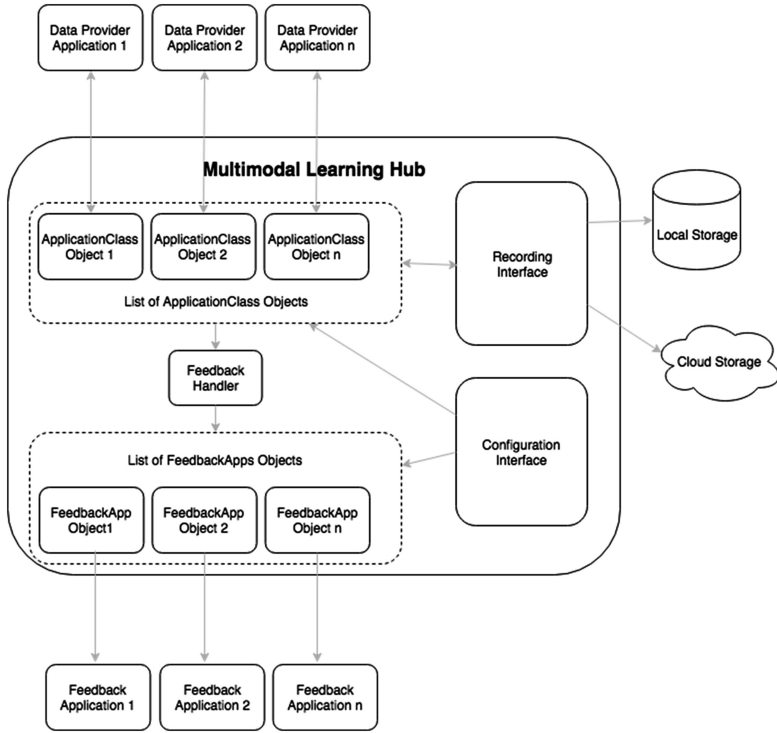


Fig. 1. Sketch of the MLH architecture

Name	Path	Remote	TCPListener	TCPSender	TCPFile	UDPListener	UDPSender	Used
MyoTest	C:\Users\FW\id\source\repos\MyoStandalone\MyoStandalone\MyoTest\MyoTest\bin\Debug\MyoTest.exe	<input type="checkbox"/>	11004	12004	13004	14004	15004	<input checked="" type="checkbox"/>
PresentationTrainer	192.168.171.127	<input checked="" type="checkbox"/>	11001	12001	13001	14001	15001	<input checked="" type="checkbox"/>
LeapExample	C:\Users\FW\id\source\repos\LeapExample\LeapExample\bin\Debug\LeapExample.exe	<input type="checkbox"/>	11005	12005	13005	14005	15005	<input checked="" type="checkbox"/>

Fig. 2. The configuration of the Data Providers

the communication between these applications and the MLH, including the automatic discovery of the defined communication channels. Currently, the MLH solution includes a dynamic library for .Net projects and for Windows Universal Platform.

Data Storing: As discussed previously each Data Provider Application retrieves a different type of data at a different rate. In order to fuse the data coming from different providers in one unified multimodal recording we used the following recording format:

A multimodal recording is composed of a collection of RecordingObjects. Each provider generates one RecordingObject. A RecordingObject is composed by a recordingId, an applicationName (name of the Data Provider Application) and a collection of FrameObjects. Each FrameObject consist of a frameStamp (relative timestamp since the beginning of the recording) and a dictionary containing the name of the

attributes stored for each frame and the current values of these attributes. Figure 3 displays an example of a `RecordingObject` already stored in a JSON<sup>2</sup> format. Once the recording stops all `RecordingObjects` are collected by the MLH using batch integration approach (See Sect. 3.1).

```

1 {
2   "recordingID": "14H2M15S",
3   "applicationName": "MyoTest",
4   "frames": [{
5     "frameStamp": "00:00:00.0030000",
6     "frameAttributes": {
7       "GripPressure": "2",
8       "OrientationW": "0.7064209",
9       "OrientationX": "-0.4743652",
10      "OrientationY": "0.5076904",
11      "OrientationZ": "0.1347656"
12    }
13  }]
14 }
```

**Fig. 3.** Example of a JSON string containing a one frame long *RecordingObject* for a MYO Band application

Multimodal Data Synchronization: Sharing a common data format among the different Data Provider Applications is a step required in order to fuse these streams of data in one Multimodal Learning Experience. A second step required is to share a common time reference among the Data Provider Applications. The MLH achieves this shared time reference among the Data Provider Applications by sending them a *StartRecording* instruction whenever a recording starts. The Dynamic Libraries linked to the Data Provider Applications, receive this instruction, take note of their current time, and store it as the *Starting Time* of the recording. During the recording, once a Data Provider has a frame ready to be stored, the Dynamic Library linked to it subtracts the *Starting Time* from the current time and uses this result to provide the frame with a timestamp. By assuming the clocks from the Data Providers run at the same speed, and that all Data Providers received the *StartRecording* instruction almost at the same time, this strategy allows a good enough synchronization of the recorded multimodal data.

*Immediate Feedback:* For behaviors that can be corrected immediately, immediate feedback has proven to be more effective than delayed feedback [18]. This type of feedback provides learners the opportunity to change behaviors while practicing a skill and helps learners to avoid repeating some mistakes that are outside of learners' awareness [19]. Immediate feedback requires from tutors (human or artificial) to

<sup>2</sup> <https://www.json.org/>.

analyze in real-time the learner’s performance, identify mechanisms to improve this performance and transmit feedback instructions that will allow the learner to improve this performance. Learners, on the other hand, need to perform a learning task, pay attention to the feedback provided by tutors and adapt their behavior accordingly. The processes of providing and receiving immediate feedback are both limited by the computational power of the tutor (artificial or human tutor) and the learner. Therefore, immediate feedback needs to be simple in order to be effective. As feedback increases in complexity, the effectivity of delayed feedback over immediate feedback also increases [19].

To keep things simple from the side of the tutor it is recommended to only transmit critical data from the Data Provider Applications to the MLH. An example of critical data could be the instruction to “Speak Louder” in case a microphone application detects that the learner is speaking too soft during the specific learning task. The Dynamic Libraries linked to the Data Provider Applications communicate these types of instructions to the MLH via UDP sockets.

The MLH can then forward the received instructions to applications design to provide feedback to learners. These applications can be ambient displays [20], augmented reality glasses [21], etc. Establishing the communication between the MLH and the immediate feedback applications is a very similar process to the one for establishing the link between the MLH and providers. Before starting a recording, the user configures the feedback applications that will be used. Similar to the case of the data providers, for Feedback Applications, we propose the use of dynamic libraries that handle the communication between Feedback Applications and the MLH (.NET and Windows Universal Platform dynamic libraries are already included in the MLH solution).

### 3 Reliability of the Multimodal Learning Hub

Generating a unified Multimodal Learning Experience out of the data collected by multiple Data Provider Applications presented the biggest challenge in the design and development of the MLH. This challenge leads us to our main research question:

- RQ1: how can the MLH create multimodal learning experiences out of the data captured by multiple Data Provider Applications?

Storing and synchronizing the data captured by the Data Provider Applications are the main issues that had to be addressed in order to answer RQ1. These two issues allowed us to derive the following research questions:

- RQ1a: How can the MLH reliably store the data captured by multiple Data Provider Applications?
- RQ1b: How can the MLH reliably synchronize the data captured by multiple Data Provider Applications?

In the following subsection of this article, we present a series of tests that we conducted in order to identify a reliable solution for the integration (storage and synchronization) of multimodal data.



### 3.1 Multimodal Data Integration Strategies

We identified two main multimodal data integration strategies that could be used for the creation of a Multimodal Learning Experience: real-time data integration and batch data integration. Real-time data integration means that the integration of data is performed during the recording. This strategy facilitates the problem of data synchronization and allows for real-time data analysis. We designed and implemented three solutions for real-time data integration: *Data Collector*, *Direct Push*, and *MQTT Push*. The first two solutions use UDP sockets as communication Protocol. The first solution, *Data Collector*, constantly loops through all the *ApplicationClass* objects (see Sect. 2.3) and appends their respective available data frames to the ongoing recording. Each loop has its own timestamp. For the second real-time data integration solution, *Direct Push*, whenever an *ApplicationClass* object receives a new data frame, it adds a timestamp to it and appends it to the ongoing recording. The third solution, *MQTT Push*, is similar to the *Direct Push* with the difference that the communication between MLH and Data provider Applications uses the MQTT<sup>3</sup> communication protocol, which is a machine-to-machine connectivity protocol designed to minimize network bandwidth whilst attempting to ensure reliability, and has already been used in scenarios where multiple generic sensors have to communicate to a broker.

*Batch Data Integration* means that the data captured by each of Data Provider Applications is stored independently. The integration of this data is done once the recording of the Multimodal Learning Experience finishes. In the case of our solution for the *Batch Data Integration*, each Data Provider Application is responsible for its own recording. Whenever an application has a frame ready, it tags the frame with a timestamp and then appends the tagged frame to its own recording. Once the multimodal recording is finished, each Data Provider Application sends its own recordings to the MLH. In order to reliably synchronize the data through our *Batch Data Integration* solution, it is very important that all Data Providers share a common time reference when timestamping their recorded frames.

### 3.2 Method

We conducted some test runs comparing the previously described solutions for data integration, with the purpose to provide answers to our research questions. For these test runs, it was important to investigate how reliable are the proposed solutions in terms of their capacity to store and synchronize data.

For these tests, we used three different Data Provider Applications: a LEAP Motion<sup>4</sup> Data Provider, a MYO Band<sup>5</sup> Data Provider, and finally the Presentation Trainer [11], which uses a Microsoft Kinect V2 to collect data. The test runs consisted of creating 30 to 40 s recordings with each of the strategies using five different Data Provider configurations:

---

<sup>3</sup> <http://mqtt.org/>.

<sup>4</sup> <https://www.leapmotion.com/>.

<sup>5</sup> <https://www.myo.com/>.

- Singular Configurations
  - only LEAP Motion
  - only MYO Band
  - only Presentation Trainer
- Second Configuration
  - LEAP Motion and MYO Band
- Third Configuration
  - LEAP Motion, MYO Band, and Presentation Trainer

We conducted three test runs for each of the solutions and Data Provider configurations. For the test runs, the hub integrating the data ran on a Windows 7 machine with an Intel Core i7 at 2.50 GHz processor with 16 GB of memory. The LEAP and the MYO controller applications ran on the same computer. The Presentation Trainer ran on a separate Windows 10 computer with an Intel Core i5 at 3.1 GHz processor with 16 GB of memory.

To explore the reliability of the integration solutions in terms of the data storage, we evaluated the recorded files generated during the test runs.

For the Batch Data Integration solution, we also analyzed the synchronization of the captured data. In order to do this analysis we developed two programs:

- A Screen Capture program, which generated a video file of the recordings and was used as a regular Data Provider application.
- A visual test tool<sup>6</sup>, which helps to visualize the multimodal recordings. This tool is able to plot the multimodal data while displaying a media file (video or audio) that belongs to the same multimodal recording.

For the synchronization analysis, we created recordings using the following configuration: MLH and MYO Band Controller running on the Windows 7 computer, and Presentation Trainer and Screen Capture program running on the Windows 10 computer.

### 3.3 Results

Table 1 displays the results of the test runs of the Singular Configurations (only LEAP, only MYO, and only Presentation Trainer) for each of the integration solutions. By looking at the framerate (frames/seconds) recorded during the test runs it is possible to observe that the *Batch Data Integration* solution got the highest frame rates for all the different Data Provider Applications. The *MQTT solution* was the best of the real-time solutions performing very similar to the batch solution when used with the LEAP and the Presentation Trainer applications. In the case of the MYO application, its frame rate seemed to be too high for it to be reliably handled by the real-time integration solutions.

Table 2 displays the results for test runs of the Second Configuration (LEAP and MYO Data Provider Applications running simultaneously). During this test runs, the batch solution outperformed the real-time solutions. Results show that for this Second Configuration, the Batch Data Integration maintained a similar frame rate when

---

<sup>6</sup> <https://github.com/janschneiderou/LearningHub/tree/master/VisualTest>.

**Table 1.** Results of the test runs for single Data Provider configurations.

Strategy		Data collector	Direct Push	MQTT	Batch integration
LEAP	Average File Size	5576 kb	7760 kb	6635 kb	1497 kb
	Average Recording Duration	32.56	27.81	33.33	34.22
	Average Frames Stored	882.67	1228.33	1895.33	2236.67
	Average Framerate (f/s)	27.05	44.35	56.88	65.85
MYO	Average File Size	991.7 kb	749 kb	992.33 kb	2298 kb
	Average Recording Duration	30.94	33.43	32	30.43
	Average Frames Stored	1560.67	1172.33	2446.33	6043.33
	Average Framerate (f/s)	50.19	34.93	76.7	198.59
Presentation trainer	Average File Size	1173 kb	1210 kb	1172.66 kb	1463.33 kb
	Average Recording Duration	31.93	28.72	30	33.53
	Average Frames Stored	692.66	728	912	1030
	Average Framerate (f/s)	21.68	25.34	30.42	30.72

**Table 2.** Results of the test runs for the LEAP and MYO Configuration

Strategy		Data collector	Direct push	MQTT	Batch integration
LEAP	Average Frames Integrated	871	610	1543	1800
	Average Framerate (f/s)	27.23	20.04	44.34	57.2
MYO	Average Frames Integrated	871	542	728.67	6189
	Average Framerate (f/s)	27.23	17.8	21.83	196.75
Totals	Average File Size	1934 kb	3055 kb	9634 kb	14428 kb
	Average Recording Duration	32	30.4	34.33	31.5
	Average Total Framerate (f/s)	27.23	37.85	66.17	254

compared to the frame rate obtained during the runs for the Singular Configurations. In contrast, the real-time integration solutions presented a considerable reduction in their frame rates.

Results for the Third Configuration (LEAP, MYO and Presentation Trainer simultaneous) are displayed in Table 3. These results show a similar trend to the one observed for the Second configuration. For this Third Configuration, the frame rates obtained for the *Batch Data Integration* solution remained stable when compared to the results obtained for the Singular Configurations. In contrast, the real-time integration solutions continued showing a reduction in the obtained framerate when compared to the Second and Singular configurations.

**Table 3.** Results of the test runs for the LEAP, MYO and Presentation Trainer Configuration

Strategy		Data collector	Direct push	MQTT	Batch integration
LEAP	Average Frames Integrated	1445.33	536.67	658	1794.67
	Average Framerate (f/s)	37.6	16.97	18.35	57.08
MYO	Average Frames Integrated	1445.33	570.33	640	6260
	Average Framerate (f/s)	37.6	18.03	17.95	199.1
Presentation trainer	Average Frames Integrated	1445.33	407.33	580.33	964.26
	Average Framerate (f/s)	37.6	12.87	16.53	30.67
Totals	Average File Size	4133 kb	2934 kb	5575 kb	15976 kb
	Average Recording Duration	38.4	32	34.33	31.44
	Average Total Framerate (f/s)	37.6	47.9	52.83	286.85

Overall, the results of the test runs show that the *Batch Data Integration* scaled properly when introducing simultaneous Data Provider Applications, and therefore can be used to reliably store data (See RQ1a). This is in contrast to the real-time solutions that show a reduction in performance with the introduction of simultaneous Data Provider Applications.

With RQ1a answered through the use of the *Batch Data Integration*, we moved to RQ1b and investigated the synchronization of the Batch Data Integration solution. To conduct this investigation, we analyzed test runs using a recording configuration with MYO Band, Presentation Trainer, and a ScreenCapture program. Using the visual test tool we plotted: the Orientation Y values retrieved with the MYO (The MYO was worn on the right arm), the Right Hand Y values and the Left Hand Y retrieved by the Presentation Trainer. The tool also displayed the video recorded by the ScreenCapture application. As seen in Fig. 4 the values of the Orientation Y by the MYO align with the Right Hand values obtained by the Presentation Trainer. The figure also shows how the plotted values align with the current frame of the recorded video when the hand is raised the corresponding hand Y values also increase.

**Fig. 4.** Screenshots of the visual tool displaying the multimodal recording.

## 4 Discussion

This article presents the description of the MLH, a tool that supports the collection and integration of multimodal data from customizable data provider configurations with the purpose to generate Multimodal Learning Experiences out of Meaningful Learning Tasks. One of the main challenges for creating these experiences is expressed in our RQ1 and concerns with the integration of multimodal data coming from multiple and generic Data Provider Applications. To give an answer to RQ1 we designed, developed and tested different possible solutions. Results from our tests allowed us to answer our derived research questions. First (RQ1a), results show that the tested real-time integration solutions are prone to lose data. This tendency to lose data increases with the addition of Data Provider Applications. As aimed to support customizable solutions, it is important for the MLH to scale in terms of the Data Provider Applications that can be used for the generation of a Multimodal Learning Experiences. Results of the test runs show how a *Batch Data Integration* solution is suitable for this scalability.

A concern that we had regarding the *Batch Data Integration* solution, was its capacity to synchronize data from multiple Data Provider Applications. The results presented in this study showed that our proposed data structure for storing multimodal data different from video and audio (See *RecordingObject* Sect. 2.3) and our Multimodal Data Synchronization strategy (See Sect. 2.3) was able to integrate and synchronize recordings of three different Data Provider Applications, hence providing a satisfactory answer to RQ1b.

With the satisfactory answers to our research questions, we consider that the MLH has reached a state where it can be tested by capturing multimodal experiences of Meaningful Learning Tasks such as calligraphy exercises, reanimation training, public speaking, group problem-solving tasks, etc. Testing the MLH in real learning scenarios will help identify its limitations, such as its reliability to forward immediate feedback to generic Feedback Applications. Testing the MLH in real learning scenarios will also provide important information on how to create generic platforms that can be used to capture Multimodal Learning Experiences.

The MLH is our first step in creating customizable and reusable components for MMLA. It is important to mention that the MLH addresses only one of the multiple challenges that have to be taken into account in an MMLA solution. It only deals with the capture and integration of multimodal data for the particular scenario of Meaningful Learning tasks. The *Batch Data Integration* solution of the MLH might not be suitable for other learning scenarios, such as capturing and integrating multimodal data of students' activities (lecture assistance, reading time, sleeping time, etc.) throughout a whole semester. Moreover, as mentioned before, the capture and integration of multimodal data are just one of the many challenges that have to be addressed by an MMLA solution. Some other challenges include the Analysis of multimodal data, storing historical multimodal data, and providing effective interventions for learners. We consider the development and research of customizable MMLA components such as the MLH will contribute to the creation of common specifications, best practices, and standards for MMLA, which in turn will help learners to receive digital support for their ubiquitous learning activities.

## References

1. Siemens, G., Long, P.: Penetrating the fog: analytics in learning and education. *Educ. Rev.* **46**, 30 (2011)
2. Arnold, K.E., Pistilli, M.D.: Course signals at purdue: using learning analytics to increase student success. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 267–270. ACM, New York (2012)
3. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students “game the system.” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 383–390. ACM, New York (2004)
4. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 170–179. ACM, New York (2013)
5. Swan, M.: Sensor Mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *J. Sens. Actuator Networks* **1**, 217–253 (2012)
6. Worsley, M.: (Dis) engagement matters: identifying efficacious learning practices with multimodal learning analytics. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 365–369 (2018)
7. Blikstein, P.: Multimodal learning analytics. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 102–106 (2013)
8. Schneider, J., Börner, D., van Rosmalen, P., Specht, M.: Augmenting the senses: a review on sensor-based learning support. *Sensors* **15**, 4097–4133 (2015)
9. Dermody, F., Sutherland, A.: A multimodal system for public speaking with real time feedback, pp. 369–370 (2015)
10. Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., Castells, J.: The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 360–364 (2018)
11. Schneider, J., Börner, D., van Rosmalen, P., Specht, M.: Can you help me with my pitch? studying a tool for real-time automated feedback. *IEEE Trans. Learn. Technol.* **9**, 318–327 (2016)
12. Hoque, M. (Ehsan), Courgeon, M., Martin, J.-C., Mutlu, B., Picard, R.W.: MACH: my automated conversation coach. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp 2013*, p. 697. ACM Press, New York (2013)
13. Alexandersson, J., Aretoulaki, M., Campbell, N., Gardner, M., Girenko, A., Klakow, D., Koryzis, D., Petukhova, V., Specht, M., Spiliotopoulos, D., Stricker, A., Taatgen, N.: Metalogue: a multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management. In: *2014 International Conference on Intelligent Environments*, pp. 365–368 (2014)
14. Rodríguez-Triana, M.J., Prieto, L.P., Martínez-Monés, A., Asensio-Pérez, J.I., Dimitriadis, Y.: The teacher in the loop: customizing multimodal learning analytics for blended learning. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 417–426 (2018)
15. Dillenbourg, P.: The evolution of research on digital education. *Int. J. Artif. Intell. Educ.* **26**, 544–560 (2016)
16. Di Mitri, D., Schneider, J., Specht, M., Drachsler, H.: From signals to knowledge. A conceptual model for multimodal learning analytics, *JCAL* (2018)

17. King, P.E., Young, M.J., Behnke, R.R.: Public speaking performance improvement as a function of information processing in immediate and delayed feedback interventions. *Commun. Educ.* **49**, 365–374 (2000)
18. Coulter, G.A., Grossen, B.: The effectiveness of in-class instructive feedback versus after-class instructive feedback for teachers learning direct instruction teaching behaviors. *Eff. Sch. Pract.* **16**, 21–35 (1997)
19. Hattie, J., Timperley, H.: The power of feedback. *Rev. Educ. Res.* **77**, 81–112 (2007)
20. Börner, D., Kalz, M., Specht, M.: Beyond the channel: a literature review on ambient displays for learning. *Comput. Educ.* **60**, 426–435 (2013)
21. Guest, W., et al.: Affordances for capturing and re-enacting expert performance with wearables. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017. LNCS*, vol. 10474, pp. 403–409. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_34](https://doi.org/10.1007/978-3-319-66610-5_34)



# How Teachers Prepare for the Unexpected Bright Spots and Breakdowns in Enacting Pedagogical Plans in Class

Ghita Jalal<sup>1</sup>(✉), Valentin Lachand<sup>1</sup>, Aurélien Tabard<sup>2</sup>, and Christine Michel<sup>1</sup>

<sup>1</sup> Univ Lyon, INSA-Lyon, CNRS, LIRIS, UMR5205, 69621 Lyon, France  
{ghita.jalal, valentin.lachand, christine.michel}@liris.cnrs.fr

<sup>2</sup> Univ Lyon, Université Lyon 1, CNRS, LIRIS, UMR5205, 69621 Lyon, France  
atabard@liris.cnrs.fr

**Abstract.** When teachers plan pedagogical activities, they define the pedagogical strategies, resources, and tools they will use. But, as they run these activities in class, they have to adjust their plans, according to available resources, and to live breakdowns. Teachers have very little time to adjust their plans in class, and existing tools offer very little support for live changes. We conducted contextual interviews with eight middle and high school teachers to better understand their practices in planning and enacting pedagogical activities. We identify a set of breakdowns in conducting their activities, and the strategies teachers develop to cope with them. Teachers use digital tools to keep a trace of their plans and to improve their enactment strategies. They design plans students can enact directly, or define the content, the structure, or both, with students in class. Most enactment issues are software and hardware breakdowns. Based on our findings, we propose implications for the design of novel tools to support teachers in enacting their plans in class. These tools should capture traces of the activity as it happens. They should support externalizing plans, and sharing them with students. Ultimately, planning and enactment tools should support richer cross-device interactions.

**Keywords:** Interviews · Qualitative study · Teacher practices  
Planning · Teaching tools

## 1 Introduction

Pedagogical plans are externalizations of learning activities as teachers anticipate them. As they enact these plans during the session, teachers refine, adapt and reflect on them on the go. Sharples [20] describes the complexity of the teacher's role: *“not only [s/he] has to prepare lesson plans, accommodate formal curricula, and follow regulations on health, safety and discipline, but also understand and manage a variety of technologies such as interactive whiteboards, desktop and laptop computers.”*



As they plan learning activities, teachers know their plans are likely to change as the activity unfolds. Yet, they need to prepare the structure and content they intend to include in the session. As they enact their plans in class, teachers know more about the activity. They can refine their plans, or adjust them depending on the situation. Yet, teachers often make these changes in a few seconds or minutes, while running the session at the same time. This is especially challenging when taking into account the pedagogical and technical constraints teachers manage at the same time when they run their sessions [8].

There is a tension between planning and enacting pedagogical activities [4]. As a result, teachers need to switch between routines and improvisations. Routines are practices they developed over the years. Improvisations are quick fixes they put in place, during the session, to respond to events they did not expect in their plans [15]. After class, teachers have more time to revisit their plans, edit them, rethink their routines, and evaluate their improvisations. Yet, they do not have access to context elements they experienced first-hand, during the session.

In this paper, we investigate the gap between the plans teachers create, and how they enact them in class. Our end goal is to propose interactive tools that support teachers in enacting and adapting their plans in class. We conducted contextual interviews with middle and high school teachers. We report on their routines and practices as they plan and enact pedagogical activities. We focus on breakdowns and bright spots in enacting teacher plans, and propose design recommendations to create tools to support teachers in the transition from planning to enactment.

## 2 Related Work

Scripting [23] and Orchestration [5] provide descriptive and generative guidelines to design tools that support teachers in planning and enacting their plans in class. We discuss how teachers use existing tools, the limitations of these tools, and how paradigms such as scripting and orchestration can help us understand teachers' practices in planning and enacting pedagogical activities in class.

### 2.1 Plans and Action in Social Sciences

Plans and action have long been used in sociology to describe and formalize the tension between how plans condition and define action. Akrich [1] compared plans to interaction “scripts” or “scenarios” that await for actors to enact them, and transform them into technical objects [1]. Suchman’s work on situated action emphasized how plans are not enough to ensure successful interaction: plans unfold as “*ad-hoc responses to the actions of others and to the contingencies of particular situations*” [22]. Streibel discussed and interpreted plans and situated action in learning [21]. Instructional plans determine the cognitive model of human learning, but cannot control situated learning [21]. These theories describe how plans and action interplay in users’ practices. While plans condition action, they do not determine how it unfolds.

We use these theories to frame our empirical findings, while focusing on what field observations can teach us about the design of novel interactive tools to support teachers' transition from plans to action.

## 2.2 Plans in Pedagogical Situations

In educational settings, Dore describes teachers' plans and pedagogical strategies as "*techniques and means used to reach [an] educational goal*" [10]. Several models describe pedagogical plans. The narrative model [10, 16] structures pedagogical plans at three levels: courses, activities and steps. A learning scenario describes course elements. These include domain knowledge, curriculum, aimed age, school level, and learning goals. It also describes elements more specific to each activity. These include required skills, teacher and student tools, phases, and assessment.

Models such as LOM (Learning Object Metadata), SCORM (Sharable Content Object Reference Model) or IMS-LD (Instructional Management Systems-Learning Design) [16] base their structure on these principles. These models describe pedagogical objectives and individual learning activities. Yet, they grow in complexity when describing collaborative activities where students' and teachers' roles are dynamic. Also, these pedagogical models do not account for the changes in pedagogical plans, and the challenges teachers face in enacting them in class.

Scripting is another approach to define plans for collaborative activities. It focuses on the way students collaborate [7]. CSCL scripts define more precisely how group members interact to solve a problem. There are two levels of scripts: micro-scripts and macro-scripts [6]. Micro-scripts are models students need to internalize (local perspective), such as argumentation or dialogue models whereas macro-scripts are pedagogical models (global perspective). One of the main differences between micro-scripts and macro-scripts is duration. Micro-scripts are short-termed and students need to internalize them. Macro-script cover longer periods and are directly linked to pedagogical objectives.

Kobbe [14] identifies the following script components: activity participants, groups and roles assigned to group members (roles are "*associated with privileges, obligations and expectations*"), and activities. In this model, scripts structure pedagogical activities and learners' resources. For Dillenbourg and Hong [6], script components are: activity type, sequencing in time, participants' roles, distribution and activity representation. This model mostly adds pedagogical objectives to Kobbe's model.

Scripts describe plans teachers design to anticipate the dynamics of collaborative activities. Learning design provides a broader perspective on planning. The term design here refers to: "the process of mapping and/or actually developing specific resources for teaching or learning" [13].

### 2.3 Plans in Practice

Few empirical studies focus on how teachers use theoretical models in practice, to script real pedagogical activities. Dore et al. found that the narrative reference model can guide training to clarify a teaching frame for students. But, they still need assistance for novel forms of training at school or outside [10]. Rodríguez-Triana et al. [17] conducted two studies of an implementation of a model combining learning design (scripting) and learning analytics (conducting). They found that designing scripts with monitoring information helps the teacher anticipate what can happen in class during the scripting phase.

### 2.4 Conducting Pedagogical Activities

Research on learning has explored planning and conducting pedagogical activities since its inception. In particular, the orchestration metaphor has been increasingly used to describe the “live” management of unfolding activities in the classroom [7]. Work on orchestration proposes principles to structure a training timeline (or graph) to support teachers in conducting educational activities. This structure takes into account a number of practical constraints (length, curriculum, number of students, etc.) [5] and ways to improve activity progress (continuity, awareness, relevance, etc.) [8].

*Primo-scripting* is an orchestration phase where the teacher identifies constraints and pedagogical objectives [23]. In primo-scripting, teachers create a scenario with available resources and strategies to implement this scenario in the classroom. *Run-time scripting* is an orchestration phase where teachers edit scripts live. It helps them reconsider their activity’s structure, implementation and teaching objectives [23]. Orchestration becomes challenging when there is a division between learning at school and outside [9] (e.g., homework). Sharples et al. proposed shared orchestration [11, 19, 20] as a new way of conducting activities where teachers and learners can orchestrate their own activities.

Orchestration tools support enacting pedagogical activities in class [8, 9, 18]. Live monitoring dashboards give teachers feedback about learners’ progress in multi-device contexts [15], but at the cost of extra mental workload [20]. Tangible devices create ambient awareness for teachers [9]. For example, Lantern [8], an orchestration lamp, changes color to inform teaching assistants about students’ progress in problem solving sessions.

### 2.5 Transitioning from Scripting to Orchestration

Scripting tools support creating plans before class, while orchestration tools support enacting these plans in class. Yet, scripting and orchestration do not support the transition from planning to enacting plans in class. To our knowledge, teachers have little to no technical support in managing this transition. Orchestration literature also rarely discusses what happens after class. Orchestration systems do not focus nor support teachers’ post-session reflections to adapt and reuse their plans for future sessions. We focus on how teachers currently manage this

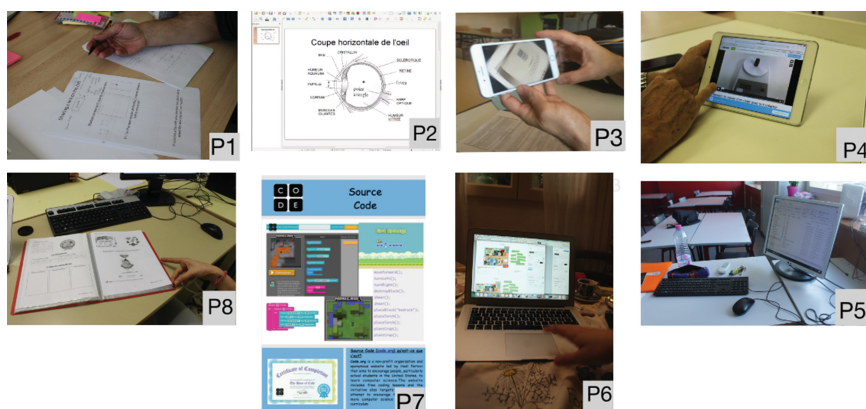
transition as they prepare and run learning activities. We follow a methodology similar to [24], to investigate teachers practices with and without digital tools. In the following we present results of contextual interviews with teachers, and highlight bright spots and breakdowns in their attempts to enact their plans.

### 3 Study

We conducted contextual interviews to better understand the interplay between digital tools and teachers' practices in planning and enacting pedagogical activities. We focused on moments where pedagogical plans did not proceed as intended and on how teachers dealt with these situations.

#### 3.1 Participants and Procedure

We interviewed eight French middle and high school teachers (3 women, 5 men; age 26–50; 5 in middle school, 3 in high school) about their practices in planning and enacting pedagogical activities. Teaching topics include French literature, Physics, Chemistry, History, English, German, Biology and Computer Science.



**Fig. 1.** During the interviews, teachers showed us how they created their plans, and described how they used them to enact the session.

We conducted semi-structured interviews with participants in their classroom or office for about one hour. We asked participants to walk us through the planning and enactment steps of a specific teaching session. We also asked them to show us the documents they created before, during and after the session. We probed for situations where planning or enacting was particularly effective, but also when it was extremely difficult.

### 3.2 Data Collection and Analysis

We recorded audio for each interview and took written notes. We also recorded videos of participants’ interactions with the documents they had created, and photographed relevant elements of their classroom settings (position of student tables, interactive board, tablets, routers). We transcribed the eight interviews, and extracted examples of pedagogical moments -stories- where teachers enacted plans they created before the session. We used thematic analysis [2] to extract themes that describe how teachers plan and enact pedagogical activities. We considered how teachers plan their sessions before class, and how they use these plans as they enact their session with students. We also identified main types of breakdowns teachers reported as they attempted to enact their plans in class, and the tools (digital or physical) they used to plan and run the activity. We created a visual representation of each story [12] to validate it, and to gather more contextual information in a second meeting with the participant.

In the following, we present and discuss how teachers use current tools to plan and enact pedagogical activities in real classroom situations.

## 4 Results

We extracted 48 stories in total (between 2 and 11 stories by participant). Each teacher in our interview walked us through a session they recently run with their students. These narrative descriptions of teachers’ actions to prepare and run pedagogical activities helped us identify several activity structures. All participants alternated group and individual activities in their sessions.

In the following, we report on how teachers in our interviews planned their sessions, and how they enacted them in class. We focus on the tools they used, and on the breakdowns and bright spots in their enactment strategies.

### 4.1 How Do Teachers Plan a Pedagogical Activity?

All participants planned their sessions before class. Teachers in our interviews used different names to describe the pedagogical plans they created. P2, a physics teacher, called the plan: “a connecting thread”, and a “contract” between him and students. P3, a history teacher, talked about a “work plan”, referring to the technique he used to construct pedagogical activities [3].

**Table 1.** Types of tools used to run pedagogical activities in class (percentages)

Planning tool	Teacher tool				Student tool			
	No tool	Digital	Physical	<b>Total</b>	No tool	Digital	Physical	<b>Total</b>
No plan	1.9	1.3	9.4	<b>22.6</b>	0.0	13.2	9.4	<b>22.6</b>
Digital	26.4	35.8	11.3	<b>73.6</b>	0.0	41.5	32.1	<b>73.6</b>
Physical	1.9	0.0	1.9	<b>3.8</b>	0.0	0.0	3.8	<b>3.8</b>

Teachers used digital tools to plan their sessions in almost 3/4 of the situations they described, whether they used digital tools to conduct the activity in class (35.8%) or not (26.4% + 11.3%). They also created digital plans in sessions where they did not assign students to use digital tools (32.1%).

In about a third of the stories (35.8%), teachers created a plan directly in a digital tool, and later used the same tool to conduct their session. For example, P5, an English teacher, planned a session in h5p<sup>1</sup>, an online teaching tool for creating interactive content. P5 and her students both used the same tool to run the activity in class.

Students could follow the teacher's plan autonomously in around 1/4 of the stories. In these cases, teachers let students run the plan, and provided feedback as needed. Some teachers (32.1%) also created digital plans, and printed them for students to use in class. Teachers used these digital versions to keep a trace of the session progress for future years. For some courses (22.6%), teachers did not create representations of their plans before the session (Table 1). In these cases, teachers had in mind the structure they would follow. They established routines they followed in several sessions. Both teachers and students were aware of these routines. For example, P4, a physics teacher, always starts with questions, follows with a short experiment, and another series of questions. The session structure in this case is implicit. P4 does not create a representation of the plan before class, but he and the students know how the session will proceed.

Our results suggest that teachers often create representations of their plans ahead of time. Digital plans helped teachers integrate the session structure and student activity in the same tool. Few digital plans could be run autonomously by the students.

## 4.2 What Goes in Teacher Plans?

The teachers we interviewed left parts of their plans open, and defined them in class, as the activity unfolded. We found that preparation could be organized around the structure of the activity or its content.

**Planning Content First, and Defining Structure in Class:** More than half participants (five out of eight), prepared or created content before class, only to decide in class on how they would present it to students. For example, P5, an English language teacher, used a Web application, *Genially*, to add dynamic links to a painting. In class, she decided of the order in which she opened and presented the links based on her discussion with students.

Teachers prepare the content, and use it to guide the discussion, depending on their interaction with students in class. Two participants provided examples where they prepared several versions of the content. In class, they decided which version to use depending on how the session unfolded. P4, a physics teacher, created many versions of the same activity, with different levels of difficulty.

---

<sup>1</sup> [www.h5p.org](http://www.h5p.org).

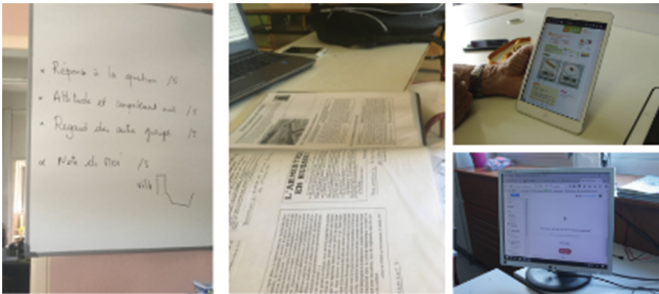
He started with a less detailed version, and provided more details as he perceived students struggling.

**Planning Structure First, and Defining Content in Class:** Half participants (four out of eight) represented the session’s structure in their plans, and then created the content with students, in class. For instance, P5, an English language teacher, came to class knowing the structure, but created the content in class, with students. She created a mind map, before class. Then, in class, she filled in the content with students: *“I wanted to know the vocabulary they already know.”* (P5) One participant created both the structure and content before class. Only to re-create the content with students in class. As she explained: *“I was cheating, I led them where I wanted them to go”* (P8).

Teachers set constraints in the planning phase. They do not create fully detailed plans. At the same time, they do not leave the session totally open. They keep a level of freedom for them to adjust the plan according to what happens in class.

**4.3 Enactment Bright Spots and Breakdowns**

During the interviews, all eight teachers presented a version of their plan for the session we discussed. They used the plan in class with students. P1, a French literature teacher, writes the plan on the blackboard before the session begins, and presents it to students to start the session. Participants often used a printed version of the plan, or a digital version on a mobile device, such as a tablet. In some cases, they also had versions of the plan on a static computer in the classroom (Fig. 2).



**Fig. 2.** Teachers externalized their plans on different media. They showed us plans on the physical board, on paper, on a digital tablet, or on the classroom’s computer.

Half our participants presented pedagogical situations where they changed their plans in class. New content, structure, or live events lead them to change the plan they initially anticipated. Teachers make room for changes in their plans, and attempt to work around live breakdowns to reach their pedagogical objectives.

To understand if planning helped teachers succeed in conducting their session, we extracted satisfaction statements from participants' stories about the session they described and mapped these statements to the tools they used to plan this specific session (table 2). Bright spots and breakdowns in participant stories are time, space and resources issues:

- *Time issues are about the expected versus the actual time it took the teacher to run the session.*
- *Physicality issues [5] include situations where teachers needed to be mobile in the classroom.*
- *Resource issues are about content transfer and distribution. For example, these issues include accessing plans the teachers created at home on the school computer, managing software versions, and distributing content on several devices.*

**Table 2.** Teachers' satisfaction of their enactment strategies by plan type (Percentages)

Plan type	resources		time		space	
	(-)	(+)	(-)	(+)	(-)	(+)
No plan	3.8	18.9	5.7	17.0	1.9	20.8
Digital	18.9	54.7	20.8	52.8	3.8	69.8
Physical	0	3.8	0	3.8	0	3.8
<b>Total</b>	<b>22.6</b>	<b>77.4</b>	<b>26.4</b>	<b>73.6</b>	<b>5.7</b>	<b>94.3</b>

To understand if teachers managed to run their session with the tools they planned to use, we identified stories where breakdowns occurred and mapped them to the tools teachers and students used during the session (Table 3).

**Enactment Bright Spots:** In most stories (more than 70%), teachers were satisfied with their enactment strategies. They were satisfied of their strategies in managing resources (77.4%), time (73.5%), and space (94.3%) (Table 2).

In many cases (more than 50%), interviewees chose to plan their sessions with digital tools (Table 2). Teachers did not articulate their involvement during the session explicitly in the interviews. Yet, their plans reflected different levels of involvement in the activity in class. In most stories, teachers closely monitored the activity in class.

Teachers in our interviews gave various examples of routine plans they reused to conduct several pedagogical activities. They created digital plans to keep a trace of their practices, and to improve their enactment strategies over time. Teachers also created detailed digital plans for students to run independently. *“I prefer this type of activities because students can finish them at home.”* (P5). Other teachers used this strategy to have more time in class to answer students' questions. For example, P5 created an interactive video for students using



*Edpuzzle*, a teaching web application for interactive videos. Using this tool, she could see, in class, sections of the video students viewed the most. P5 adjusted her plan to spend more time on these problematic sections. Other teachers created detailed plans to make sure they covered all educational objectives for the session. P2, a physics teacher, created a plan to guide him through the session. He created a checklist with important points to search for in students’ answers to his questions in class. He used a printed version of the checklist during the session.

**Table 3.** Breakdowns and Bright Spots by tool type (Percentages)

	Teacher tool	Student tool		<b>Total</b>
		Digital	Physical	
Breakdowns	No tools	3.8	3.8	7.5
	Digital	35.8	1.9	37.7
	Physical	0.0	9.4	9.4
	<b>Sub-total</b>	<b>39.6</b>	<b>15.1</b>	<b>54.7</b>
Bright spots	No tools	5.7	17.0	22.6
	Digital	9.4	0.0	9.4
	Physical	0.0	13.2	13.2
	<b>Sub-total</b>	<b>15.1</b>	<b>30.2</b>	<b>45.3</b>

In other cases, the plan was limited or intentionally open for the teacher to add explanations, details, or examples. Teachers gave instructions live, and recreated the content with students in class. Their goal was to maintain interaction and student involvement during the session. For example, P8, a biology teacher, provided students with a “session plan”, with the structure of the activities they will run in class. Students filled in the plan with answers to P8 questions during the session. Then, P8 copied the answers in class, and uploaded a version of the “session plan” to the school’s digital system.

Enacting teacher plans in class requires them to take into account potential breakdowns. They should be able to adjust their plans live, to use alternative tools, and to change instructions and content depending on unexpected events during the session.

**Breakdowns in Enacting Pedagogical Plans:** Interviews with teachers revealed different types of breakdowns in enacting plans they created before class (Table 4). Most breakdowns in enacting the session are related to time (26,4%) and resources (22,6%) (Table 2). In more than half of the stories (54.7%), teachers did not manage to run their sessions as intended (Table 3). Breakdowns were more frequent (35.8%) when teachers planned and run the session with digital tools (Table 3).

**Table 4.** Number of stories and participants per breakdown type

Breakdowns	Stories	Participants
Software	16	7
Hardware and Network	10	6
Content and instructions	3	2

Content and instruction breakdowns are cases where teacher plans did not correctly respond to unexpected live events in class. We found examples of pedagogical moments where teachers changed instructions and content live, based on students' feedback in class. For example, P3 a history teacher, assigned a group to work on writing a biography. As they started, P3 realized that students were writing a full textual biography. He adjusted the instructions to ask for the birth and death dates, and for major events in the life of the character. Similarly, P5, an English teacher, created an activity around a Martin Luther King video. Although she designed the activity for students to regulate on their own, she stopped in class after each video section: *"This video content is too difficult, I want to explain the words as we go"* (P5).

More than half participants (six out of eight), provided examples of hardware breakdowns. In these cases, teachers' plans broke when they moved them across different devices. They are also linked to content access from different locations, and from different devices. For example, P2, a physics teacher, replicated his plans on dropbox, and on a USB stick, to avoid loosing them when moving them out of the classroom computer. Several teachers also reported on hardware breakdowns related to sharing hardware among students or student groups. For example, P1, a French literature teacher, used a personal tablet for an activity where student groups created a movie. In class, P1 needed to make sure all groups had access to the tablet when they wanted to start filming.

Software breakdowns were more recurrent in participant stories. Almost all participants (seven out of eight) provided examples of specific moments in the session, where they did not manage to enact their plans because of software breakdowns. For example, P8, a biology teacher, could not access, or edit her plans in class, because the installed version on the classroom's computer does not open her files.

Teachers who used software tools less often still presented software breakdowns. Yet, these breakdowns were less frequent compared to teachers who tried and used more software tools to plan and enact their sessions. P5, an English teacher, reported on four different alternatives to plan an activity with interactive videos. In one activity, she used Edpuzzle<sup>2</sup>, an educational software that supports annotating videos, cutting video sections, and adding questions. The problem with Edpuzzle: *"students need to open a new window, they go out of moodle"*<sup>3</sup> (the educational platform used in her institution)" (P5). In a similar

<sup>2</sup> <https://edpuzzle.com/>.

<sup>3</sup> <https://moodle.org/>.

activity, P5 used h5p, another interactive tool for teachers. While h5p is a moodle plugin, she needed to spend time in class explaining how students can access the different parts of the video, and how they could use the codes she generated ahead of time. A third option she used for this type of activities consisted of cutting the video using MovieMaker, and adding the questions on a MS Word document. She would play the video section, and follow with the questions. The fourth option she presented consisted of preparing the questions before the session, based on the video content. In class, she would play the video, and stop manually at the end of the first section, and ask the questions. As she presented this alternative, P5 said: *“I am getting a mobile keyboard. It will be great for this type of activities. I will be able to stop the video without having to stay close to the computer”*.

## 5 Implications for Design

We believe teaching tools should account for the challenges teachers face as they transition from plans to action, and we propose specific guidelines to support this transition.

### 5.1 Capturing the Activity, as it Happens

Teachers used digital tools to keep traces of the activity after its end. While they managed to keep track of the structure and content they followed, they could not keep a trace of the changes they made to the plan, and of how they responded to unexpected events. These traces could help teachers better adapt to a specific student group, classroom, or content. They could also help teachers reflect on their practices, and improve them for upcoming years.

Pedagogical activity planning and enactment tools should support capturing content and instructions as they are enacted in class. Most teachers in our interviews added content, instructions, and changed their activity soon after the session. For example P3, a history teacher, mentioned: *“I take notes on a sheet of paper. Then, I edit my plan in the evening. I do this the same day.”*

Capturing traces could take different forms. Teachers could take pictures, record audio, or create their own way of capturing traces they find important during the session. Tracing can also happen implicitly, as the teacher or students change the plan in class. Then, after the session, the teacher can compare the planned session, to what actually happened in class and improve the plan for future sessions.

### 5.2 Externalizing and Sharing Plans

All teachers in our interviews shared their plans with students before class. While the teacher and students pointed at several resources and in-class activities, they both came back to the plan, on regular basis, to track and regulate their progress. Teachers need a shared representation of the plan from where they can point

to other content, questions and instructions. In this representation, the plan becomes a shared communication channel between the teacher and students. They both create, edit, and complement the plan as the session unfolds.

### 5.3 Supporting Richer Cross-Device Interaction

Several breakdowns occurred when teachers moved files from one device to another. This resulted in losing formatting, content, or wasting time. The multiplicity of applications and devices in teachers' practices raises interoperability problems planning and enactment tools should address. They should support teachers in organizing resources in the planning phase, and link these resources to the plan as it is enacted.

Several teachers emphasized issues related to limited storage space and network speed. Planning and enactment tools should be designed around the storage and network constraints of the classroom environment.

## 6 Conclusions and Future Work

In this paper, we investigate ways in which teachers manage the transition between the plans they design before class, and the activity as it unfolds in class with students. Our primary focus is on current tool use, and how it can inspire the design of novel tools to support teachers in transitioning from plans to action.

We conducted contextual interviews with middle and high school teachers, and found that most plans are created using digital tools. We identified breakdowns and bright spots in current teachers' practices before and during the session. We found that most breakdowns occurred when teachers used digital tools before and during the session. Our findings confirmed several initial hypotheses on unexpected events during class. These unexpected events make the transition from plans to action complex for teachers. For example, network and hardware issues were recurrent in participant stories, and often created problems with document transfer across different devices. We found that teachers create planning strategies to work around these problems. For example, they include lightweight versions of the content they want to use in their plans. They make sure the content is accessible in class, while still positioning it correctly in the session structure.

Khakaj et al. also conducted contextual interviews with teachers to investigate how they collect data about students [24]. Our approach focuses on how teachers tool use affects their planning and enactment practices over time. We look at current breakdowns, but also analyze situations in which teachers succeeded in enacting their plans, and how their strategies in these situations could inspire the design of novel tools to support them in the transition between pedagogical plans and action.

In addition to the empirical work we present in this paper, we are building prototypes to demonstrate our design implications. We are currently running

co-design sessions with teachers, and working together to investigate how to best integrate prototypes in their current practices in planning and enacting pedagogical activities.

## References

1. Akrich, M.: Comment décrire les objets techniques? *Techniques et culture* **9**, 49–64 (1987)
2. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006)
3. Dalland, C.P., Klette, K.: Individual teaching methods: work plans as a tool for promoting self-regulated learning in lower secondary classrooms. *Educ. Inq.* **7**(4), 28249 (2016)
4. Dillenbourg, P.: Trends in orchestration. Second research & technology scouting report. Technical report (2011)
5. Dillenbourg, P.: Design for classroom orchestration. *Comput. Educ.* **69**, 485–492 (2013)
6. Dillenbourg, P., Hong, F.: The mechanics of CSCL macro scripts. *Int. J. Comput.-Support. Collab. Learn.* **3**(1), 5–23 (2008)
7. Dillenbourg, P., Jermann, P.: Designing Integrative Scripts. In: Fischer, F., Kollar, I., Mandl, H., Haake, J.M. (eds.) *Scripting Computer-Supported Collaborative Learning*, vol. 6, pp. 275–301. Springer, US (2007)
8. Dillenbourg, P., Zufferey, G., Alavi, H.S., Jermann, P., Do, L.H.S., Bonnard, Q., Cuendet, S., Kaplan, F.: Classroom orchestration: the third circle of usability. In: *Connecting CSCL to Policy and Practice: CSCL Conference Proceedings. Volume I – Long Papers*, vol. 1, pp. 510–517. Hong Kong, China (2011)
9. Dimitriadis, Y., Prieto, L.P., Asensio-Pérez, J.I.: The role of design and enactment patterns in orchestration: Helping to integrate technology in blended classroom ecosystems. *Comput. Educ.* **69**, 496–499 (2013)
10. Doré, S., Basque, J.: Le concept d’environnement d’apprentissage informatisé. *Int. J. E-Learn. Distance Educ.* **13**(1), 40–56 (2007)
11. Fong, C., Cober, R.M., Moher, T., Slotta, J.D.: The 3R orchestration cycle: fostering multi-modal inquiry discourse in a scaffolded inquiry environment. In: *Proceedings from the Annual meeting CSCL Conference* (2015)
12. Jalal, G., Maudet, N., Mackay, W.E.: Color portraits: from color picking to interacting with color. In: *Proceedings of the 33rd Conference on Human Factors in Computing Systems*, pp. 4207–4216. ACM CHI 2015. ACM, New York (2015)
13. Kali, Y., McKenney, S., Sagy, O.: Teachers as designers of technology enhanced learning. *Instr. Sci.* **43**(2), 173–179 (2015). <https://doi.org/10.1007/s11251-014-9343-4>
14. Kobbe, L.: Framework on multiple goal dimensions for computer-supported scripts, kaleidoscope, d21. 2.1 (2006)
15. Looi, C.K., Song, Y.: Orchestration in a networked classroom: where the teacher’s real-time enactment matters. *Comput. Educ.* **69**, 510–513 (2013)
16. Paquette, G., Léonard, M., et al.: *Modèles et métadonnées pour les scénarios pédagogiques* (2013)
17. Rodríguez-Triana, M.J., Martínez-Monés, A., Asensio-Pérez, J.I., Dimitriadis, Y.: Scripting and monitoring meet each other: aligning learning analytics and learning design to support teachers in orchestrating CSCL situations. *Br. J. Educ. Technol.* **46**(2), 330–343 (2015)

18. Roschelle, J., Dimitriadis, Y., Hoppe, U.: Classroom orchestration: synthesis. *Comput. Educ.* **69**, 523–526 (2013)
19. Sharples, M., Scanlon, E., Paxton, M., Kerawalla, L., Feisst, M., Gaved, M., Wright, M., Collins, T., Anastopoulou, S., Mulholland, P.: nQuire: technological support for personal inquiry learning. *IEEE Trans. Learn. Technol.* **5**, 157–169 (2012)
20. Sharples, M.: Shared orchestration within and beyond the classroom. *Comput. Educ.* **69**, 504–506 (2013)
21. Streibel, M.J.: Instructional plans and situated learning: the challenge of suchman's theory of situated action for instructional designers and instructional systems. *J. Vis. Lit.* **9**(2), 8–34 (1989)
22. Suchman, L.A.: *Human-machine reconfigurations: plans and situated actions*. Cambridge University Press, Cambridge (2007)
23. Tchounikine, P.: Clarifying design for orchestration: orchestration and orchestrable technology, scripting and conducting. *Comput. Educ.* **69**, 500–503 (2013)
24. Khakaj, F., Alevan, V., McLaren, B.M.: How teachers use data to help students learn: contextual inquiry for the design of a dashboard. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) *Adaptive and Adaptable Learning*, pp. 340–354. Springer International Publishing, Cham (2016)



# Evaluating the Robustness of Learning Analytics Results Against Fake Learners

Giora Alexandron<sup>1</sup>(✉), José A. Ruipérez-Valiente<sup>2</sup>, Sunbok Lee<sup>3</sup>,  
and David E. Pritchard<sup>2</sup>

<sup>1</sup> Weizmann Institute of Science, Rehovot, Israel  
`giora.alexandron@weizmann.ac.il`

<sup>2</sup> Massachusetts Institute of Technology, Cambridge, MA, USA  
`{jruipere,dpritch}@mit.edu`

<sup>3</sup> University of Houston, Houston, TX, USA  
`slee96@uh.edu`

**Abstract.** Massive Open Online Courses (MOOCs) collect large amounts of rich data. A primary objective of Learning Analytics (LA) research is studying these data in order to improve the pedagogy of interactive learning environments. Most studies make the underlying assumption that the data represent truthful and honest learning activity. However, previous studies showed that MOOCs can have large cohorts of users that break this assumption and achieve high performance through behaviors such as Cheating Using Multiple Accounts or unauthorized collaboration, and we therefore denote them *fake learners*. Because of their aberrant behavior, fake learners can bias the results of Learning Analytics (LA) models. The goal of this study is to evaluate the robustness of LA results when the data contain a considerable number of fake learners. Our methodology follows the rationale of ‘replication research’. We challenge the results reported in a well-known, and one of the first LA/Pedagogic-Efficacy MOOC papers, by replicating its results *with* and *without* the fake learners (identified using machine learning algorithms). The results show that fake learners exhibit very different behavior compared to true learners. However, even though they are a significant portion of the student population (~15%), their effect on the results is not dramatic (does not change trends). We conclude that the LA study that we challenged was robust against fake learners. While these results carry an optimistic message on the trustworthiness of LA research, they rely on data from one MOOC. We believe that this issue should receive more attention within the LA research community, and can explain some ‘surprising’ research results in MOOCs.

**Keywords:** Learning analytics · Educational data mining · MOOCs  
Fake learners · Reliability · IRT

## 1 Introduction

The high resolution behavioral data that MOOCs collect provide new opportunities to study learners behavior, in order to improve the pedagogy of interactive

learning environments, and to develop data-driven tools for personalization and analytics [10,20]. The implicit assumptions behind such research are typically that the data collected represent genuine learning behavior, and that there are hidden causal relationships between learners behavior and their success, which can be discovered using Educational Data Mining (EDM).

*Fake Learners.* However, several studies revealed that there are a considerable amount of users who use cynical means to succeed in the courses, such as Cheating Using Multiple Accounts [3,4,15,18], or unauthorized collaboration [19]. Such users break the ‘genuine learning behavior’ assumption, thus we refer to them as *fake learners*. The data in fake learners logs is largely an artifact with respect to explaining their performance. As was pointed out in [4], this can bias LA and EDM results. For example, fake learners typically make minimal interaction with the learning materials, yet show high success; this can lead to false conclusions regarding the effectiveness of different learning paths or the pedagogic efficacy of course materials. However, this issue remains an open question.

*Research Questions.* The goal of the current research is to address this issue directly by measuring the effect of fake learners on LA research. Specifically, we study the following Research Questions (RQs):

1. (RQ1) What is the difference between the ‘fake’ and ‘true’ learners with respect to the amount of use of different course materials (e-text, videos, checkpoint items, homework, and quizzes), and to various performance measures?
2. (RQ2) What is the effect of fake learners’ data on the results of a correlation study, such as the relationships between resource use and performance?

To answer these, we challenge the findings reported in one of the first studies of pedagogic efficacy in MOOCs [7], by replicating its results with and without fake learners data. To identify the fake learners, we use the algorithms published in [4,19].

*Findings in brief.* In the course that we study about ~15% of the certificate earners are fake learners (of the types that we can detect; we expect that there are more that are still under the radar). With respect to RQ1, they have a very distinguished learning behavior (e.g., a much lower use of course materials). With respect to RQ2, their data effect the correlations that were studied in [7] in a way that we interpret as not very significant (i.e., no ‘change of trend’).

*Our contribution.* Due to the large amount of fake learners that were reported in MOOCs, the risk that fake learners’ data can bias LA discoveries raises doubts on the trustworthiness of such studies. However, identifying and removing such learners from the data requires sophisticated algorithms that are not available off-the-shelf. We build upon our previous research on both identifying fake learners, and pedagogic efficacy in MOOCs, to make a stride in the direction of evaluating the robustness of LA research against fake learners. To the best of our knowledge, this is the first rigorous attempt to study this issue.



*A broader perspective.* This research also touches upon two issues that we believe should receive much more attention within the LA and EDM communities. One is *verification* and *validation* of computational models that rely on noisy data that its quality can be affected by malicious or otherwise unusual behavior. Second is *replication research* as a scientific methodology to explore and confirm the generalizability of LA and EDM results to different educational contexts and their stability under various conditions.

## 2 Methodology

In this section we describe in brief the experimental setup and the EDM procedures that are used. Some of the methodological contents of this section have been reused from previous work [4, 18].

### 2.1 Experimental Setup

The context of this research is MITx MOOC 8.MReVx, offered on edX.org in Summer 2014<sup>1</sup>. The course attracted 13500 registrants, of which 502 earned a certificate. Gender distribution was 83% males, 17% females. Education distribution was 37.7% secondary or less, 34.5% College Degree, and 24.9% Advanced Degree. Geographic distribution includes US (27% of participants), India (18%), UK (3.6%), Brazil (2.8%), and others (total of 152 countries).

The course covers the standard topics of a college introductory mechanics course with an emphasis on problem solving and concept interrelation. It consists of 12 required and 2 optional weekly units. A typical unit contains three sections: Instructional e-text/video pages (with interspersed concept questions, aka Checkpoints), homework, and quiz. Altogether there are 273 e-text pages, 69 videos, and about 1000 problems.

### 2.2 Data Mining

**Identifying Fake Learners.** We define ‘fake learners’ as users who use unauthorized methods to improve their grade in a way that does not rely on learning (or pre-knowledge). Currently, we have means to identify two types of such methods.

1. **Cheating Using Multiple Accounts:** This refers to users who maintain multiple accounts: A *master* account that receives credit, and *harvesting* accounts/s used to collect the correct answers (typically by relying on the fact that many questions provide the full answer, or at least True/False feedback, after exhausting the maximum number of attempts) [3, 18]. We note that in this method the multiple accounts are used by the same person. Using the algorithm described in [4, 18], we identified 65 (~13%) of the certificate earners who used this method. Hereafter we use the term CAMEO that was suggested by [15] for this phenomenon.

<sup>1</sup> <https://courses.edx.org/courses/MITx/8.MReVx/2T2014/course/>.

2. **Collaborators:** MOOC learners might work in study groups or with peers to submit assignments together. These associations are found using the algorithm described in [19] that relies on dissimilarity metrics and a data-driven method to find accounts that tend to submit their assignments in close proximity in time. Sometimes these associations represent real learning collaboration between peers taking a MOOC together and working towards a common goal, in other occasions they may represent more unethical and systematic dishonest behaviors, such as one learner passing the correct quiz responses to a friend every week. Overall, we identified 20 (~4%) of the certificate earners who submitted a significant portion of their assignments with peers. As there might be some overlapping between the detection of the two methods, we give the CAMEO algorithm priority as it represents a more specific behavioral pattern. Among the collaborators, 11 also used the CAMEO method. Hereafter we refer as ‘collaborators’ to the 9 accounts who were not CAMEO users.

### 3 Results

The results are organized as follows. First, we examined the differences between fake and true learners with respect to fundamental behavioral characteristics. Then, we examine the effect of these differences on correlations that seek to associate behavior and performance.

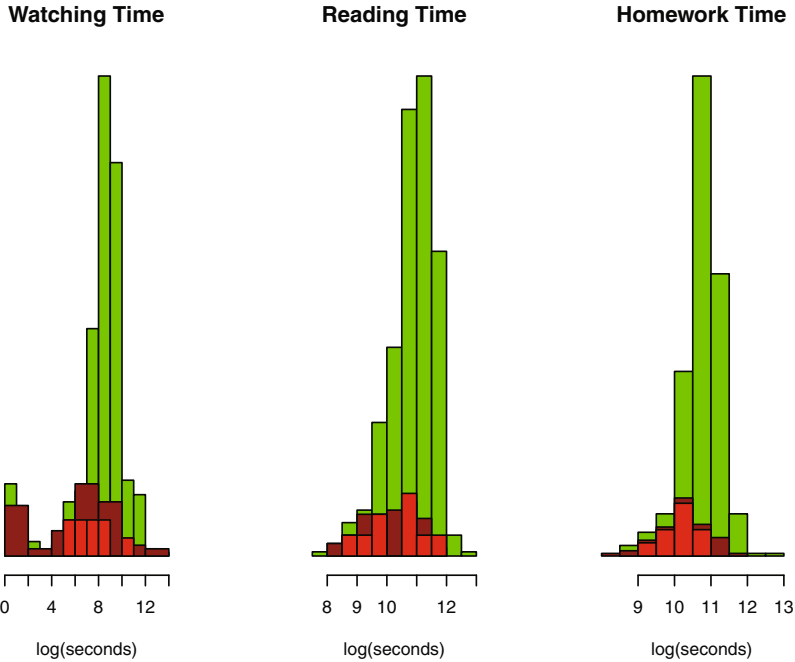
#### 3.1 Differences in Behavioral Characteristics

**Time on Course Resources.** The first measure that we examine is the amount of time that the fake learners spent on different course resources, compared to the true learners. We quantify *Reading Time* (time that the users spent on explanatory pages), *Watching Time* (time spent on videos), and *Time on Homework* (time spent in pages that contain homework items). Table 1 presents, per resource type, the mean time spent by fake/true learners, and *p-value* for the hypothesis that the fake learners spent less time on this type of resource.

**Table 1.** Time on resources.

Item Type	True learners	Fake Learners	<i>p-value</i>
Reading time	17.8	9.9	<0.001
Watching time	3.4	2.1	<0.1
Homework time	14.4	9.2	<0.001

From the table, it is quite clear that fake learners spent less time on the instructional resources. A more detailed illustration of the differences between the groups, also separating the fake learners into their subgroups, is presented in Fig. 1.



**Fig. 1.** Time on Resources: True learners in green; CAMEO in dark-red; Collaborators in red (Color figure online)

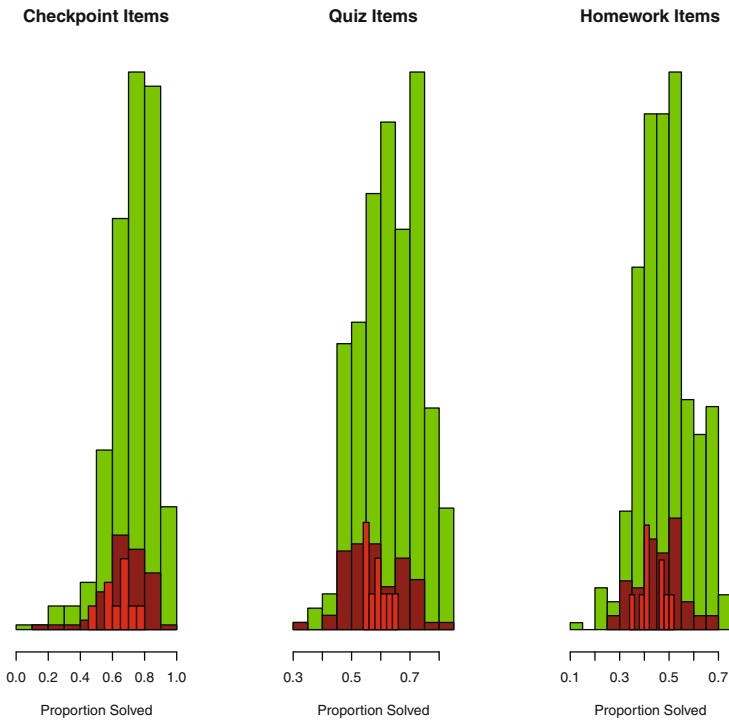
Overall, the behavior of the two subgroups within the fake learners cohort – cheaters and collaborators, is quite similar (confirmed with *t-test*).

**Proportion of Items Solved.** Next, we measure the proportion of assessment items that students attempted (either correct or incorrect). As explained in Subsect. 2.1, the course contains mainly three types of assessment items: Checkpoint, Homework, and Quiz. The reason for analyzing them in separate is that their different characteristics with respect to weight (points for solving them), and the easiness of getting the correct answer without effort (e.g., whether it is possible to receive the full answer after exhausting the possible attempts). Our assumption is that fake learners would factor that into their decision of whether to spend time on these items. For example, as checkpoint items have low weight, we assume that fake learners would show less interest in solving them. Quiz items have high weight, but are harder to cheat upon (no ‘show answer’, only True/False feedback). Homework offers relatively high weight and have ‘show answer’ enabled, which probably makes them ideal for fake learners (high ‘return on investment’). Table 2 contains the proportion of items solved by each group, and the *p-value* that the fake and true learners have a different distribution.

**Table 2.** Proportion of items solved.

Item Type	True learners	Fake learners	<i>p-value</i>
Quiz	0.63	0.58	<0.001
Homework	0.49	0.46	<0.01
Checkpoint	0.73	0.67	<0.01

Again, there is a clear difference between the groups, with fake learners trying less items. We also examine the distribution in more detail, and separate the fake learners into their two subgroups. This is shown in Fig. 2. Again, we do not see a significant difference between cheaters and collaborators in each of these metrics. In Sect. 4 we analyze these results and discuss the characteristics which make certain questions more attractive for fake learners.



**Fig. 2.** Proportion of Items Solved: True learners in green; CAMEO in dark-red; Collaborators in red (Color figure online)

**Performance Measures.** Student performance can be measured in various ways. We focus on the following metrics.

- Grade: Total points earned in the course (60 points is the threshold for certificate)
- Ability: Student’s skill in a 2PL Item-Response Theory (IRT) model, based on first attempt, with population containing the certificated users ( $N=502$ ), and items that were answered by at least 50% of these users. We chose IRT because students’ IRT ability scores are known to be independent of the problem sets each student tried to solve [9]. Missing items were imputed using a mean imputation. We used R’s *TAM* package<sup>2</sup>.
- Weekly Improvement: Per student, this is interpreted as the slope of the regression line fitted to the *weekly* IRT ability measures (e.g., fitting 2PL IRT on each week of the course in separate). One of the important issues that must be addressed during the calculation of the IRT slopes is to set up the common scale across weekly IRT scores. IRT is a latent variable model, and a latent variable does not have any inherent scale. Therefore, each IRT estimation defines its own scale for the latent variable. Equating is the process of transforming a set of scores from one scale to another. We used mean and sigma equating to set up a common scale across weekly IRT scores. The equated IRT slope captures *the change* in students’ relative performance during the course. For example, a student who has average performance in all the weeks, will have 0 relative improvement.
- Proportion Correct on First Attempt (CFA): The proportion of items, among the items that the student attempted, that were answered correctly on the first attempt.
- Mean Time to First Attempt (TTF): The average time it took a student between seeing the item (operationalized as entering into the page in which the item resides, or in case of multiple items in page, answering the previous item), and making the first attempt.
- Mean Time on Task (TOT): The average time the student spent on an item (e.g., sum of time for all attempts).

The mean values for these performance measures, and the *p-value* for the hypothesis that fake and true learners have different distribution, are presented in Table 3.

According to the table, fake learners are significantly faster (on both measures), but on the other metrics do not differ significantly from the true learners. However, it turns out that on these metrics there is a significant difference *within* the fake learners cohort, between the CAMEOers and the collaborators. This is demonstrated in Fig. 3. CAMEOers have higher grade (0.85 vs. 0.77), ability (0.21 vs.  $-0.66$ ), and CFA (0.79 vs. 0.67), than collaborators, all with significant *p-values*.

In fact, on ability and CFA, we get that CAMEOers  $>$  true learners  $>$  collaborators, with ability = (0.21,  $-0.07$ ,  $-0.66$ ), and CFA = (0.79, 0.76, 0.67),

<sup>2</sup> <https://cran.r-project.org/web/packages/TAM/TAM.pdf>.

**Table 3.** Performance of true and fake learners.

Measure	True learners	Fake learners	<i>p-value</i>
Grade	0.85	0.83	0.27
Ability	-0.07	0.1	0.23
Weekly Improvement	0.01	0.09	<0.05
Proportion CFA	0.76	0.77	0.42
Mean TTF	112s	72s	<0.001
Mean TOT	150s	97	<0.001

respectively. The *p-value* for these are borderline (<0.1 for CAMEOers vs. true learners, and < 0.2 for true learners vs. collaborators), but it demonstrates that on these metrics the fake learners have different behaviors, in which their average is quite similar to the average behavior of the true learners.

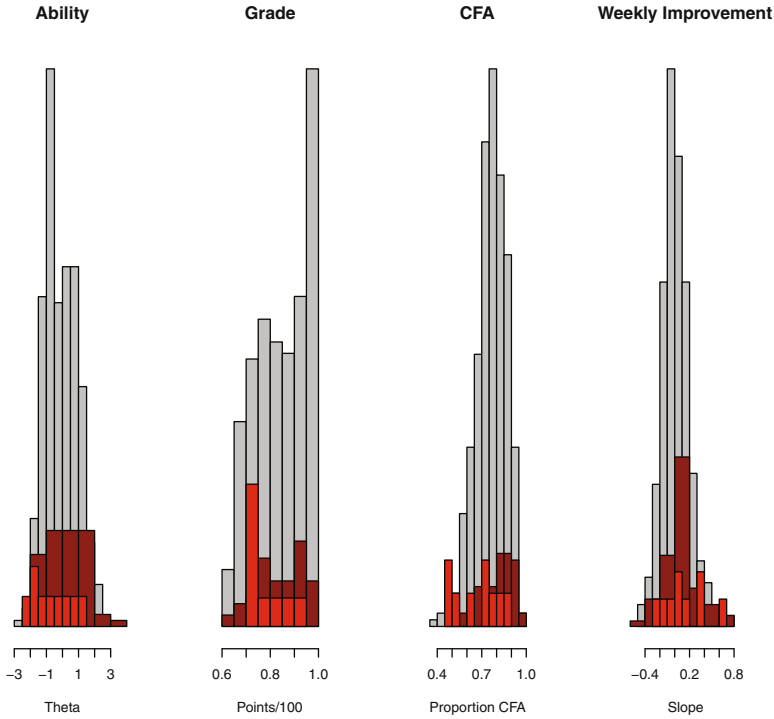
The fact that CAMEOers can have higher ability, yet the same grade, as true learners, is due to the nature of IRT, which weighs items according to their empirical behavior, and due to the fact that we train the IRT models on first attempts data. CAMEOers choose items strategically, and have very high proportion of CFA.

**Summary of Differences.** Overall, we see that fake learners spent much less time on course resources, and attempted less items. In the case of response time, we see that fake learners are much faster to solve exercises correctly. Regarding success metrics, we see that on average there is no significant difference between true and fake learners with respect to grade, ability, and CFA. However, a finer look into the subgroups reveals that CAMEOers have higher ability and CFA, and collaborators have lower ability and CFA, than true learners (though strictly speaking the *p-value* for this ordering is slightly above the 0.05 customary threshold).

### 3.2 Correlation Study

Next, we examine the effect of the differences in the behavioral metrics presented above on fundamental relationships – between response time and success, and between resource use and aggregated performance in the course.

**Response Time Vs. Success.** One of the issues of interest in education research is the relation between *response time*, and the likelihood of making a correct attempt. On one hand, better students might be faster (between-person differences), but on the other hand, spending more time on the question increases the probability of finding the correct solution (within-person effect) [11]. This is under the assumption that students try to learn. However, the performance of fake learners is affected by other factors. Figure 4 (left) shows the relation between *proportion CFA*, and *mean time to first attempt*, for the fake and true

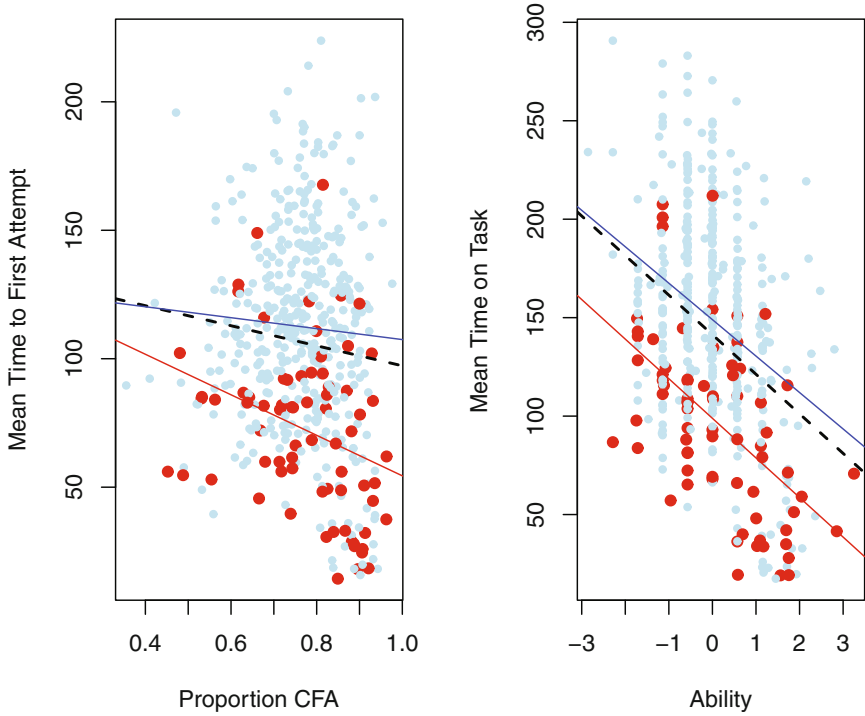


**Fig. 3.** Performance Measures: True learners in gray; CAMEO in dark-red; Collaborators in red (Color figure online)

learners (red and light-blue dots, respectively). The red, steep regression line is of fake learners; the blue, moderate line is of true learners; the dashed line is for the entire population. The difference between the blue and the dashed lines is how much the fake learners ‘pull’ the correlation down. Figure 4 (right) shows the same for the effect of fake learners on the relation between *IRT ability*, and *mean time on task* (the difference between *time on task* and *response time* is that the former is the time for all attempts, while the latter is only the time till the first attempt; most items in the course allow multiple attempts, and there is no penalty for using them).

As can be seen in both figures, the relationship between speed and performance is very different for fake and true learners, however the fake learners cohort is not big enough (about 15%) to change the overall trend dramatically.

**Resource Use and Aggregated Performance Measures.** The relation between the time students spend on different types of instructional materials, and their performance on various metrics, was studied in [7]. This is one of the first MOOCs EDM research papers, and it studied core questions related to the effectiveness of online learning materials. Here, we replicate the specific relationships studied in that paper, and how they change when removing fake learners.

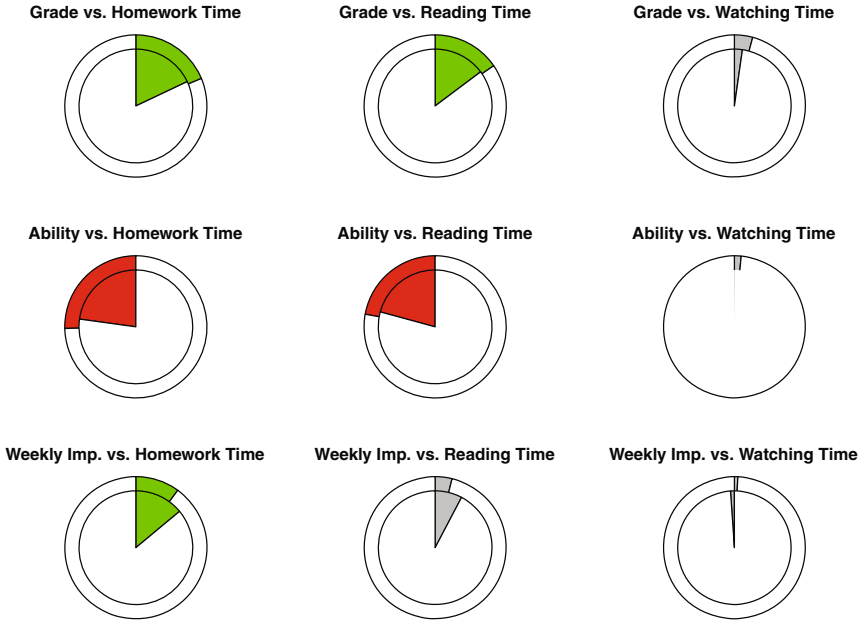


**Fig. 4.** The effect of fake learners on the relation between time and measures of skill

The results are presented in Fig. 5, which also adopts the visualization style used in [7]. It shows the relation between the amount of time spent on various course resources, and certain performance metrics. For each pie, the outer circle is the whole group, and the inner is the same measure, **after removing the fake learners from the data**. The angle of the piece is the size of the correlation. Clockwise angle represents positive correlation (colored with green), and counter clockwise represents negative correlation (colored in red). Gray color means  $p - value > 0.05$ . **The difference between the angle of the outer circle, and the angle of the inner one, is the effect of fake learners' data on the correlation.**

Let us examine the correlations with  $p - value < 0.05$  (colored with red/green). With respect to Grade vs. Homework and Reading Time, there is almost no effect (angle of inner and outer piece is almost identical). With respect to Ability vs. Homework and Reading Time, we see a *negative* correlation, which is *reduced* when removing fake learners. With respect to Weekly Improvement vs. Homework Time, we see a *positive* correlation, which *increases* when removing fake learners.





**Fig. 5.** Effect of fake learners on correlation between performance and time on course resources

## 4 Discussion

The findings reported in Sect. 3.1 indicate that fake learners tend to spend much less time than true learners on course materials, and to attempt fewer assessment items. Most likely, this is the result of being interested in easy ways to achieve a certificate. Since they have means other than learning to find correct answers, they can score well without spending a lot of time on the learning materials. Since the threshold for earning a certificate is 60% of the points, they can be selective with the items they choose to solve, and concentrate on ones they can solve more easily, either legitimately or not. This explains why they attempt less items, and also their *performance metrics* – why their *grade* is slightly lower, and why their *IRT ability* and *proportion CFA* are slightly higher, than the true learners (*grade* is very sensitive to the number of items solved, but *IRT* and *CFA* are basically not). Since they solve many of the questions using non-legitimate means, their response time is much faster than the other learners (very fast response time is a hallmark of cheating [17]).

Due to these differences, fake learners can bias various statistics and affect data-driven research results and decision-making processes. This depends not only on the behavioral differences, but also on the size of the cohort. In the course that we study, the population of the fake learners is roughly 15% of the certificate earners, which is a considerable sub-population.

Our results show mild effect on the strength of the relationships, so our conclusion is basically that the *correlation study* was robust against fake learners. This suggests that even a sub-population of  $\sim 15\%$  with very different characteristics is still not a threat to ‘average’ correlations. However, attempts to study selected groups like ‘efficient learners’ (e.g., learners who are fast and successful) would be very prone to distortions due to fake learners cohorts. We would also caution against doing expert-novice studies by classifying the very top students (containing a high percentage of fake learners, especially CAMEOers) as representative of ‘experts’. Also, we can expect that the percentage of fake learners, and subsequently their effect, may rise as the reward for good performance is raised (e.g. in getting a grade from a college) [17].

From a systematic point of view, using ‘black box’ computational models that rely on data requires taking proper steps to verify that the data are trustworthy. This was already acknowledged in other domains (and is considered a major challenge), but to date received only minor attention in the LA and EDM research community. We believe that verification, validation, and quality assurance of LA and EDM models and results should receive much more attention, and that this is an important part in the process of becoming a mature field. Our study makes an initial stride in this direction.

The main limitation of our research is that it is based on a single course and examines a limited set of learning analytics. Future research can examine a wider set of courses and challenge additional reported studies that could have been affected. Also, while the definition of ‘fake learners’ is broad and refers to various types of cynical learning behaviors, our results are based on the limited set of such behaviors that we currently know how to detect. We hope that future research will shed light on more types of ‘fake learning’ behaviors, and on ways to detect and prevent them.

## 5 Related Work

EDM and LA are emerging disciplines that aim to make sense of educational data in order to better understand teaching and learning, with the applied goal of improving the pedagogy of online learning environments, and developing ‘smart’ content and tools [10,20]. In particular, open learning environments such as MOOCs, where the large enrollment, wide scope (typically, a few weeks course), variety of learning materials, the relative freedom for learners to navigate, and the high-resolution data being collected, provide “unparalleled opportunities to perform data mining and learning experiments” [7] (pp. 1). A partial list of studies includes comparing active vs. passive learning [13], how students use videos [1,12], which resources are helpful [2,7,8,14], and many others.

The basic assumption behind most EDM/LA studies (though this assumption is typically not articulated), is that the data represent genuine learning behavior of individuals. This assumption is broken by fake learners, e.g., users who succeed in the course using means such as Cheating Using Multiple Accounts [3,4,15,18], or conducting some sort of collaboration [19]. In the context of Intelligent Tutoring Systems and K12 learners, Baker *et al.* [6] defined a related

phenomenon termed *gaming the system*, which they describe as “Attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning [...]”. This makes this behavior a sort of ‘fake learning’, however, *gaming the system* is not interpreted as illegitimate, and is more associated with frustration, lack of motivation, and inadequate design of the learning environment [5].

To the best of our knowledge, the influence of fake learners (and more generally, aberrant behavior) data on the reliability of models and results was not studied within the EDM/LA community. More generally, this issue can be seen as an instance of what Cathy O’Neil calls “Weapons of math destruction” [16]: Data-driven algorithms that make wrong decisions due to bugs, wrong assumptions on the data or the process that generated them, etc. An example within the context of education is the reported incident of a teacher who was fired because of a ‘performance assessment’ algorithm which yielded that her class did not improve enough during the school year<sup>3</sup>. She argued that the previous year’s test scores were artificially raised by cheating (possibly by a teacher who wanted to increase his/her evaluation). In social-media, the Facebook–Cambridge Analytica data scandal<sup>4</sup> demonstrates how fake accounts can be used to collect data and affect social trends.

## References

1. Alexandron, G., Keinan, G., Levy, B., Hershkovitz, S.: Evaluating the effectiveness of educational videos. In: EdMedia (2018) (To appear)
2. Alexandron, G., Pritchard, D.: Discovering the pedagogical resources that assist students in answering questions correctly a machine learning approach. In: Proceedings of the 8th International Conference on Educational Data Mining, pp. 520–523 (2015)
3. Alexandron, G., Ruiperez-Valiente, J.A., Pritchard, D.E.: Evidence of MOOC students using multiple accounts to harvest correct answers, learning with MOOCs II (2015)
4. Alexandron, G., Ruipérez-Valiente, J.A., Chen, Z., Muñoz-Merino, P.J., Pritchard, D.E.: Copying@Scale: using harvesting accounts for collecting correct answers in a MOOC. *Comput. Educ.* **108**, 96–114 (2017)
5. Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K.: Why students engage in “Gaming the System” behavior in interactive learning environments. *J. Interact. Learn. Res.* **19**(2), 162–182 (2008)
6. Baker, R.S.J.D., De Carvalho, A.M.J.B., Raspat, J., Alevan, V., Corbett, A.T., Koedinger, K.R.: Educational software features that encourage and discourage “gaming the system”. In: Proceedings of the 2009 Conference on Artificial Intelligence in Education, pp. 475–482 (2009)

<sup>3</sup> [https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/g1QAwwZpvR\\_story.html?utm\\_term=.21c5f0af7fd3](https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/g1QAwwZpvR_story.html?utm_term=.21c5f0af7fd3).

<sup>4</sup> [https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge\\_Analytica\\_data\\_scandal](https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal).

7. Champaign, J., Colvin, K.F., Liu, A., Fredericks, C., Seaton, D., Pritchard, D.E.: Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In: Proceedings of the first ACM conference on Learning @ scale conference - L@S 2014 (March), pp. 11–20 (2014)
8. Chen, Z., Chudzicki, C., Palumbo, D., Alexandron, G., Choi, Y.J., Zhou, Q., Pritchard, D.E.: Researching for better instructional methods using AB experiments in MOOCs: results and challenges. *Res. Pract. Technol. Enhanc. Learn.* **11**(1), 9 (2016)
9. De Ayala, R.: *The Theory and Practice of Item Response Theory. Methodology in the social sciences.* Guilford Publications, New York (2009)
10. U.S. Department of Education, O.o.E.T.: *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief* (2012)
11. Goldhammer, F.: Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimental control. *Measur. Interdisc. Res. Perspect.* **13**(3–4), 133–164 (2015)
12. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: *Understanding in-video dropouts and interaction peaks in online lecture videos* (2014)
13. Koedinger, K.R., Mclaughlin, E.A., Kim, J., Jia, J.Z., Bier, N.L.: *Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC*, pp. 111–120 (2015)
14. MacHardy, Z., Pardos, Z.A.: *Toward the evaluation of educational videos using Bayesian knowledge tracing and big data.* In: Proceedings of the Second (2015) ACM Conference on Learning @ Scale, L@S 2015, pp. 347–350. ACM (2015)
15. Northcutt, C.G., Ho, A.D., Chuang, I.L.: *Detecting and preventing “multiple-account” cheating in massive open online courses.* *Comput. Educ.* **100**(C), 71–80 (2016)
16. O’Neil, C.: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, New York (2016)
17. Palazzo, D.J., Lee, Y.J., Warnakulasooriya, R., Pritchard, D.E.: *Patterns, correlates, and reduction of homework copying.* *Phys. Rev. ST Phys. Educ. Res.* **6**, 010104 (2010)
18. Ruiperez-Valiente, J.A., Alexandron, G., Chen, Z., Pritchard, D.E.: *Using multiple accounts for harvesting solutions in MOOCs.* In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S 2016, pp. 63–70 (2016)
19. Ruipérez-Valiente, J.A., Joksimović, S., Kovanović, V., Gašević, D., Muñoz Merino, P.J., Delgado Kloos, C.: *A data-driven method for the detection of close submitters in online learning environments.* In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 361–368 (2017)
20. Siemens, G.: *Learning analytics: the emergence of a discipline.* *Am. Behav. Sci.* **10**, 1380–1400 (2013)



# Where Is the Learning in Learning Analytics?

## A Systematic Literature Review to Identify Measures of Affected Learning

Justian Knobbout<sup>(✉)</sup> and Esther van der Stappen<sup>ID</sup>

Research Centre for Learning and Innovation, HU University of Applied Sciences, P.O. Box 182, 3500 AD Utrecht, The Netherlands  
{justian.knobbout, esther.vanderstappen}@hu.nl

**Abstract.** Learning analytics is the analysis and visualization of student data with the purpose of improving education. Literature reporting on measures of the effects of data-driven pedagogical interventions on learning and the environment in which this takes place, allows us to assess in what way learning analytics actually improves learning. We conducted a systematic literature review aimed at identifying such measures of data-driven improvement. A review of 1034 papers yielded 38 key studies, which were thoroughly analyzed on aspects like objective, affected learning and their operationalization (measures). Based on prevalent learning theories, we synthesized a classification scheme comprised of four categories: learning process, student performance, learning environment, and departmental performance. Most of the analyzed studies relate to either student performance or learning process. Based on the results, we recommend to make deliberate decisions on the (multiple) aspects of learning one tries to improve by the application of learning analytics. Our classification scheme with examples of measures may help both academics and practitioners doing so, as it allows for structured positioning of learning analytics benefits.

**Keywords:** Learning analytics · Systematic literature review  
Data-driven intervention · Measures of affected learning · Enhanced learning

## 1 Introduction

Learning analytics make use of student data to improve learning and the environment in which this takes place. This improvement is achieved via data-driven interventions, which are an important step in the learning analytics process. Presently, much learning analytics activities are aimed at enhancing academic achievement [16]. Learning, however, is more than its mere outcome in the form of scores and grades. In this study, we research what other measures of affected learning can be identified in existing learning analytics literature. We conduct a systematic literature review and synthesize the results in order to provide an answer to the research question: *In what way does existing learning analytics literature measure affected learning?*

We structure the results of our study based on a classification scheme which is derived from prevalent learning theories. Our research supports both academics and

practitioners in their work as it provides (1) different types of affected learning which can be the target of learning analytics activities and (2) actual measures of these effects which help to determine the benefits of learning analytics on learning.

We structure the remainder of this paper as follows. First, we provide a short overview on the background of the study. We then describe in detail the methodology, followed by an elaboration on the results. Finally, we provide recommendations for future research and discuss the limitations of our study.

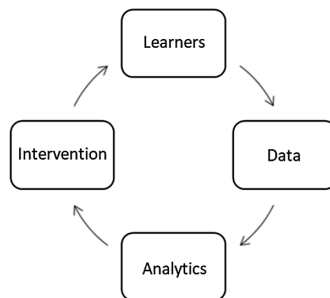
## 2 Background

In this section, we give an overview of learning analytics, its process and goals. Furthermore, we introduce a classification scheme to classify and analyze the key studies found during the literature review.

### 2.1 Learning Analytics

Learning analytics is “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which it occurs” [43]. We can achieve optimized learning in various ways, e.g., personalize learning, enhance instructor performance or improve curricula [33]. Learning analytics takes place at the micro and meso level within educational institutes, so the focus is on the learner and its surroundings [48]. Analytics at macro (or institutional) level is usually referred to as academic analytics.

The Learning Analytics Cycle [10] describes the process of turning data into action and involves four steps, which are: (1) students generate learner data, (2) the infrastructure captures, collects and stores the data, (3) the collected data is analyzed and visualized, and (4) the design and use of data-driven pedagogical interventions based on the analysis and visualizations (see Fig. 1). The cycle then starts again, enabling the measurement of effects caused by the performed interventions. This analysis, however, requires measures that allow for comparison of affected learning.



**Fig. 1.** Learning Analytics Cycle [10].

A systematic review of learning analytics literature by Papamitsiou and Economides [35] classifies studies by learning setting, analysis method, and research objectives. That study shows that learning analytics uses a wide variety of techniques and is not limited to only Learning Management Systems (LMSs), but can also be applied within other Virtual Learning Environments (VLEs), such as web-based education, social learning, and cognitive tutors. The objectives of the studies are diverse and include e.g., student behavior modelling, prediction of performance, prediction of dropout and retention, and increased (self-) reflection and (self-) awareness. These goals are achieved via pedagogical interventions. Interventions are an important part of the learning analytics process, since in this step, information is turned into action. Learning analytics interventions can be defined as “the surrounding frame of activity through which analytic tools, data, and reports are taken up and used” [52]. In our research, we analyze in what way the effects of interventions are measured by selecting key studies which report on empirical results, often from (quasi-)experimental settings and case studies applying data-driven pedagogical interventions. To categorize the various types of measures, we first synthesize a classification scheme from the extant literature.

## 2.2 Classification Scheme

To evaluate whether learning is indeed affected, we should be able to measure the effect of these interventions, by measuring the observed difference in learning. This raises the question in what way(s) learning can be measured. Below, we discuss several prevalent learning theories from this perspective.

Biggs’s 3P model [6] describes the educational system based on three factors: (1) presage factors which affect learning, (2) the learning process, and (3) the desired learning outcomes – see Fig. 2.

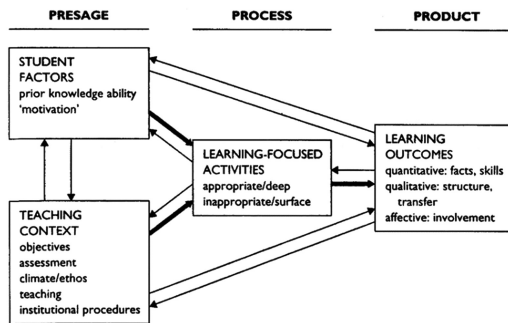


Fig. 2. 3P model [6].

Within *Presage*, Biggs presents a distinction between students and teaching context, which is also present in learning analytics literature. Siemens and Long [44] distinguish between learning analytics at course level and departmental level. Furthermore,

Van Barneveld et al. [48] propose a conceptual framework where learning analytics focusses on both learners and department. The latter is a broad concept, as it includes the context in which the learning takes place or, in other words, the ‘environment’ where the aforementioned learning analytics definition refers to. Departmental variables may consider a more long-term effect of learning analytics, which has been posed as an important feature of future learning analytics research [16].

Learning can either be described as a process or as the outcome of this process: a (relatively permanent) change in a person’s behavior, knowledge and/or skills [8]. As mentioned by Kolb [24]: “learning is best conceived as a process, not in terms of outcomes”. Learning outcomes – or ‘products’, to use the term provided by Biggs [6] – are often operationalized by performance indicators derived from assessment of learning, such as grades, degrees, and so on. A concept closely related to performance is (academic) achievement: performance outcomes that indicate the extent to which a person has accomplished specific goals that were the focus of activities in instructional environments, specifically in school, college, and university [46]. These ‘specific goals’ that students try to achieve are often formulated by the instructor as learning outcomes, which can be defined as a way “to express what the students are expected to achieve and how they are expected to demonstrate that achievement” [55]. In order to do so, learning outcomes should be clearly measurable – either direct or indirect – and used to gauge whether students can move to a higher level. Direct measures specifically assess learning as students must demonstrate this by performance of a task [38]. Indirect measures only give a general indication of learning and may include questionnaires and self-reports. Although grades may seem to be a direct measure, this is debatable. Grades can be regarded as a proxy for learning and therefore be an indirect measure, as they often comprise a combination of learning outcomes or included non-related corrections like extra credits for certain activities [14]. Therefore, as learning involves more than just a grade at the end of a course, we take a broader view at learning and include measures related to the process as well.

Based on the literature described above, we now discern two dimensions: (1) level of learning analytics and (2) learning as a (supported) process or learning as a result. Combining these two dimensions, we propose a classification scheme to classify learning analytics measures – see Fig. 3. We use this scheme to classify the measures of affected learning we find in our literature study.

	<b>Process</b>	<b>Performance</b>
<b>Student level</b>	Learning process	Student performance
<b>Departmental level</b>	Learning environment	Departmental performance

**Fig. 3.** Classification scheme for (measures of) learning analytics effects.



### 3 Method

In this section, we will provide a detailed description of the method used for our systematic literature review. The method applied in this literature review builds on other systematic literature reviews in the learning analytics domain (cf. [7, 33, 35, 39]). In our study, we aim at providing an answer to the following research question: *In what way does existing learning analytics literature measure affected learning?*

#### 3.1 Literature Sources

During the literature review, papers from seven different databases are sourced: (1) Learning Analytics and Knowledge (LAK) is the main conference in the learning analytics field. Organized for the first time in 2011, it produced an extensive amount of proceeding papers ever since. In this study, we include the LAK conference proceeding papers. (2) SpringerLink is the world's most comprehensive online collection of scientific, technological and medical journals, books and reference works, including the EC-TEL proceedings. (3) The Association for Computing Machinery (ACM) database is a large, comprehensive database focused on computing and information technology. (4) IEEE Xplore is technical-oriented database and contains papers related to, among others, computer science. (5) ScienceDirect is Elsevier's leading information solution for researchers and includes over 3,800 journals. (6) The Education Resources Information Center (ERIC) database is focused on educational literature and resources. (7) Learning Analytics Community Exchange (LACE) was a European Union funded project and one of the project aims was to collect evidence of the effects learning analytics have on education. In the study at hand, we include papers which relate to the proposition "Learning analytics improve learning outcomes".

#### 3.2 Search Terms

To search the aforementioned databases for literature related to measures of affected learning, different search terms are used. The search terms are formulated based on a priori analysis of relevant papers. Generally, the search includes the terms "learning analytics" AND student\* AND (achievement OR "student learning" OR "learning goal" OR "learning outcome" OR performance OR "student success"). When allowed for by the search engine, we specifically search the abstracts for student\* and ("learning analytics") to ensure we get learning analytics-related articles.

#### 3.3 Selection of Papers and Inclusion Criteria

The aim of this study is to identify measurable effects of data-driven interventions in real-life educational settings. It is a first step to identify what types of measures are currently used for effects of learning analytics endeavors. These insights will allow the learning analytics community to develop (possibly standardized) instruments to measure these effects, in turn creating opportunities for the replication or reproduction of results and performing meta-analyses on effect sizes. We therefore focus on studies reporting on quantitative results, as they provide us with actual measures of learning

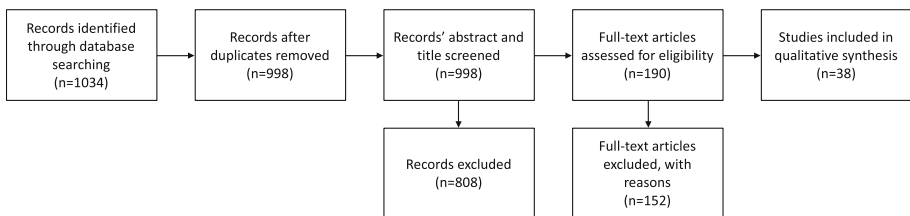
which can be calculated and can ultimately be applied in a standardized way. The following inclusion criteria are used during our search process:

- Paper is written in English;
- Paper must either be a conference proceeding paper or journal paper;
- Paper is published between 2011 and July 2017;
- Paper must describe interventions performed based on data analysis;
- Paper must present empirical data;
- Paper must report on quantitative results.

From the papers found in the previous step, the title and abstract are read to determine whether it meets the inclusion criteria. Papers clearly not meeting the criteria are dismissed. If the abstract and title do not provide enough information to make the selection, the paper is scanned – especially the method and result section – to make a better-informed decision. In a second selection round, the remaining papers are entirely read and again gauged against our inclusion criteria. To ensure the objectivity of the selection, a subset of the retrieved articles was handled separately by a second researcher and the results were discussed. No conflicts were observed in the selection of key studies by the two researchers. The key studies are all included in the analysis phase of the review. From these papers, we extracted and collected: author(s); title and subtitle; year; research objectives; level of analytics (descriptive, predictive or prescriptive); measure or indicator of improved learning; and operationalization of these measures and indicators. We analyzed these data to synthesize the results presented in the next section.

## 4 Results

This section presents the results of our literature review. From the 1034 hits on the search terms in the seven databases, 38 key studies meet the inclusion criteria – see Fig. 4. A retention of around 4% sounds rigid, however, other literature reviews in the learning analytics domain like Bodily and Verbert [7] and Ruiz-Calleja et al. [39] show similar results of 10% and 3%, respectively.



**Fig. 4.** Search process results.

#### 4.1 Classifying Key Studies Based on Affected Learning

Using the classification scheme introduced in Sect. 2.2, we now classify the key studies based on the different measures of affected learning.

**Learning Process.** The learning process relates to learning-focused activities. Learning analytics key studies within this category try to affect different tasks which can be distinguished during this process like the planning of coursework [21], supporting self-regulated learning [32, 34, 41, 42], time management skills [47], discussion board posts quantity and quality [4], engagement with assignment [28], number of readings [29], plagiaristic behaviors [1], and choosing to solve more difficult questions [11]. One of the major objectives of the key studies in this category is increase of (self) reflection and (self) awareness. By providing students with the right visualizations, they can take control of their own learning, thereby improving the learning process.

**Student Performance.** Containing 19 key studies, this category is by far the largest in our research. Most studies relate to academic performance, achievement, grades or scores [9, 11, 12, 15, 17, 22, 23, 25, 36, 37, 52, 54, 55]. Other mentioned forms of affected learning in this category are learning gains [40], content mastery [27], students predicting their own final scores [2] or the quality of a written computer program [5]. Remarkably, some of the key studies claim to affect aspects which one would expect in the learning process category – e.g., supporting self-regulated learning [31], time management skills [47] – but the effects that are measured fall in the student performance category (e.g., grades or scores). That is, the product or outcome of the learning process is measured rather than the actions performed during this learning process. Objectives of the key studies in this category include the increase of (self) reflection and (self) awareness, prediction of performance, recommendation of resources, and student behavior modeling.

**Learning Environment.** Although the optimization of the learning environment is explicitly mentioned in the commonly accepted definition of learning analytics [43], with only five key studies this category is the smallest within our research. The learning environment is affected by providing teachers with tools to intervene on problematic groups [49, 50], assessment time savings [18], improvement of course quality and outcomes [45], and teachers attention [30]. The sole objective of studies in this category is the improvement of assessment and feedback services.

**Departmental Performance.** Instead of focusing on individual students, departmental performance mostly relate to the success of students as a group [3, 12, 17, 22, 26], to student retention [13, 20], or to financial benefits of Early Warning Systems [19]. The prediction of performance, dropout and retention are the most common objectives of the key studies in this category.

#### 4.2 Measures of Affected Learning

The previous paragraph describes the aspects of learning which learning analytics literature aims to affect. We regard these aspects as the dependent variables of these studies. The operationalization of the dependent variables are measures of affected

learning and can be used to describe changes caused by learning analytics. We use our classification scheme to give an overview of the measures used in the key studies (see Table 1).

**Table 1.** Measures of affected learning.

Learning process	Counts of events, centrality measures	Siadaty et al. [42]
	Making predictions about grades by students	Holman et al. [21]
	Number of posts, discourse features	Beheshitha et al. [4]
	Plagiarised post ratios	Akçapınar [1]
	Pre- and/or post-questionnaire scores	Melero et al. [32] Siadaty et al. [41]
	Revision of artefact made	Manske and Hoppe [28]
	Social Network Analysis (SNA) indicators	Marcos-García et al. [29]
	Study time	Tabuenca et al. [47]
	Time spent on solving questions, higher level of difficulty of questions	David et al. [11]
	Use of metacognitive tools, application of self-regulated-learning (SRL) cycle (plan, learn, assess, reflect)	Nussbaumer et al. [34]
Student performance	Answers to reference questions	Papoušek et al. [36]
	Depth, rarity, quality, and specificity of program	Berland et al. [5]
	Difference between pre- and post-test	Perikos et al. [37] Sharma et al. [40]
	Grades	Grann and Bushway [17] Diana et al. [15] Jayaprakash et al. [22] Khan and Pardo [23] Whitelock et al. [51] Tabuenca et al. [47] Kumar et al. [25] McKenzie et al. [31]
	Test scores	McKenzie et al. [31] Cheng and Liao [9] Ben David et al. [11] Yamada et al. [54] Xiong et al. [53]
	Mastery scores	Lonn et al. [27]
	Score of the game	Arguedaset al. [2]
	Learning environment	Detection of problematic student groups
Number of messages sent by teacher		Van Leeuwen et al. [49] Van Leeuwen et al. [50]
Teacher attention, teacher interaction		Martinez-Maldonado et al. [30]
Time it takes a teacher to assess a student		Groba et al. [18]
Validity, reliability of exam		Smolin and Butakov [45]

(continued)

**Table 1.** (continued)

Departmental performance	Grades	Lauría et al. [26] Davis et al. [12] Arnold and Pistilli [3]
	Course completion rates	Davis et al. [12] Herodotou et al. [20]
	Withdrawal rates	Lauría et al. [26] Jayaprakash et al. [22] Arnold and Pistilli [3]
	Reregistration rate	Grann and Bushway [17]
	Revenue from student enrollment	Harrison et al. [19]
	Student retention	Dawson et al. [13]

## 5 Conclusions and Discussion

The aim of this study was to provide an answer to the research question: *In what way does existing learning analytics literature measure affected learning?* The first conclusion is that, from 1034 articles on learning analytics, only 38 describe quantitative, measurable effects of complete learning analytics cycles in education. This is a noticeable shortcoming, since studies in which both qualitative and quantitative results were present also satisfied our inclusion criteria. By analyzing these 38 key studies, we identified different measures of learning which can be affected with learning analytics. The measures are positioned according to a classification scheme: learning process, student performance, learning environment, and departmental performance. Our study allows for improved positioning of learning analytics research based on concrete measures, which helps learning analytics research and endeavors to be better compared. This systematic literature review shows that key studies mostly relate to the categories *student performance* and *learning process*. This was to be expected, as learning analytics particularly aims at learners and learning at the micro level. Only four papers report on measures in more than one category [11, 17, 22, 47], even though cross-categorical learning analytics provide a better, multi-perspective view on learning as it includes both process and performance or multi-level measures.

### 5.1 Recommendations

In order to justify the use of data analytics within educational processes, the effects of learning analytics on learning must be clear and well-defined. Some of the analyzed papers do report on potential improvements gained by data-driven intervention but do not describe their actual effect in terms of measures of affected learning. By describing those effects, more evidence about the benefits of learning analytics on education can be gathered, consequently strengthening the field in general. We suggest the use of our research outcomes for reporting on and comparing learning analytics results in both research and practice.

Gašević et al. [16] urge us to remember that “learning analytics are about learning”. In line with this statement, and based on the outcomes of this study, we recommend

learning analytics researchers and educational institutes to move away from mere performance-based evaluation of learning analytics projects and include measures related to learning processes and learning environment as well, as that is also a core objective of learning analytics [43]. Moreover, by optimizing the learning environment, learners are provided with better and more prompt feedback, while instructors can make more accurate decisions. Regardless of the dominant learning theory within an institute, a more complete view on learning is taken by looking at measures from multiple categories of our classification scheme.

## 5.2 Limitations and Future Work

Our goal was to identify measures of affected learning and group these based on a classification scheme. In order to do so, we only included empirical, quantitative results from data-driven interventions in our study. However, several studies use tools, techniques or methods as an intervention, even though they do not rely on data analytics itself. These papers then use data to describe the effect the intervention has on learning. Although this provides insight in the variables used to measure affected learning, these studies were disregarded as they do not meet our inclusion criterion demanding data-driven interventions, which is an important step within the learning analytics process. Future research might adopt broader inclusion criteria and extend the current findings with a larger set of key studies, thereby enhancing our results and identifying more and different measures of affected learning.

Finally, this study revealed that in recent learning analytics literature, no default set of constructs exists from which the dependent variable for a study can be selected. In Table 1, we see several different terms for closely related concepts, while the operationalizations also differ between studies. Building on the classification scheme in Fig. 3, a next step would be to devise an ontology of constructs - with operationalizations in the form of measures or instruments - for learning analytics benefits, in order to facilitate the reproducibility of empirical learning analytics research.

## References

1. Akçapınar, G.: How automated feedback through text mining changes plagiaristic behavior in online assignments. *Comput. Educ.* **87**, 123–130 (2015)
2. Arguedas, M., Daradoumis, T., Xhafa, F.: Analyzing the effects of emotion management on time and self-management in computer-based learning. *Comput. Hum. Behav.* **63**, 517–529 (2016)
3. Arnold, K.E., Pistilli, M.D.: Course signals at purdue: using learning analytics to increase student success. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 267–270. ACM (2012)
4. Beheshitha, S.S., Hatala, M., Gašević, D., Joksimović, S.: The role of achievement goal orientations when studying effect of learning analytics visualizations. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 54–63. ACM (2016)

5. Berland, M., Davis, D., Smith, C.P.: AMOEBA: designing for collaboration in computer science classrooms through live learning analytics. *Int. J. Comput. Support. Collab. Learn.* **10**(4), 425–447 (2015)
6. Biggs, J.B., Telfer, R.: *The Process of Learning*. McGraw-Hill/Appleton & Lange (1987)
7. Bodily, R., Verbert, K.: Trends and issues in student-facing learning analytics reporting systems research. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 309–318. ACM (2017)
8. Braungart, M., Braungart, R.: *Applying learning theories to healthcare practice* (2007)
9. Cheng, H., Liao, W.: Establishing an lifelong learning environment using IOT and learning analytics. In: *14th International Conference on Advanced Communication Technology (ICACT)*, pp. 1178–1183. IEEE (2012)
10. Clow, D.: The learning analytics cycle: closing the loop effectively. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 134–138. ACM (2012)
11. David, Y.B., Segal, A., Gal, Y.K.: Sequencing educational content in class rooms using Bayesian knowledge tracing. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 354–363. ACM (2016)
12. Davis, D., Jivet, I., Kizilcec, R.F., Chen, G., Hauff, C., Houben, G.: Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 454–463 (2017)
13. Dawson, S., Jovanovic, J., Gašević, D., Pardo, A.: From prediction to impact: evaluation of a learning analytics retention program. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 474–478. ACM (2017)
14. DePaul: Direct versus indirect assessment of student learning (n.d.). <https://resources.depaul.edu/teaching-commons/teaching-guides/feedback-grading/Pages/direct-assessment.aspx>. Accessed 28 Apr 2018
15. Diana, N., Eagle, M., Stamper, J.C., Grover, S., Bienkowski, M.A., Basu, S.: An instructor dashboard for real-time analytics in interactive programming assignments. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 272–279 (2017)
16. Gašević, D., Dawson, S., Siemens, G.: Let’s not forget: learning analytics are about learning. *TechTrends* **59**(1), 64–71 (2015)
17. Grann, J., Bushway, D.: Competency map: visualizing student learning to promote student success. In: *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, pp. 168–172. ACM (2014)
18. Groba, A.R., Barreiros, B.V., Lama, M., Gewerc, A., Mucientes, M.: Using a learning analytics tool for evaluation in self-regulated learning. In: *Frontiers in Education Conference (FIE)*, pp. 1–8. IEEE (2014)
19. Harrison, S., Villano, R., Lynch, G., Chen, G.: Measuring financial implications of an early alert system. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 241–248. ACM (2016)
20. Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., Hlosta, M., Naydenova, G.: Implementing predictive learning analytics on a large scale: the teacher’s perspective. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 267–271. ACM (2017)
21. Holman, C., Aguilar, S.J., Levick, A., Stern, J., Plummer, B., Fishman, B.: Planning for success: how students use a grade prediction tool to win their classes. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pp. 260–264. ACM (2015)

22. Jayaprakash, S.M., Moody, E.W., Lauría, E.J., Regan, J.R., Baron, J.D.: Early alert of academically at-risk students: an open source analytics initiative (2014)
23. Khan, I., Pardo, A.: Data2U: scalable real time student feedback in active learning environments. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 249–253. ACM (2016)
24. Kolb, D.: *Experiential Learning as the Science of Learning and Development*. Prentice Hall, Englewood Cliffs (1984)
25. Kumar, V., Boulanger, D., Seanosky, J., Panneerselvam, K., Somasundaram, T.S.: Competence analytics. *J. Comput. Educ.* **1**(4), 251–270 (2014)
26. Lauría, E.J., Moody, E.W., Jayaprakash, S.M., Jonnalagadda, N., Baron, J.D.: Open academic analytics initiative: initial research findings. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 150–154. ACM (2013)
27. Lonn, S., Aguilar, S.J., Teasley, S.D.: Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Comput. Hum. Behav.* **47**, 90–97 (2015)
28. Manske, S., Hoppe, H.U.: The “Concept Cloud”: supporting collaborative knowledge construction based on semantic extraction from learner-generated artefacts. In: IEEE 16th International Conference on Advanced Learning Technologies (ICALT), pp. 302–306. IEEE (2016)
29. Marcos-García, J., Martínez-Monés, A., Dimitriadis, Y.: DESPRO: a method based on roles to provide collaboration analysis support adapted to the participants in CSCL situations. *Comput. Educ.* **82**, 335–353 (2015)
30. Martínez-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)
31. McKenzie, W.A., Perini, E., Rohlf, V., Toukhsati, S., Conduit, R., Sanson, G.: A blended learning lecture delivery model for large and diverse undergraduate cohorts. *Comput. Educ.* **64**, 116–126 (2013). <https://doi.org/10.1016/j.compedu.2013.01.009>
32. Melero, J., Hernández-Leo, D., Sun, J., Santos, P., Blat, J.: How was the activity? A visualization support for a case of location-based learning design. *Br. J. Edu. Technol.* **46**(2), 317–329 (2015)
33. Nunn, S., Avella, J.T., Kanai, T., Kebritchi, M.: Learning analytics methods, benefits, and challenges in higher education: a systematic literature review. *Online Learn.* **20**, 2 (2016)
34. Nussbaumer, A., Hillemann, E., Gütl, C., Albert, D.: A competence-based service for supporting self-regulated learning in virtual environments. *J. Learn. Anal.* **2**(1), 101–133 (2015)
35. Papamitsiou, Z.K., Economides, A.A.: Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* **17**(4), 49–64 (2014)
36. 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)
37. Perikos, I., Grivokostopoulou, F., Hatzilygeroudis, I.: Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic. *Int. J. Artif. Intell. Educ.* **27**(3), 475–514 (2017)
38. Rome, M.: Best practices in student learning and assessment: creating and implementing effective assessment for NYU schools, departments and programs (2011)



39. Ruiz-Calleja, A., Prieto, L.P., Ley, T., Rodríguez-Triana, M.J., Dennerlein, S.: Learning analytics for professional and workplace learning: a literature review. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 164–178. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_13](https://doi.org/10.1007/978-3-319-66610-5_13)
40. Sharma, K., Alavi, H.S., Jermann, P., Dillenbourg, P.: A gaze-based learning analytics model: in-video visual feedback to improve learner’s attention in MOOCs. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 417–421. ACM (2016)
41. Siadaty, M., Gašević, D., Hatala, M.: Associations between technological scaffolding and micro-level processes of self-regulated learning: a workplace study. *Comput. Hum. Behav.* **55**, 1007–1019 (2016)
42. Siadaty, M., Gašević, D., Hatala, M.: Measuring the impact of technological scaffolding interventions on micro-level processes of self-regulated workplace learning. *Comput. Hum. Behav.* **59**, 469–482 (2016). <https://doi.org/10.1016/j.chb.2016.02.025>
43. Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S.P., Shum, S., Ferguson, R., Duval, E., Verbert, K., Baker, R.: Open learning analytics: an integrated & modularized platform (2011)
44. Siemens, G., Long, P.: Penetrating the fog: analytics in learning and education. *EDUCAUSE Rev.* **46**(5), 30 (2011)
45. Smolin, D., Butakov, S.: Applying artificial intelligence to the educational data: an example of syllabus quality analysis. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 164–169. ACM (2012)
46. Steinmayr, R., Meißner, A., Weidinger, A.F., Wirthwein, L.: Academic achievement. Oxford Bibliographies (2014). <https://doi.org/10.1093/obo/9780199756810-0108>
47. Tabuenca, B., Kalz, M., Drachler, H., Specht, M.: Time will tell: the role of mobile learning analytics in self-regulated learning. *Comput. Educ.* **89**, 53–74 (2015)
48. Van Barneveld, A., Arnold, K.E., Campbell, J.P.: Analytics in higher education: establishing a common language. *EDUCAUSE Learn. Initiative* **1**, 1–11 (2012)
49. Van Leeuwen, A., Janssen, J., Erkens, G., Brekelmans, M.: Teacher regulation of cognitive activities during student collaboration: effects of learning analytics. *Comput. Educ.* **90**, 80–94 (2015)
50. Van Leeuwen, A., Janssen, J., Erkens, G., Brekelmans, M.: Supporting teachers in guiding collaborating students: effects of learning analytics in CSCL. *Comput. Educ.* **79**, 28–39 (2014)
51. Whitelock, D., Twiner, A., Richardson, J.T., Field, D., Pulman, S.: OpenEssayist: a supply and demand learning analytics tool for drafting academic essays. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, pp. 208–212. ACM (2015)
52. Wise, A.F., Zhao, Y., Hausknecht, S.N.: Learning analytics for online discussions: embedded and extracted approaches. *J. Learn. Anal.* **1**(2), 48–71 (2014)
53. Xiong, X., Wang, Y., Beck, J.B.: Improving students’ long-term retention performance: a study on personalized retention schedules. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, pp. 325–329. ACM (2015)
54. Yamada, M., Kitamura, S., Matsukawa, H., Misono, T., Kitani, N., Yamauchi, Y.: Collaborative filtering for expansion of learner’s background knowledge in online language learning: does “top-down” processing improve vocabulary proficiency? *Educ. Technol. Res. Dev.* **62**(5), 529–553 (2014)
55. Yassine, S., Kadry, S., Sicilia, M.: A framework for learning analytics in moodle for assessing course outcomes. In: IEEE Global Engineering Education Conference (EDU-CON), pp. 261–266. IEEE (2016)



# Can I Have a Mooc2Go, Please? On the Viability of Mobile vs. Stationary Learning

Yue Zhao<sup>1</sup>(✉), Tarmo Robal<sup>2</sup>, Christoph Lofi<sup>1</sup>, and Claudia Hauff<sup>1</sup>

<sup>1</sup> Delft University of Technology, Delft, The Netherlands  
{y.zhao-1,c.lofi,c.hauff}@tudelft.nl

<sup>2</sup> Tallinn University of Technology, Tallinn, Estonia  
tarmo.robal@ttu.ee

**Abstract.** The use of mobile technology has become an ubiquitous part of our daily lives and enables us to perform tasks on-the-go and anytime that once were possible only on stationary devices. This shift has also affected the way we learn. The use of mobile devices for learning on-the-go requires users to multitask and divide attention between several activities, at least one of which (the learning activity) with high cognitive load. Massive Open Online Courses (MOOCs) have become a popular way for people around the world to learn outside of the traditional and formal classroom setting. While most MOOC platforms today offer specific apps to learn via mobile devices, the learning situation and its effect on learners while using mobile devices on-the-go has not been studied in full. In contrast to most existing mobile learning studies which were conducted in the lab, we focus on real-life situations commonly experienced by learners while they learn on-the-go. In a study with 36 participants and four mini-MOOCs deployed on edX, we investigate the differences in MOOC learners' performance and interactions in two different learning situations with mobile devices (stationary learning and learning on-the-go) and under two environmental variables (daylight and crowdedness).

**Keywords:** Mobile learning · MOOCs · Divided attention

## 1 Introduction

With the rapid advancement of mobile technology, the use of mobile devices has become ubiquitous around the world—about 98% of the population in developed countries, and 50% of the population in developing countries had mobile-broadband subscriptions in 2017 [18]. This development has affected the way people exploit mobile technology to learn new skills—a significant number of people use mobile devices for learning. A 2012 survey on lifelong learning by Tabuenca et al. [23] found that 56% of learners used their smartphone on a daily basis, whilst a study on mobile language learning by Dingler et al. [5] in 2017 reported that about 38% of learning sessions took place while in transit.

According to O'Malley et al. [14], mobile learning refers to “*any sort of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies.*”

The start of the MOOC movement in 2011 vastly widened the learning opportunities for people across the world outside of a formal education setting. While in the early years MOOC platforms lacked support for mobile devices, by 2015, most well-known platforms (such as edX, Coursera and Udacity) offered a mobile learning experience [12], either in the form of responsive web pages or native mobile apps (for Android and iOS), thus further expanding the possibilities to learn anywhere and anytime.

Critical for mobile learning [16, 19–21] is the *learning situation*—a set of environmental and intentional constraints [2]—in which learning occurs. A learner’s available time, the employed device type(s), and the frequency of interventions or distractions are only a few of those constraints that affect learning. One common learning situation for MOOC learners is *stationary learning*: here, learners use a device with a large screen to access course materials whilst being stationary in a comfortable environment (e.g. at their desk), enabling them to focus on the learning activity. In the mobile learning situation<sup>1</sup>, the conditions are quite different—mobile devices have considerably smaller screens and they are used in various and possibly changing environments which require learners to multitask (e.g., learning whilst walking or transiting). In terms of learning, this situation results in an increase in interruptions and distractions [19], an increase in cognitive load [3, 5, 23], and increased frustration [4].

Existing works on mobile learning in MOOCs focus on the design and delivery of course content for mobile devices [12, 17] as well as the learning experience on mobile devices [4, 15, 24, 24]; the latter though is typically studied *in the lab*, instead of real (urban) environments. Thus, little is known about how multitasking and a multitude of overlapping *real-life conditions* affect MOOC learning on-the-go compared to stationary learning. This knowledge gap serves as the core motivation for our work.

More specifically, we focus on the impact of the learning situation on learners’ performance and interactions, the effect of different environmental variables on the learning on-the-go process, and the correlation between learners’ perceived workload and their performance/interactions. We analyzed the data we collected from a user study with 36 participants, each of whom completed two mini-MOOCs (one in stationary and one in the on-the-go condition<sup>2</sup> at specific times of the day to control for daylight and crowdedness), guided by the following research questions:

**RQ1:** To what extent does learning on-the-go (compared to stationary learning on a mobile device) affect MOOC learners’ learning gain, learning efficiency and interactions with the course content?

<sup>1</sup> In the remainder of this paper, we refer to learning in a non-stationary situation with a mobile device as *learning on-the-go*.

<sup>2</sup> In this condition our participants physically explored the university campus.

**RQ2:** How do learners perceive their workload (physical as well as mental) in the stationary and learning on-the-go conditions and how does it relate to their learning performance and interactions?

## 2 Background

Our research addresses the following aspects of online learning: multitasking and attention fragmentation, and the use of mobile devices in different learning situations, with a focus towards learning in MOOCs.

**Multitasking and Divided Attention.** Interacting with a mobile device while on-the-go requires the ability to multitask and divide one’s attention between several tasks efficiently at once. Multitasking—the act of attempting to engage simultaneously in two or more tasks that have independent goals [7]—is directly connected to our research on mobile learning from MOOCs.

Multitasking is tightly coupled with the attention level and situational awareness. Studies on walking and mobile use have highlighted the increase of cognitive load and a necessity to divide attention, thus forcing mobile users to correct their gait and walk slower while performing tasks on mobile devices [10, 11].

Multitasking also incurs a cost on performance and accuracy for other tasks as our ability to effectively process two or more attention-demanding tasks simultaneously is limited [7], and performance across two concurrent tasks is optimized based on perceived priorities [6]. Thus, switching between activity contexts (e.g. in the on-the-go setting switching between reading the slides, paying attention to the traffic, listening to the video lecture) lowers task effectiveness. Harvey and Pointon [9] investigated the effect of fragmented attention on mobile web search tasks in three different contexts (walking on a treadmill, navigating through an obstacle course, and sitting down) and found that the contextual situation affects user (search) task performance—walking affected participants’ objective and perceived search performance negatively. In addition, participants who performed searches while on the move reported a higher difficulty and cognitive workload in performing the tasks than those sitting. In MOOC learning, which requires a high degree of attention and commitment, this indicates a potential for less effective learning in the on-the-go condition compared to the stationary one. Xiao and Wang [24] investigated the impact of divided attention on the learning process and learning outcomes for mobile MOOCs, and proposed to detect divided attention via monitoring learners’ heart rate. In their study with 18 participants under lab conditions, they observed divided attention to hurt learners’ performance.

With respect to multitasking and fragmented attention our study explores the effect and extent learning on-the-go has on learners’ ability to comprehend course content, and on their cognitive learning performance.

**Mobile Learning.** Mobile learning (i.e. learning with a mobile device) stresses the possibility to learn across time and space, and commonly assumes that learners are on the move [21]. What mainly distinguishes mobile learning from traditional classroom learning is the variety and unpredictability of the situations in which learning can take place [19] which places different demands on learners' attention level, body posture, environment, and social context whilst learning.

Mobile technology has enabled context-sensitive learning and the use of sensor data of mobile devices to enrich the learning experience [20]. Dingler et al. [5] implemented an Android app to collect sensor data (e.g., location, ringer mode, motion) in order to detect learners' contexts and boredom levels during microlearning sessions on mobile devices. Based on a user study, the authors concluded that while on mobile and in transit people are more open to engage in quick learning sessions, and context information retrieved from phone sensors can be helpful for mobile learning.

Learning tasks that are cognitively demanding (e.g., reading and writing scientific essays) seem to be incompatible with the use of mobile phones while on-the-go, whereas activities that are less cognitively demanding (e.g., social networking, texting, taking pictures) are compatible with body movement [3]. Music et al. [13] attempted to detect changes in user attention by exploiting smartphone accelerometers to trace changes in user gait patterns as a response of interaction with a mobile device. In a traditional study setting (e.g. a library, classroom), the use of mobile phones whilst learning has been found to be a distraction for most learners [1]; the same can be said about the mobile MOOC setting as incoming notifications, messages, news, etc. can take learners' focus away from the actual learning task.

The mobile devices themselves also affect learners' perceptions. Dalipi et al. [4] studied learners' experience by comparing desktop and mobile platforms of three well-known MOOC environments (edX, Coursera, and Udacity). They found that learners were more satisfied with the respective desktop variants; mobile platforms with their small screens and a lack of external input devices caused negative emotions as a number of tasks, which were easy on the desktop variants, were rather difficult to execute on the mobile variants. In a similar vein, Becking et al. [2] argue that learning situations for learning on-the-go are uncomfortable because of the lack of space for taking notes, and the potential for interruptions.

In our study, we explore learning with a mobile device in two different settings: (i) on-the-go and (ii) in a seated and more convenient condition close to traditional online learning, yet with a mobile device. In the former condition, we do not confine our participants to the lab (e.g. by using a treadmill or an obstacle course), but instead ask them to physically explore the university campus whilst learning.

### 3 Study Design

#### 3.1 Learning Situations

Inspired by the mobile search study conducted by Harvey and Pointon [9] (who found walking to impact participants workload perception and search effectiveness), we investigate whether learning on-the-go has any measurable impact on learning gain, effectiveness and perceived workload compared to stationary learning in the MOOC setting. We consider the following two learning situations (or scenarios) in our user study:

**Stationary Scenario (StaSc):** Learners study MOOCs while sitting in the office with a mobile device. This scenario is used as the baseline in order to measure the impact moving around has on learning.

**Moving Scenario (MovSc):** Learners study MOOCs with a mobile device while on-the-go. Participants are asked to learn whilst walking from one building to another on campus at their normal walking speeds, while paying attention to the traffic.

To eliminate the effects of learning behaviors unrelated to the use of mobile devices (e.g., taking notes on a piece of paper) and of different types of mobile devices, we instructed our study participants to perform all learning tasks exclusively on the same mobile device<sup>3</sup> in both StaSc and MovSc. We hypothesized—in line with the findings in [24]—that compared to StaSc, the necessary multitasking and the possible interruptions and distractions in MovSc negatively affect MOOC learners’ learning gain. We also hypothesized that participants in MovSc require more time to consume the course materials (due to the divided attention) than those in StaSc. In line with the previous hypothesis, we anticipated participants in MovSc to revisit the video page more often and rewind the video more often than those in StaSc to refresh their memory (which was impaired due to the distractions on-the-go).

#### 3.2 Learning Materials

We prepared four mini-MOOCs on different topics (Table 1) for our user study and deployed them on edX Edge, a low-visibility clone of the edX platform.

All four mini-MOOCs have the same structure: one lecture video and 20 knowledge questions about the video content. To ensure similar difficulty across the four mini-MOOCs, we selected them from a pool of introductory MOOC video lectures produced by the Delft University of Technology for the edX platform. We chose those four based on their similar *amount of unfamiliar terminology* as labelled by three annotators with computer science degrees. Each question is a multiple-choice question (almost all with four answer options in addition to *I don’t know*), created by two of this paper’s authors. These questions are not

---

<sup>3</sup> A Samsung S5 smart-phone with 1080\*1920 pixels, 5.1” display screen, 2GB RAM, 2.50 GHz CPU, Google Android 6.0.1 and the Chrome browser installed.

only used in the mini-MOOCs (right after the video lecture) but also in the pre-study questionnaire, which enables us to compute the knowledge gain in a straight-forward manner. This setup also means that the questions cover key knowledge concepts discussed in the respective lecture, instead of specific video details (such as the number of instructors, or the color of the background). Each question can be attempted once in the pre-study questionnaire and MOOC.

The pre-study questionnaire thus contained  $4 \times 20 = 80$  questions about the four topics; we used the answers to those questions to select for each study participant the two mini-MOOCs with the *lowest* prior knowledge levels. This setup leads to large potential knowledge gains. Table 1 lists the pre-study knowledge scores for the four mini-MOOCs across our 36 participants. Note that the maximum obtainable score for the questionnaire was 20 for each topic. The Qubit topic proved to be the most difficult, with more than half of the participants answering 0 or 1 question correctly; in contrast, water quality aspects proved to be the easiest topic with half the participants answering between 7 and 11 questions correctly.

**Table 1.** Overview of our mini-MOOCs, the video length per MOOC and the minimum/median/maximum of participants’ prior knowledge test scores on the topics. The highest possible score per topic is 20.

Mini-MOOC	Video length	Pre-study scores		
		Min.	Median	Max.
Radioactive decay	6m53s	0.0	3.0	9.0
Qubit	12m24s	0.0	1.5	16.0
Water quality aspects	10m45s	1.0	7.0	11.0
Sedimentary rocks	5m03s	0.0	4.0	10.0

### 3.3 Environmental Conditions

In our study, next to stationary and on-the-go, we focus on the impact of two additional environmental variables—the *light condition* and the *crowdedness of the surrounding*. It is known that daylight can affect the visibility of the screen on mobile devices [25] and the visibility of the surroundings during learning. The crowded learning situation may lead to intensive interruptions and distractions in MovSc. We thus hypothesized daylight and crowdedness to lead to reduced learning gains. Note that these environmental conditions only apply to MovSc.

Study participants were randomly assigned to one of four groups based on the time of the experiments for MovSc: (i) 8:45 am (crowded time with daylight), (ii) 11:00 am (uncrowded, daylight), (iii) 5:45 pm (crowded, no daylight<sup>4</sup>), and (iv) 8:00 pm (uncrowded, no daylight). Table 2 shows the distribution of study participants across the four groups.

<sup>4</sup> We conducted this user study in December 2017 and January 2018 in Delft, the Netherlands.

**Table 2.** Number of participants under different experimental conditions.

Mini-MOOC	MovSc				StaSc
	Daylight & Crowded	Daylight & Uncrowded	Dark & Crowded	Dark & Uncrowded	
Radioactive decay	3	1	4	2	15
Qubit	3	5	3	4	13
Water quality aspects	0	2	0	0	2
Sedimentary rocks	2	0	3	4	6
Total	8	8	10	10	36

### 3.4 User Study Steps

In our experiments, each participant was guided through the following steps.

1. Pre-study questionnaire: 80 knowledge questions plus questions on demographics, experience with mobile devices, mobile learning and MOOCs;
2. In random order, complete **StaSc** and **MovSc** with the two mini-MOOCs that exhibited the lowest prior knowledge levels. During a mini-MOOC, participants were allowed to switch between the video and questions. Each of the two scenarios was assigned a 30 min time block.
3. Post-MOOC questionnaires: after each of the two scenarios a NASA TLX workload assessment form<sup>5</sup> [8] had to be completed. It assesses the workload during learning in each scenario on six aspects: mental demand, physical demand, temporal demand, performance, effort, and frustration.

### 3.5 Metrics

We now describe how we measured participants’ learning gain, learning efficiency and interactions. To measure the statistical significance of the difference between groups of learners, we employed the Mann-Whitney U test.

In our study we use *absolute learning gain (ALG)* and *realized potential learning (RPL)* to measure participants’ **learning gain** [22]. *ALG* refers to the number of questions that were answered *incorrectly* in the pre-study questionnaire and *correctly* in the mini-MOOC, normalized by the total number of questions (20). *RPL* refers to the *absolute learning gain* normalized by the maximum possible learning gain<sup>6</sup>.

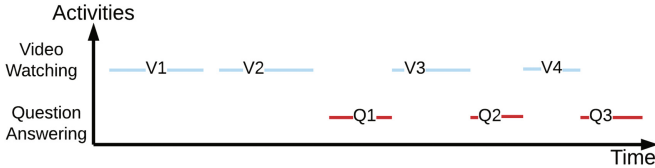
We measure **learning efficiency** through the efficiency of (i) course material consumption and (ii) learning gain. For the former, the time participants spend on watching videos (i.e., *video duration* and *normalized video duration*) and

<sup>5</sup> <http://www.nasatlx.com/>.

<sup>6</sup> For example, if in the pre-study questionnaire a learner answered 2 out of 20 questions correctly, the maximum possible learning gain is 18. If in the MOOC quiz two more questions are answered correctly, then *ALG* is  $\frac{2}{20}$  and *RPL* is  $\frac{2}{18}$ .



answering questions (i.e., *question duration*) are calculated—as we deploy our mini-MOOCs on edX Edge, we have access to all tracking data logged by edX. As shown in Fig. 1, *video duration* ( $VD$ ) refers to the minutes a participant spent watching the lecture video. *Normalized video duration* ( $NVD$ ) refers to  $VD$  normalized by the video length, which measures the proportion of the video consumed. *Question duration* ( $QD$ ) refers to the minutes a participant spent on the questions, including any time spent on video rewinding. To compute the **efficiency of the learning gain**, we divide  $RPL$  by  $VD$  and  $NVD$ .



**Fig. 1.** An example of a participant’s learning progress. In this example, *video duration* ( $VD$ ) is  $V_1 + V_2 + V_3 + V_4$ , *initial video watching duration* is  $V_1 + V_2$ , *video rewinding duration* ( $VRD$ ) is  $V_3 + V_4$ , *question duration* ( $QD$ ) is  $Q_1 + V_3 + Q_2 + V_4 + Q_3$ , and *question answering duration* is  $Q_1 + Q_2 + Q_3$ .

As **interactions** metrics we consider those that lead the participant away from the default mini-MOOC path (i.e. watch the video lecture and answer the 20 quiz questions). Specifically, we use the times participants revisit the video page during question answering (i.e., *#video page revisiting*,  $\#V\_revisit$  in short) and the minutes participants spent on video rewinding for questions (i.e., *video rewinding duration*,  $VRD$  in short) as metrics.

### 3.6 Study Participants

We recruited study participants from within TU Delft’s faculty of Electrical Engineering, Mathematics and Computer Science through flyers and mailing lists. 36 learners participated in our study: 9 women and 27 men. Their average age was 24.4 (std. dev. 2.7; min. 19; max. 30). Most participants were Master students, the highest educational degree (so far) was: high school (5 participants), Bachelor’s degree (21) and Master’s degree (10). On average, the participants had been using smart-phones for 7 years; all indicated to use them daily. 27 participants had used their mobile device for a learning activity within the last seven days before the user study. 26 participants had registered to at least one MOOC, 13 had made use of their mobile devices to learn in a MOOC and 11 participants had successfully completed at least one MOOC.

On average, each participant took about two hours to complete the entire experiment (recall, that each mini-MOOC was given a thirty minute time limit, however additional time was required for the pre-study questionnaire, switching scenarios, explanations by the experimenter, post-MOOC questionnaires and so on). Participants received a payment of €15. To motivate participants to learn, we provided a bonus payment of €5 for the participant achieving the highest learning gain overall.

## 4 Results

### 4.1 RQ1: Learning Gain, Efficiency and Interactions

In Table 3 (rows 1 & 2) we report our learning gain metrics across the two learning scenarios and the different environmental conditions, aggregated over all participants and topics. We find that, overall the learning gain achieved in the *MovSc* setting ( $ALG = 0.47$ ) is slightly lower than in *StaSc* ( $ALG = 0.5$ ). The difference is not significant though; similarly, the environmental conditions exhibit no consistent tendency. More concretely, as in our setup (20 questions per mini-MOOC), an  $ALG$  value of 0.05 represents one question answered correctly in the mini-MOOC but not the pre-study questionnaire, the recorded difference between *StaSc* and *MovSc* means that on average not quite one more question is answered correctly in the stationary learning scenario—this is in contrast to our hypotheses, where we expected to find considerable differences in learning gain across the two learning scenarios. The findings also hold for  $RPL$ ; here a value of 0.05 means that 5% of those questions not answered correctly in the pre-study questionnaire are answered correctly in the mini-MOOC.

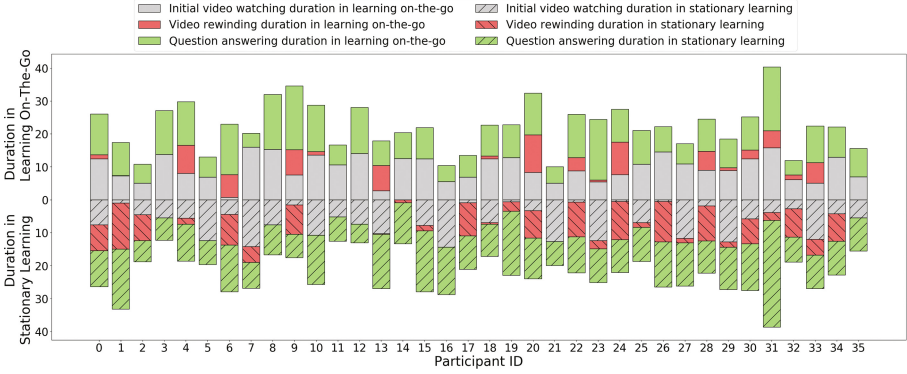
In terms of **learning efficiency**, the results in Table 3 (rows 3 to 7) show that in line with our hypotheses, participants in the *MovSc* scenario did take slightly more time to consume the lecture videos than those in the *StaSc* scenario. Importantly, participants spent significantly more time on questions in *StaSc* (on average 16 min) than in *MovSc* (13 min), a finding that corresponds to the results in [9] where stationary and on-the-go mobile web search tasks were compared. This result can be explained by the fact that a comfortable and stationary environment allows participants to engage with in-depth tasks requiring a lot of focus. Remember though, that this additional time spent on questions did not result in significantly higher learning gains as seen in our previous analyses. Once again, when considering the impact of the environmental variables, we do not observe a consistent trend, one way or another.

To determine the **efficiency of learning gain**, we measure how much participants learn from video watching. We hypothesized that *MovSc* has a negative impact on participants’ efficiency of learning gain.  $RPL/VD$  refers to participants’ learning gain *per minute of video watching*. We find that on average participants in *StaSc* reach a 40% higher efficiency (statistically significant) than in *MovSc*. We again did not observe clear trends for the different environmental variables.

When we consider **learners’ interactions** in Table 3 (rows 8 & 9) it is evident that on average participants in *StaSc* spend nearly twice as much time rewinding the videos than those in *MovSc*. The same trend holds for the number of times participants revisit the video playing page during question answering. Both of these findings indicate that in *StaSc* participants put more effort on finding relevant information for question answering than in *MovSc*. In order to understand participants’ interactions in more detail, in Fig. 2 we plot on a per-participant basis their (i) video watching duration before they start question

**Table 3.** The average value and standard deviation of metrics about participants' learning gain, learning efficiency and interactions under different experimental variables. † indicates significance at  $p < 0.1$  level. ‡ indicates significance at  $p < 0.05$  level. ◊ indicates significance at  $p < 0.01$  level.

Metrics	Learning Situation		MovSc with different environmental variables			
	StaSc (S)	MovSc (M)	Daylight & Crowded (DIC)	Daylight & Uncrowded (DIU)	Dark & Crowded (DkC)	Dark & Uncrowded (DkU)
ALG	0.504(±0.130)	0.474(±0.145)	0.463(±0.155)	0.463(±0.074)	0.480(±0.164)	0.485(±0.178)
RPL	0.575(±0.140)	0.533(±0.164)	$DkU†S†0.484(±0.161)$	0.550(±0.125)	0.536(±0.177)	0.554(±0.195)
VD (minutes)	10.796(±3.929)	11.883(±4.125)	10.881(±4.577)	$S†13.179(±1.937)$	11.068(±5.131)	12.463(±4.139)
NVD	1.304(±0.572)	1.407(±0.519)	$DkU†1.312(±0.468)$	$DkU†1.187(±0.189)$	1.457(±0.716)	$S†1.609(±0.486)$
QD (minutes)	16.284(±6.754)	$S†12.581(±6.323)$	$S†12.142(±6.983)$	13.913(±6.551)	$S†12.703(±6.833)$	$S†11.745(±5.916)$
RPL/VD	0.074(±0.080)	$S†0.053(±0.029)$	0.053(±0.029)	$S†0.043(±0.013)$	0.063(±0.040)	0.050(±0.026)
RPL/NVD	0.583(±0.531)	$S†0.419(±0.170)$	$S†DU†0.384(±0.126)$	0.475(±0.138)	0.459(±0.236)	$S†DU†0.363(±0.141)$
VRD (minutes)	4.515(±4.514)	$S◊2.284(±3.416)$	$S†2.102(±3.523)$	$S†2.048(±3.485)$	$S†2.698(±4.406)$	$S†2.203(±2.568)$
#V_Revisit	5.056(±5.270)	$S◊2.250(±2.708)$	$S†2.500(±3.546)$	$S◊1.125(±1.356)$	$S†2.700(±3.093)$	$S†2.500(±2.506)$



**Fig. 2.** The time participants spend on difference activities in StaSc and MovSc.

answering (i.e., *initial video watching duration*), (ii) their *video rewinding duration* during question answering and (iii) their time spent on question answering only (i.e., *question answering duration*).

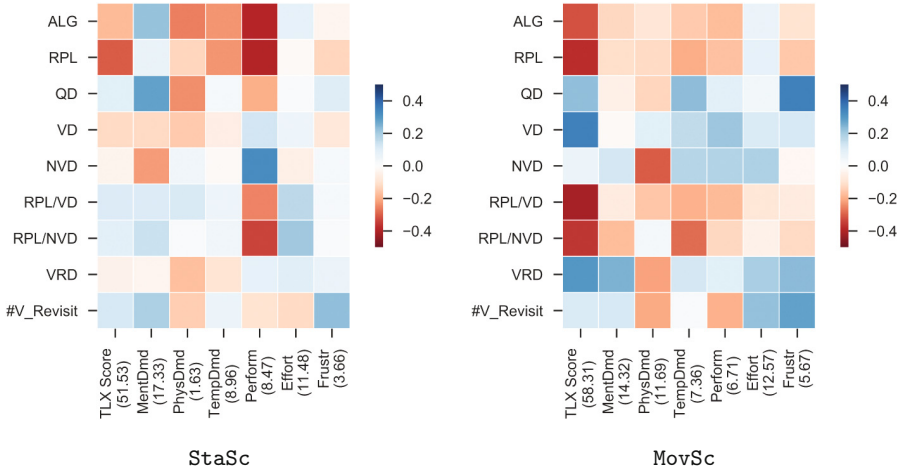
Compared to StaSc, it is evident that participants in the MovSc scenario tend to spend more time on video watching before they start question answering and less time on question answering. During question answering, most participants in MovSc revisited the video playing page fewer times and spent less time on video rewinding than in StaSc. This finding shows that participants in MovSc tend to switch less between the video playing page and the question page than those in StaSc. An explanation for the long question answering duration in StaSc can be that question answering is an activity with higher cognitive demand than video watching, which is not as compatible as video watching with walking with a mobile device [3].

### 4.2 RQ2: Learning and Perceived Workload

We now investigate the relationships between participants’ learning and their workload perception. Concretely, we report the Pearson correlation coefficient between our learning & interaction metrics and the six aspects of workload participants self-reported via the NASA TLX form. The results are shown in Fig. 3; here, *TLX score* is the overall score of workload, and *MentDmd*, *PhysDmd*, *TempDmd*, *Perform*, *Effort*, *Frustr* are participants’ workload scores on mental demand, physical demand, temporal demand, performance, effort, and frustration respectively.

When comparing StaSc and MovSc we observe sensible results with respect to mental demand and physical demands: in both scenarios the mental demand was found to be the most important one, followed by the physical demand in MovSc (in contrast to StaSc, where the physical demand received the lowest average weighting).

In StaSc we find performance (*How successful were you in accomplishing what you were asked to do?* with answer options ranging from *Poor* to *Good*)



**Fig. 3.** Linear correlation coefficient between participants’ learning performance, interactions and their perceived workload as measured through the NASA TLX form. The x-axis label also shows the average score of each workload dimension across our participants.

to be negatively correlated with learning gain, i.e. our participants were not able to estimate their own learning success very well. In contrast, performance is positively correlated with *normalized video duration*, indicating that participants estimated their learning performance to at least some extent based on how much of the video content they watched.

In the MovSc scenario, participants were also not able to self-estimate their learning gains (we found a slight negative correlation between *ALG/RPL* and performance); most interesting though is the positive correlation between *frustration* and question duration, i.e. the longer participants in the on-the-go condition spent answering questions, the more frustrated they felt (though overall frustration was not a major workload dimension).

## 5 Conclusions and Future Work

In this paper, we investigated to what extent learning on-the-go (compared to stationary learning on a mobile device) and its requirement for divided attention and multitasking affects MOOC learners’ learning gain, learning efficiency and interactions with course content. Our investigation included a foray into the influence environmental variables (light conditions and crowdedness) have on mobile learning. A second research question we considered is the relationship between learners’ perceived workload and their learning.

In order to explore these questions, we designed a user study with 36 participants; each participant “followed” two mini-MOOCs deployed on the edX Edge

platform: one in the on-the-go condition (learning on a mobile device while walking) and one in the stationary condition (learning on a mobile device while being stationary). We measured participants’ learning through a set of pre/post-study multiple choice question sets. Our analyses resulted in the following key findings:

- On average, learning on-the-go (*MovSc*) results in a lower ( $-6\%$  in *ALG*) learning gain than stationary learning (*StaSc*) with a mobile device.
- Compared to *MovSc*, *StaSc* participants spent 29% more time on answering questions and reached a 40% higher learning efficiency.
- When it comes to workload perception, participants in both conditions were not able to estimate their performance (wrt. learning gain) well; *MovSc* participants reported higher physical demands and slightly higher frustration than participants in the *StaSc* condition, though the differences in learning gains were small (first key finding).
- The environmental variables we investigated (daylight and crowdedness) did not have a consistent impact on any of the metrics investigated.

Our study has several limitations, among them the size of the user study (36 participants in total) which provides us with trends but few significant differences. A second limitation is the simplification of the on-the-go scenario to a walk on the campus (which does improve though—in terms of realism—on the lab conditions in prior studies). As pointed out by Becking et al. [2], the learning situation might be more complicated and unstable in many situations. Learners may walk, wait or take a bus or train while learning with a mobile device. Additionally, we only considered two environmental variables—the light condition and the crowdedness; other variables such as the weather and the temperature (recall that we conducted the experiments during December/January, i.e. the winter season in Europe) were not considered, although they are likely to also affect our participants’ behaviour. For example, two participants who were assigned the 8pm timeslots for the study told us that they aimed to finish their learning sessions as quickly as possible due to the bad weather. In the future to measure learners’ interactions in more complex learning situations, a dedicated mobile app may be needed to record fine-grained details of learners’ contexts and actions whilst on-the-go.

**Acknowledgements.** This research has been partially supported by the EU Widening Twinning project TUTORIAL, the Leiden-Delft-Erasmus Centre for Education & Learning and NWO project SearchX (639.022.722).

## References

1. Beasley, R.E., McMain, J.T., Millard, M.D., Pasley, D.A., Western, M.J.: The effects of college student smartphone use on academic distraction and dishonesty. *J. Comput. Sci. Coll.* **32**(1), 17–26 (2016)
2. Becking, D., et al.: Didactic profiling: supporting the mobile learner. In: *E-Learn 2004*, pp. 1760–1767 (2004)

3. Castellano, S., Arnedillo-Sánchez, I.: Sensorimotor distractions when learning with mobile phones on-the-move. In: 12th International Conference Mobile Learning 2016. International Association for Development of the Information Society (2016)
4. Dalipi, F., Imran, A.S., Idrizi, F., Aliu, H.: An analysis of learner experience with MOOCs in mobile and desktop learning environment. In: Kantola, J.I., Barath, T., Nazir, S., Andre, T. (eds.) *Advances in Human Factors, Business Management, Training and Education*. AISC, vol. 498, pp. 393–402. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-42070-7\\_36](https://doi.org/10.1007/978-3-319-42070-7_36)
5. Dingler, T., et al.: Language learning on-the-go: opportune moments and design of mobile microlearning sessions. In: *MobileHCI 2017*, pp. 28:1–28:12 (2017)
6. Farmer, G.D., Janssen, C.P., Nguyen, A.T., Brumby, D.P.: Dividing attention between tasks: testing whether explicit payoff functions elicit optimal dual-task performance. *Cogn. Sci.* **42**(3), 820–849 (2017)
7. Gazzaley, A., Rosen, L.D.: *The Distracted Mind: Ancient Brains in a High-tech World*. MIT Press, Cambridge, MA (2016)
8. 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)
9. Harvey, M., Pointon, M.: Searching on the go: the effects of fragmented attention on mobile web search tasks. In: *SIGIR*, pp. 155–164 (2017)
10. Krasovskiy, T., Weiss, P., Kizony, R.: Effect of aging, mixed reality and dual task prioritization on texting while walking. In: *ICVR 2017*, pp. 1–6 (2017)
11. Licence, S., Smith, R., McGuigan, M.P., Earnest, C.P.: Gait pattern alterations during walking, texting and walking and texting during cognitively distractive tasks while negotiating common pedestrian obstacles. *PLOS ONE* **10**(7), 1–11 (2015)
12. Marco, F.A., Penichet, V.M., Gallud, J.A.: What happens when students go offline in mobile devices? In: *MobileHCI 2015*, pp. 1199–1206 (2015)
13. Music, J., Stancic, I., Zanchi, V.: Is it possible to detect mobile phone user’s attention based on accelerometer measurement of gait pattern? In: *ISCC*, pp. 522–527 (2013)
14. O’Malley, C., et al.: *Guidelines for learning/teaching/tutoring in a mobile environment* (2005)
15. Pham, P., Wang, J.: Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In: *ICMI 2016*, pp. 37–44 (2016)
16. Rajasingham, L.: Will mobile learning bring a paradigm shift in higher education? *Educ. Res. Int.*, 1–10 (2011)
17. Renz, J., Staubitz, T., Meinel, C.: *Mooc to go*. International Association for Development of the Information Society (2014)
18. Sanou, B.: *ICT facts and figures 2017*. International Telecommunication Union (ITU) Fact Sheet (2017)
19. Sharples, M., Arnedillo-Sánchez, I., Milrad, M., Vavoula, G.: Mobile learning. In: Balacheff, N., Ludvigsen, S., de Jong, T., Lazonder, A., Barnes, S. (eds.) *Technology-enhanced Learning*, pp. 233–249. Springer, Dordrecht (2009). [https://doi.org/10.1007/978-1-4020-9827-7\\_14](https://doi.org/10.1007/978-1-4020-9827-7_14)
20. Sharples, M., Kloos, C.D., Dimitriadis, Y., Garlatti, S., Specht, M.: Mobile and accessible learning for MOOCs. *J. Interact. Media Educ.* **1**(4), 1–8 (2015)
21. Sharples, M., Taylor, J., Vavoula, G.: A theory of learning for the mobile age. In: *The SAGE Handbook of E-Learning Research*, pp. 221–247 (2007)
22. Syed, R., Collins-Thompson, K.: Retrieval algorithms optimized for human learning. In: *SIGIR 2017*, pp. 555–564 (2017)
23. Tabuenca, B., Ternier, S., Specht, M.: Everyday patterns in lifelong learners to build personal learning ecologies. In: *mLearn 2012*, pp. 86–93 (2012)

24. Xiao, X., Wang, J.: Understanding and detecting divided attention in mobile mooc learning. In: CHI 2017, pp. 2411–2415 (2017)
25. Xue, J., Chen, C.W.: Mobile video perception: new insights and adaptation strategies. *IEEE J. Sel. Top. Sig. Process.* **8**(3), 390–401 (2014)





# Validation of the Revised Self-regulated Online Learning Questionnaire

Renée S. Jansen<sup>(✉)</sup>, Anouschka van Leeuwen, Jeroen Janssen,  
and Liesbeth Kester

Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands  
r. s. jansen@uu.nl

**Abstract.** Self-regulated learning (SRL) is essential for students in online education to be successful. The Self-Regulated Online Learning Questionnaire was developed to measure SRL in online educational contexts. In this paper, a revised version of the questionnaire is presented and tested with three datasets. The scale ‘metacognitive skills’ is split into three subscales: metacognitive activities before, during, and after a learning task. Next to the three scales measuring metacognitive activity, the questionnaire contains scales measuring time management, environmental structuring, persistence, and help seeking. The revised questionnaire was found to have improved validity, usability, and reliability.

**Keywords:** Questionnaire · Online education · Blended learning  
MOOCs

## 1 Introduction

In online and blended learning, learners have more autonomy than in face-to-face education [1, 2]. This increase in autonomy makes it essential for learners to be actively involved in their own learning process, meaning that they self-regulate their learning [3, 4]. To accurately measure learners’ self-regulated learning in online education, the Self-regulated Online Learning Questionnaire (SOL-Q; [5]) was developed. While this questionnaire is a useful instrument, its validity, reliability, and usability could be improved. In the current paper, a revised version of the questionnaire is presented tested with three datasets.

### 1.1 Self-regulated Learning

Self-regulated learners are actively involved in their own learning process, not only during learning (performance phase), but also before (preparatory phase), and after learning (appraisal phase) [6, 7]. In the preparatory phase, learners think about what and how they will learn and the goals they have for the current learning session; they engage in (strategic) planning and goal setting. In the performance phase, learners engage in comprehension monitoring and strategy regulation. They furthermore manage their ‘resources’, including their time and study environment, as well as find

help when needed and persist when motivation drops. During the appraisal phase, learners reflect on their learning progress and their learning strategies [6].

## 1.2 Self-regulated Online Learning Questionnaire (SOL-Q)

To improve students' SRL in online education, it is important that students' SRL can be measured. The SOL-Q [5] was a first attempt at developing a questionnaire suitable to measure students' SRL in *online* learning environments. The developed questionnaire was based on several existing well-established SRL questionnaires (such as the Motivated Strategies for Learning Questionnaire; [8]): items from these questionnaires were selected and adapted to fit the context of online education. Based on exploratory and confirmatory factor analysis, an initial version of the SOL-Q was published. The SOL-Q consists of five scales: metacognitive skills (18 items,  $\alpha = .90$ , time management (3 items,  $\alpha = .71$ ), environmental structuring (5 items,  $\alpha = .67$ ), persistence (5 items,  $\alpha = .79$ ), and help seeking (5 items,  $\alpha = .83$ ).

## 1.3 Further Development of the SOL-Q

Although a satisfactory, initial version of the SOL-Q was created, the scale 'metacognitive skills' proved to be large and diverse. It consisted of items from a range of metacognitive self-regulation activities (e.g., goal setting, comprehension monitoring, reflection) and covering all SRL phases (preparatory, performance, and appraisal phase). The clustering of metacognitive items into a single metacognitive scale is not unexpected. In the SRL model presented by Zimmerman [9], significant correlations between the variables within a SRL phase are described, and Sitzmann and Ely [10] indeed found strong correlations between SRL constructs. While learners may not be able to distinguish among all the metacognitive activities, learners may be able to distinguish among the SRL *phases*. We therefore propose to split the scale 'metacognitive skills' into three separate subscales: activities before, during, and after a learning task. Not only would a separation into these three scales lead to an improvement of the face validity of the questionnaire, but it would also allow for more specific use of the questionnaire's (sub)scales, and for conclusions to be drawn about specific phases in the SRL process.

Based on the possible methodological and theoretical improvements on the scale 'metacognitive skills' outlined above, the aim of the current study is to create and test a revised version of the SOL-Q to improve its validity, reliability, and usability.

## 2 Method

### 2.1 SOL-Q Revised (SOL-Q-R)

The scale metacognitive skills within the SOL-Q was expanded and revised to generate three subscales. The existing 18 items in the scale were divided over the three subscales (i.e., before, during and after learning) based on the meaning of the item and on words signaling the timing of the activity. For instance, the item 'I am aware of what

strategies I use when I study for this online course' was placed into the subscale 'metacognitive activity during learning'. Second, the subscales were complemented to make sure all relevant aspects of metacognition were sufficiently present in each subscale. Strategic planning in the preparatory phase was not present in the existing items and only four appraisal items were present. Therefore, an item measuring strategic planning was added to the scale 'metacognitive activity before learning' ('At the start of a task I think about the study strategies I will use'), and two items measuring reflection on learning progress and learning strategies were added to the scale 'metacognitive activity after learning' ('After studying for this online course I reflect on what I have learned' and 'After learning for this online course, I think about the study strategies I used'). Specific attention was paid to words signaling timing when formulating the new items.

Furthermore, three small adaptations were made to improve the validity and reliability of the questionnaire. The first adaptation concerned the item 'I know what the instructor expects me to learn in this online course', originating from the Metacognitive Awareness Inventory scale for task definition [11]. Factor analyses during the development of the SOL-Q placed the item in the scale 'environmental structuring'. As the item does not measure environmental structuring, and is therefore also not conceptually similar to the other items in the scale, the item was removed from the questionnaire. Second, there were three negatively phrased items in the original design of the SOL-Q. These items were removed after factor analyses, as they did not fit the factor structure. Polar opposite items (i.e., 'I often feel so lazy or bored when I study for this online course, that I quit before I finish what I planned to do') are however known to result in lower internal-consistency reliabilities [12]. These three items, two in the persistence scale and one in the help-seeking scale, were rephrased to be polar positive and added to the SOL-Q-R. Finally, the time management scale was slightly adapted to improve its reliability as it was the scale with low reliability in the SOL-Q, which was likely due to the small size of the scale (3 items). Therefore, two items were added to the scale. The first item was already part of the originally developed questionnaire, but fell out during factor analyses. As the item conceptually fits in the scale, it was re-added ('I make good use of my study time for this online course'). The second item was formulated in line with the meaning of the scale ('I allocate studying time for this online course.').

The answering format was not changed for the SOL-Q-R. All questions had to be answered on a 7-point Likert scale ranging from 'not at all true for me' (= 1) to 'very true for me' (= 7). The full SOL-Q-R can be found at [SOONER.NU/SOL-Q-R](http://SOONER.NU/SOL-Q-R).

## 2.2 Participants and Procedure

The SOL-Q-R was administered to two groups of MOOC participants and one group of participants in a blended university course.

First, the questionnaire was implemented as a voluntary activity in a MOOC on Clinical Epidemiology offered by Utrecht University, The Netherlands, on Coursera. This MOOC consisted of 7 modules: an introductory module, 4 content modules, a module with a peer-graded assignment, and a module with a final exam. While students were free to decide on their own pace of studying, one module per week was

recommended. The questionnaire was added as a voluntary activity at the end of Module 2, to make sure students could reflect on their actual learning in the online course, and would not answer based on what they planned or expected to do. Complete data was gathered from 149 students. The responses of three students were considered outliers as they answered all questions identically (SD of their answers was 0). Responses of 146 students were used for analyses ( $M_{\text{age}} = 36.08$ , 48.6% male).

The questionnaire was also implemented as a voluntary activity in a MOOC on Environmental Sustainability offered by Wageningen University, The Netherlands, on edX. The MOOC consisted of seven modules: an introductory module and six content modules. In this MOOC, students were also free to study at their own pace, while one module per week was recommended. The questionnaire was added as a voluntary activity at the end of Module 2. Complete data was gathered from 73 students. Three students were considered outliers (SD = 0). Responses of 70 students were used for analyses ( $M_{\text{age}} = 39.67$  40.0% male).

The SOL-Q-R was also administered in a blended higher education course about designing educational materials at Utrecht University, the Netherlands. The course lasted 10 weeks, and followed a weekly structure of online preparation activities and face to face teacher-guided sessions (i.e., a flipped classroom design). In week 10, the students took an individual exam. The questionnaire was added as a voluntary online activity in week 4 of the course. Complete data was gathered from 94 students. One student was considered an outlier (SD = 0). Responses of 93 students were used for analyses ( $M_{\text{age}} = 23.59$ , 10.8% male).

### 2.3 Analyses

The SOL-Q and SOL-Q-R were compared based on reliability analyses. Furthermore, model fit was calculated using SPSS AMOS to test if the revised version had acceptable model fit. In line with the analyses done for the development of the SOL-Q [5], NC (normed Chi square) and RMSEA (root mean square error of approximation) were used as absolute fit statistics [13, 14].

## 3 Results

Reliability analyses were conducted to compare the internal-consistency reliabilities of the SOL-Q and the SOL-Q-R (Table 1). The results of the reliability analyses indicate higher reliabilities for the scales time management, environmental structuring, persistence, and help seeking in the SOL-Q-R. The reliability of the three metacognitive subscales are slightly lower than the reliability of the metacognitive skills scale. However, reliability is above .740 for all subscales, indicating good reliability.

An overview of the model fit statistics of the SOL-Q-R is presented in Table 2. Normed Chi square (NC) is a measure of  $\chi^2$  corrected for sample size, as  $\chi^2$  is known to be highly influenced by sample size [13]. Values of NC between 2.0 and 3.0 indicate acceptable fit and smaller values are better [13]. All tested models score below 2.0 thus indicating good fit of the SOL-Q-R in all three datasets. For RMSEA, smaller values indicate better fit and values below .08 are reasonable [15]. Based on the RMSEA

**Table 1.** Internal-consistency reliabilities of the SOL-Q and SOL-Q-R scales.

Scale	Items	1	2	3	Items	1	2	3
		$\alpha$	$\alpha$	$\alpha$		$\alpha$	$\alpha$	$\alpha$
Metacognitive skills	18	.93	.91	.88				
Activities before					7	.87	.84	.77
Activities during					7	.82	.78	.75
Activities after					6	.86	.86	.81
Time management	3	.57	.72	.71	5	.68	.72	.80
Environmental structuring	5	.78	.74	.66	4	.82	.77	.69
Persistence	5	.78	.70	.84	7	.82	.76	.88
Help seeking	5	.87	.91	.82	6	.88	.90	.84

*Note.* Dataset 1 = MOOC Clinical Epidemiology, 2 = MOOC Environmental Sustainability, and 3 = Flipped course educational materials.

statistic, the revised version shows adequate fit only in the first dataset, which is also largest. RMSEA is known to indicate poor model fit for small samples [16], which may explain the RMSEA values above .08 for dataset 2 and 3.

**Table 2.** Absolute model fit statistics of the SOL-Q-R.

	MOOC 1	MOOC 2	Blended
NC	1.797	1.700	1.713
RMSEA	.074	.101	.088

## 4 Discussion

In this paper, a revised version of the SOL-Q was presented and tested: the SOL-Q-R. The revised version has increased face validity, as the items within the scales were conceptually more similar. The separation of the large scale metacognitive skills into three smaller subscales (metacognitive activity before, during, and after learning) increases the usability of the questionnaire, as specific aspects of metacognition can be measured with the revised version. The theoretical and practical value of the questionnaire thus increases in the revised version. The results of the reliability analyses showed that the adaptations furthermore led to reliable scales overall (all  $\alpha$  above .67), with increased reliability for most scales. Model fit statistics are somewhat ambiguous, but provide no argument against acceptance of the SOL-Q-R. To conclude, the revised version of the SOL-Q is an improved version of the SOL-Q in terms of validity, reliability and usability and is therefore considered a valuable tool for researchers to measure students’ SRL in online education. The full SOL-Q-R can be found at [SOONER.NU/SOL-Q-R](http://SOONER.NU/SOL-Q-R).

## References

1. Garrison, D.R.: Self-directed learning and distance education. In: Moore, M.G., Anderson, W.G. (eds.) *Handbook of Distance Education*, pp. 161–168. Lawrence Erlbaum Associates, Mahwah (2003)
2. Wang, C.-H., Shannon, D.M., Ross, M.E.: Students' characteristics, self-regulated learning, technology self-efficacy, and course outcomes in online learning. *Distance Educ.* **34**, 302–323 (2013)
3. Kizilcec, R.F., Pérez-Sanagustín, M., Maldonado, J.J.: Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Comput. Educ.* **104**, 18–33 (2017)
4. Beishuizen, J., Steffens, K.: A conceptual framework for research on self-regulated learning. In: Carneiro, R., Lefrere, P., Steffens, K., Underwood, J. (eds.) *Self-Regulated Learning in Technology Enhanced Learning Environments*, pp. 3–19. Sense Publishers, Rotterdam (2011)
5. Jansen, R.S., van Leeuwen, A., Janssen, J., Kester, L., Kalz, M.: Validation of the self-regulated online learning questionnaire. *J. Comput. High. Educ.* **29**, 6–27 (2017)
6. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. *Theory Pract.* **41**, 64–70 (2002)
7. Puustinen, M., Pulkkinen, L.: Models of self-regulated learning: a review. *Scand. J. Educ. Res.* **45**, 269–286 (2001)
8. Pintrich, P.R., Smith, D.A.F., García, T., McKeachie, W.J.: *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning, Ann Arbor, MI (1991)
9. Zimmerman, B.J.: Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* **45**, 166–183 (2008)
10. Sitzmann, T., Ely, K.: A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. *Psychol. Bull.* **137**, 421–442 (2011)
11. Schraw, G., Dennison, R.S.: Assessing metacognitive awareness. *Contemp. Educ. Psychol.* **19**, 460–475 (1994)
12. Woods, C.M.: Careless responding to reverse-worded items: implications for confirmatory factor analysis. *J. Psychopathol. Behav. Assess.* **28**, 186–191 (2006)
13. Kline, R.B.: Details of path analysis. In: Kenny, D.A. (ed.) *Principles and Practice of Structural Equation Modeling*. The Guilford Press, New York (2005)
14. Hooper, D., Coughlan, J., Mullen, M.: Structural equation modelling: guidelines for determining model fit. *Electron. J. Bus. Res. Methods.* **6**, 53–60 (2008)
15. Gatignon, H.: Confirmatory factor analysis. In: Gatignon, H. (ed.) *Statistical Analysis of Management Data*, pp. 59–122. Springer, New York (2010)
16. Kenny, D.A., Kaniskan, B., McCoach, D.B.: The performance of RMSEA in models with small degrees of freedom. *Sociol. Methods Res.* **44**, 486–507 (2015)



# SRLx: A Personalized Learner Interface for MOOCs

Dan Davis<sup>1</sup>(✉), Vasileios Triglianios<sup>2</sup>, Claudia Hauff<sup>1</sup>, and Geert-Jan Houben<sup>1</sup>

<sup>1</sup> Web Information Systems, Delft University of Technology, Delft, The Netherlands  
{d.davis,c.hauff,g.j.p.m.houben}@tudelft.nl

<sup>2</sup> Faculty of Informatics, University of Lugano, Lugano, Switzerland  
triglv@usi.ch

**Abstract.** Past research in large-scale learning environments has found one of the most inhibiting factors to learners' success to be their inability to effectively self-regulate their learning efforts. In traditional small-scale learning environments, personalized feedback (on progress, content, behavior, etc.) has been found to be an effective solution to this issue, but it has not yet widely been evaluated at scale. In this paper we present the **Personalized SRL Support System (SRLx)**, an interactive widget that we designed and open-sourced to improve learners' self-regulated learning behavior in the Massive Open Online Course platform edX. **SRLx** enables learners to plan their learning on a weekly basis and view real-time feedback on the realization of those plans. We deployed **SRLx** in a renewable energies MOOC to more than 2,900 active learners and performed an exploratory analysis on our learners' SRL behavior.

**Keywords:** Learner modeling · Self-regulated learning  
Personalized learning

## 1 Introduction

Large-scale learning environments open up world-class educational resources to the masses. With this unprecedented scale and reach, however, come new challenges in enabling learners of diverse backgrounds to excel given the unfamiliar context of the massive online classroom. Low course completion rates—dropout rates of 95% are not uncommon [17]—highlight the need for additional support in MOOCs. Past research in this space, e.g. [12, 14, 15, 25] has explored the problems learners face when trying to succeed in these self-directed learning environments. Learners are often unable to find the time to keep up with a course, an issue related to insufficient self-regulatory abilities [12, 25]. *Self-regulated learning (SRL)* is the ability to plan, monitor, and actively control one's learning process. The discipline to plan and follow a self-imposed studying regime is a skill that is learned over time and associated with a higher likelihood of achieving self-set

---

D. Davis—The author's research is supported by the *Leiden-Delft-Erasmus Centre for Education and Learning*.

course goals in MOOCs [13, 19]. Learners who were exposed to such training during their studies tend to be more successful in MOOCs than learners without a tertiary education background. The latter though is a target population that is vital to keep the original vision of MOOCs alive: making higher education accessible to those that do not enter the traditional tertiary education system. Learners need tools that enable them to learn *how* to learn.

Today’s MOOC platforms (such as Coursera and edX) are not designed in a way that encourages learners to explicitly plan or monitor (with the help of feedback) their learning activities [7]. In general, learners are exposed to very few feedback moments to support their SRL processes.

Yeomans and Reich [24] found that a single planning prompt at the start of a MOOC can positively influence learning outcomes. We have expanded upon this concept by designing and developing the **Personalized SRL Support System**<sup>1</sup> (SRLx), an interactive widget for the edX platform that allows learners to explicitly express their motivation, *plan* their learning, *monitor* their progress towards their set goals at any point in time, and *reflect* on them. SRLx’s design was based on educational theories and findings in the SRL literature.

We deployed SRLx in a MOOC on renewable energies offered by the Delft University of Technology in 2017 with more than 2,900 active learners and empirically evaluate the following research questions:

- RQ1** To what extent do MOOC learners adopt and take advantage of a personalized SRL support tool?
- RQ2** Does SRLx support MOOC learners in promoting effective self-regulated learning behavior?

Along with the contribution of an open-sourced system architecture that provides SRL support at scale, we present the following key findings from our analysis of learners’ SRL behaviors:

- As the course progresses, learners are able to plan their time commitment more effectively.
- Learners are more conservative with the way they plan to commit time to the course compared to video and quiz activity planning.

## 2 Related Work

Zimmerman et al.’s model of self-regulated learning [27] comprises three cyclical phases: forethought, performance, and self-reflection. Learners first formulate a plan for their learning activities, they then carry out and act according to their plan, and finally they look back at their behavior and examine their strengths and areas for improvement. In this section we first examine self-regulated learning research in the classroom and then delve into SRL studies conducted within MOOCs.

---

<sup>1</sup> Open-sourced at <https://github.com/dan7davis/Lambda>.



**Self-regulated Learning in the Classroom.** Goal setting has been shown to be an important factor across all levels of education. Past research has investigated to what extent aspects such as *who* sets the goals, *when* are they set, *what* goals are set and *why* are those set influence the effectiveness of goal setting. While these studies have been conducted across a range of education levels, they have all taken place in the traditional classroom or lab setting.

Schippers et al. [23] showed that engaging and teaching undergraduate students about goal setting at the beginning of their studies has a positive impact across a prolonged period of time—after one year, a 98% reduction in the gender achievement gap and a 38% reduction in the ethnicity achievement gap was observed compared to the previous year’s cohort of students.

At the secondary education level, Zimmerman et al. [26] found that social-studies class students perform better (as measured by their final grade) when they set their own goals and benchmarks, than when having those imposed on them by teachers. Regularly reviewing and reflecting upon one’s study goals and behaviors was found by Sagotsky et al. [22] to be significantly more effective in terms of grades and study behavior than just setting goals in a user study with primary and middle school students. A similar result was found by Mahoney et al. [20] among 27 undergraduate students who were assigned to one of three experimental conditions while preparing for an exam: (i) continuous self-monitoring, (ii) intermittent self-monitoring, and (iii) receiving instructor feedback. In line with [22], students who performed self-monitoring exhibited higher levels of engagement and achievement than students who did not.

**Self-regulated Learning in MOOCs.** Due to the massive nature of MOOC platforms (supporting millions of learners), a large part of the platform development effort has to be spent on continued scalability. This leaves little time and attention for advances in platforms’ instructional designs. Prior research in the MOOC setting has so far focused on learner surveys (to elicit their SRL needs), pre-course SRL interventions, MOOC forum interventions, and the notion of learner feedback [4].

Nawrot and Doucet [21] and Hood et al. [9] surveyed MOOC learners about their experiences taking MOOCs. Proper time management was found to be a major hindrance for many MOOC learners [21]. The ability to self-regulate one’s learning was found to vary depending on learners’ professional backgrounds: higher-educated learners are better able to regulate their learning (including time management) than lower-educated learners [9].

Providing learners with visualizations of their progress enables them to *reflect* upon their learning, and an emerging body of research has begun to empirically evaluate the effectiveness of such feedback [1, 2, 6, 10]. Over time, this reflection should improve learners’ use of SRL strategies [3, 8]. One interesting finding by Kulkarni et al. [18] pertains to the timeliness of feedback and its impact on MOOC learners’ final grades: feedback (in this case on in-progress assignments) received within 24 h after assignment submission improves learning outcomes; if the feedback is delayed beyond this point, learners do not benefit from it.

According to Davis et al. [6], enabling learners to reflect weekly on their learning behavior in comparison to that of their successful peers (i.e. feedback through social comparison) led to a significant increase in passing rates among learners with high levels of prior education (Bachelor degree or higher). A drawback of this work is the need for a successful cohort to compare against and the fact that learners cannot establish their own plans and goals.

**Conclusions.** Goal setting and feedback are important techniques to improve learning outcomes in the traditional classroom. In the MOOC setting, SRL interventions have so far either been restricted to pre-course interventions or feedback. We here investigate the effect of regular planning and goal setting in the MOOC setting.

### 3 System Overview

We now first describe the client-side and server-side components of SRLx which allow for *real-time event tracking* and then turn to the design rationales behind the four front-end interfaces we developed (cf. Fig. 1).

**Client-Side.** The edX platform—on which we deployed SRLx—allows course designers to embed and execute custom HTML, CSS, and JavaScript code in edX pages, thus enabling the creation of customized interfaces and programming logic. We take advantage of this affordance and embed our client-side code in edX’s RAW HTML input elements.

We implemented two functionalities on the client-side: (i) the tracking and persisting of learners’ activities to the back-end such as quiz question submissions and video watch events (cf. Sect. 4 for an exhaustive list) via AJAX and (ii) the displaying of our front-ends for goal setting, planning & feedback and the persisting of learners’ interactions with them. We describe the activity tracking below and describe the interfaces in more detail at the end of this section.

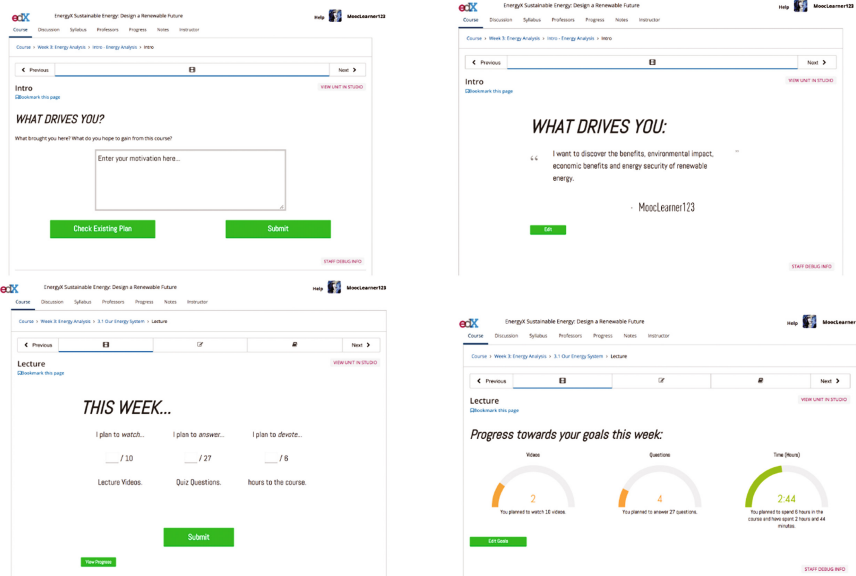
*Activity Tracking.* As SRLx provides real-time feedback based on learners’ actions on the edX platform, we had to track events such as quiz submissions and video watching events in real-time. The real-time constraint meant that we could not make use of edX’s default log data setup which distributes a MOOC’s daily logs in 24 h intervals. We therefore had to track these events ourselves as follows.

edX course components, such as videos or quizzes, are implemented via XBlocks, a component architecture based on Python, HTML, JavaScript and CSS. This allows anyone to create standalone hierarchical components that may include other XBlocks. To capture user interactions, Xblocks emit and subscribe to events using an event tracking library<sup>2</sup>. We enable real-time event tracking by using edX’s `Logger` object to subscribe to emitted events using the `listen(eventType, element, callback)` method: all Xblock fragments make

<sup>2</sup> <https://github.com/edx/event-tracking>.

use of the `Logger` object to emit events which are subsequently sent to the edX back-end via an `XMLHttpRequest`. We listen to all events of interest and forward those to our back-end.

**Back-End.** To store and retrieve learner data in real-time, we implemented an HTTPS server in Node.js and persisted the tracked events in a MongoDB database. The server uses a RESTful API to store and retrieve learner events. It supports the JSON format for both requests and responses. Along with logging edX’s learner behavior data, the SRLx server also logs all learner interactions with the SRLx interfaces.



**Fig. 1.** The four SRLx interfaces as they appear to learners on the edX platform: **motivation expression** (top-left), **motivation feedback** (top-right), **plan formulation** (bottom-left), and **plan feedback** (bottom-right).

**Front-End.** The three phases of Zimmerman’s model of self-regulated learning [27] (forethought, performance, and self-reflection) are integral to the design of SRLx’s four learner-facing interfaces shown in Fig. 1: motivation expression (forethought), motivation feedback (self-reflection), plan formulation (forethought), and plan feedback (performance and self-reflection). We now discuss them in turn.

*Motivation Expression.* This interface allows us to gain an understanding of learners’ motivations and overall forethought for their attitude towards the course. Modeled after the study planning system evaluated in [23], it is shown on the top-left of Fig. 1 and prompts learners to write about their motivation

and what brought them to the course in the first place. The key question asked to learners is *What drives you?* followed by other prompting questions to help learners express themselves: *What brought you here?* and *What do you hope to gain from this course?* Once learners have submitted their motivation it is persisted to our back-end. Learners can view and change their response any time.

*Motivation Expression Feedback.* In order to provide feedback and encourage a habit of self-reflection, we regularly make learners aware of their latest motivation response by displaying it back to them (top-right of Fig. 1) throughout each course week/unit. The response is shown as a quotation by the learner underneath the *What drives you:* text together with the learner’s edX username (to emphasize once more the source of the quotation).

*Plan Formulation.* This interface (Fig. 1 bottom-left) promotes forethought in prompting learners to formulate and state their plan for the coming course week in terms of engagement with course resources. Specifically, learners are prompted to enter the number of videos they intend to watch, quiz questions they intend to answer, and hours they intend to devote to the course this week. To aid learners in their planning, we provide the total number of videos and quizzes of the week (automatically extracted from the edX course pages) as well as the recommended time to spend in the course that week (as estimated by the course instructors).

*Plan Feedback.* To promote awareness learners’ performance and encourage self-reflection, the planning feedback interface (Fig. 1 bottom-right) consists of three gauges showing learners how well they have progressed towards the goals *they set for themselves*, removing all instructor influence. We designed the plan feedback as a data visualization dashboard that allows learners to easily draw their own insights about their progress. Previous research in data visualization for MOOC learners found that more abstract feedback (such as the “timeliness” of the quiz submissions) only benefited learners with a higher education background [6]. Since highly educated learners already have SRL abilities, we aimed to engage those learners that lack self-regulation skills and designed the interface to be clear and straight-forward to interpret.

## 4 Study Setup

**Participants.** We deployed SRLx in an edX MOOC on renewable energies offered by the Delft University of Technology. The course consists of 75 individual lecture videos and 295 graded quiz questions. A total of 8,057 learners enrolled in the course. The course started on August 29, 2017 and concluded on November 8, 2017. We made SRLx available to all learners but did not provide any additional incentive for using it.

Before the course, the learners were asked to self-report their basic demographic information. 5,349 learners at least partially complied. Of these learners, 25.3% are female; the learners’ median age is 26. We also collected information about their prior education level, as this has shown to have a significant

impact on learning outcomes and engagement with MOOCs [6]. As is common in MOOCs, we observe a great variety in this respect with learners running the gamut from high school to PhD levels of prior education: 1% had no prior formal education, 20% held at least a high school diploma, 5% an Associate’s degree, 45% a Bachelor’s degree, 26% a Master’s degree, and 3% a PhD. We consider learners’ prior education level to be *high* when they have earned at least a Bachelor’s degree, and *low* when they have not.

Given that many learners who enroll in a MOOC never enter the platform and log a session (a common occurrence in MOOCs), we narrow down the sample for analysis accordingly. Among all learners enrolled, 2,961 entered the course at least once and are therefore considered as *active learners* in our analyses.

**Measures.** To evaluate the role that SRLx plays in learners’ achievement and course engagement, we measure a number of in-course learning behaviors that are commonly used in MOOC studies as well as a number of novel measures enabled by SRLx:

- Average quiz score  $\in [0, 1]$  (proportion of attempted quiz questions answered correctly);
- Course activities:
  - Number of video interactions (play, pause, fast-forward, rewind, scrub);
  - Number of quiz submissions (submissions, correctness);
  - Number of discussion forum posts;
  - Time spent in the course;
- SRLx interactions:
  - Plan formulation (number of videos & quizzes and hours planned to spend in the course that week);
  - Motivation expression (submission text);
  - Editing (changing an established motivation or plan).

## 5 Results

In this section we analyze the deployment of SRLx along four lines: (i) course-level learning behaviors, (ii) study plan formulation tendencies, (iii) plan achievement rates, and (iv) motivations expressed over time.

### 5.1 Course-Level Learning Behaviours

In Table 1 we present summary statistics for overall course behavior among all active learners, characterized by having logged at least one session in the course. Table 2 shows the number of submissions made via SRLx.

Of the 2,961 active learners in the course, 872 (32%) engaged with SRLx at least one time (answering **RQ1**)—here characterized by having formulated at least one plan *or* submitting at least one motivation expression. While this rate of minimal engagement is substantially higher than past studies, e.g. [5], the true

**Table 1.** Overview of the average behavior of active learners. In rows 2 & 3 we partition the set of active learners into *Comply* (learners who formulated at least one plan and submitted at least one motivation expression) and *Non-Comply* (the remainder) learners.

Subset	N	Quiz Score	Session Count	SRLx Interact.	Feedback Checks	Quiz Submits	Videos Watched
Active	2,961	0.41	32.57	152.72	3.63	43.11	8.33
Comply	303	0.72	66.48	348.93	7.31	91.56	16.31
Non-Comply	2,658	0.37	28.71	130.35	3.21	37.58	7.42

rate of compliance (submitting *both* a plan and a motivation) is still very low, at 10% (303 out of 2,961 active learners).

While the top row in Table 1 represents all active learners in the course, the bottom two rows show the impact of self-selection in highlighting the difference in behavior between learners who did and did not engage with SRLx: on average, learners using SRLx (i.e. our Comply group) log more than twice as many sessions, answer nearly three times as many quizzes, answer more questions correctly and watch more than twice as many videos compared to learners in the Non-Comply group. We cannot claim that this difference is caused by the use of SRLx; rather it is at least partially a result of the self-selection of learners who would have been highly engaged and more successful in the course regardless.

However, this trend could also be partially explained by prior research on the *doer effect*, or the “...association between the number of online interactive practice activities students do and their learning outcomes” [16]. This theory states that engagement with interactive course components (such as SRLx, discussion fora, or quiz questions) has a stronger learning effect than passive activities such as reading or watching lecture videos. So while SRLx is unlikely to be the sole cause of the increase in activity between compliers and non-compliers, theory states that it likely contributed, at least in part, to the more positive learning outcomes of those who engaged with it.

When we split the engagement between the different types of interfaces (Table 2), we find that the plan formulation interface was considerably more engaging, with more than twice as many learners formulating plans (on average two plan formulations per learner) than writing up their motivation.

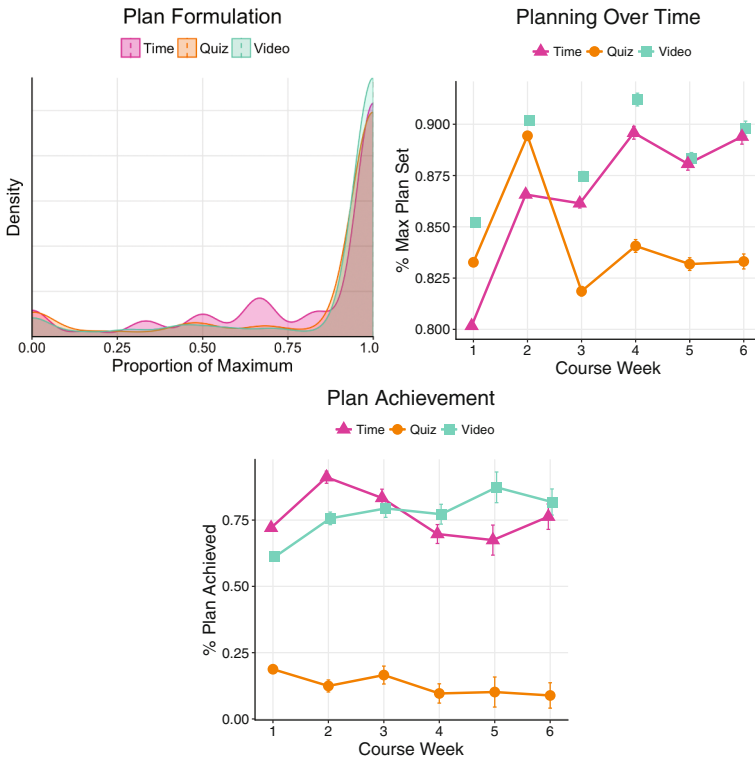
**Table 2.** Number of submissions of motivation expressions, plan formulations, and plan/expression edits. The bottom row shows the number of unique learners to have completed each action type.

	Motivation Expression	Plan Formulation	Edited
#Submissions	679	1,997	748
#Learners	396	971	338

### 5.2 Study Plan Formulation

In this analysis we focus on the plans the learners made using SRLx and thus address **RQ2**. We explore the following questions: are the learners overly ambitious with their plan formulation? Are learners able to consistently stick to their plans? Do their planning tendencies/strategies change over time? Figure 2 shows an aggregate view of all 1,997 plans submitted in the course.

Figure 2 (top left) shows the study planning behavior (in terms of time commitment, quiz submissions, and videos watched) of all learners who formulated and submitted at least one plan in SRLx. We find that the majority of plans set were for the maximum given the week’s content, i.e. most learners who submitted plans aimed at completing all quizzes, watching all videos and spending the instructor-suggested time on the course platform. At the same time in Fig. 2 (top left) we observe that the goals set pertaining to the proportion of time (from the recommended six hours per week) learners plan to commit to the course



**Fig. 2.** In clock-wise order: (i) the proportion of learners’ formulated plans set for the maximum possible value in the respective course week; (ii) the proportion of the maximum plan set by learners of each activity type over the span of all course weeks; (iii) plan achievement rates for each activity type by course week. Error bars show the standard error.

is lower than that of quiz submissions and videos. A Wilcoxon rank sum test with continuity correction ( $W = 2,210,200$ ,  $p < 0.0001$ ) indicates a significant difference between time plans ( $\bar{x} = 0.838$ ,  $\sigma = 0.34$ ) and video plans ( $\bar{x} = 0.88$ ,  $\sigma = 0.29$ ). From this analysis we conclude that learners are more conservative with the way they plan their time commitment to the course than the way they plan to engage with course materials.

To examine planning behavior at a more detailed level, in Fig. 2 (top right) we segment planning behavior by course week and illustrate the change over time. Compared to the rather steady rate of ambition (proportion of maximum plan set) with quiz plans (overall mean of 84.7% of the maximum), learners exhibited an overall trend of increasing their ambition each week for time- and video-related plans—a 9 % point increase from Week 1 to Week 6 for time plans (mean of 80% to 89%) and a 5 % point increase for video plans (mean of 85% to 90%). While these two increases can be attributed to less-ambitious learners dropping out of the course, the lower rate for quiz-related plans still holds throughout the entire course.

### 5.3 Plan Achievement

Figure 2 (bottom) shows the rate at which learners achieve each aspect of their plans each course week (**RQ2**). Whereas in the previous section we discussed how learners are conservative with their plan formulations as it pertains to time, we see in Fig. 2 (bottom) that learners are strong at achieving their plans for time commitment and video lecture viewing with high consistency across course weeks—an important insight given that poor time management has been identified by prior research [12, 13, 21, 25] as one of the primary causes of attrition in MOOCs.

It is also worth noting that the consistency and success of learners' time planning achievement is not a product of less ambitious goals being set. Refer back to Fig. 2 (top right) to see that the opposite is actually true; learners become *more* ambitious with their time plans as the course progresses, and learners are still able to achieve their plans with high consistency.

For the learners' video watching plan achievement, we observe a slight increase across the weeks with an overall mean of 63% completion. For learners' achievement of their quiz question-related plans, we observe substantially lower completion rates than those regarding time—falling from 19% in Week 1 to a mere 9% in Week 6.

We hypothesize that these results on plan achievement are a product of the difficulty of each activity type. Though not trivial, spending time in the platform requires little more than a learner's presence. Slightly more demanding is the activity type of watching lecture videos; and most challenging of all three is answering quiz questions, which is not only dependent on the previous two activities but also requires the application of newly-acquired knowledge. In other words, the rate by which learners complete their plans is commensurate with the exigency of the respective activity type.



As previous research on MOOC learners has identified achievement gaps among learners [11], we next conducted an exploratory analysis on plan completion per activity type as a function of a learner’s prior education level (with *high* education learners having earned at least a Bachelors degree, accounting for 75% of learners in the course). We observe no significant difference in plan completion rates in any of the three activity types according to a Wilcoxon rank sum test with continuity correction, thus indicating that learners are able to effectively use SRLx across a wide range of ability levels. This suggests that SRLx is equally usable and effective for learners of all prior education levels.

#### 5.4 Motivation Expression

Finally, we also conducted a preliminary analysis of the motivation texts our learners submitted. Among the 2,961 learners exposed to the SRLx interface, 396 submitted at least one motivation expression. These motivations range from learners working towards having better career opportunities to changing the world—the latter theme became markedly more prominent as the course progressed. The average word count is 23.9 (median 15, minimum 1, maximum 329). In Table 3 we randomly picked examples of *short* (at most ten words), *medium* length (up to 25 words) and *long* (26 words or more) submissions.

**Table 3.** Random sample of short, medium, and long submissions through the Motivation Expression interface.

S1	<i>Build up on sustainable energy knowledge</i>
S2	<i>I expect to get to know the future of energy</i>
M1	<i>I hope to learn more about sustainable ways of using and obtaining energy</i>
M2	<i>I want a clean planet I want to be responsible for that</i>
L1	<i>As a junior architect I am interested in learning more about the relationship between energy use and building design and how intelligent design can have positive impacts on building energy use as well as occupant health and happiness</i>

Replicating the methods in [24] applied to MOOC learner texts on course intentions, we evaluated the predictive value of the length of a learner’s text submission on their (i) current grade, (ii) average quiz question score, and (iii) total time spent in the course platform and were not able to find a significant effect in any of the metrics.

The ten most frequent terms occurring among all motivations are (in descending order): *energy*, *renewable*, *sustainable*, *knowledge*, *learn*, *future*, *course*, *hope*, *better* and *sources*. These terms speak to the motivation of many learners to use the knowledge to improve the world; interestingly, no job related term appears in this list (the term *career* occurs at rank 20), indicating that many of our

learners have an intrinsic, rather than an extrinsic motivation. They are brought to the course and engage with the materials not out of need for career change or certification (as was commonly observed among MOOC learners in previous work [15]), but rather out of a desire to be able to spark positive change in the world. Given the topic of the course and its relevance to the issues facing society today, this certainly affects learner motivation in some sense, but this also demonstrates that MOOCs can be instrumental to shaping the next generation of emerging technologies in making the subject matter accessible to the masses.

## 6 Discussion

Based on the existing literature and theory on self-regulated learning, we designed SRLx to encourage and support learners in adopting effective self-regulated learning habits in MOOCs. SRLx enables learners express their (changing) motivation and to set their own goals and track their progress towards them in real-time instead of following instructor-prescribed goals.

To evaluate the efficacy of SRLx we deployed it in a MOOC with more than 2,900 active learners to observe to what extent and how learners engage with it. Despite the inconsistencies we observed based on previous related work, learner interactions with SRLx offer novel insights about the role of motivation expression and plan formulation for MOOC learners. We find (i) that as the course progresses, learners are able to plan their time commitment more effectively, (ii) a strong trend of intrinsic motivation shared by learners with the motivation expression interface, and (iii) learners are most conservative with the way they plan to commit time to the course compared to video and quiz activity planning.

Given our findings on the progression of learner's planning strategies over time with SRLx, we are able to offer an explanation of the findings by Yeomans and Reich [24] who found that plans that were formulated about time were less likely to succeed: that intervention took place at the beginning of a course, where learners formulated time plans over the long-term—requiring the foresight of many weeks in the future; SRLx, on the other hand, allows learners to set a new plan at the beginning of each course week (short- to medium-term). Combined with our evidence that learners become more effective at plan formulation over the span of the course, we conclude that time-specific plans are likely only to be ineffective when on a long-term scale; and when used on a short- to medium-term scale, they can be effective and attainable.

Future research should implement SRLx as a randomized controlled trial, or A/B test, in MOOCs to explore questions of causality—does SRLx directly cause learners to learn and engage more? Finally, SRLx, as presented here, is completely individualistic—learners only receive feedback on their own plan formulations and motivation expressions. By making SRLx social, or showing learners the planning behavior and performance of their peers as well as their own, this could present a promising way to leverage the scale of MOOCs and improve learner performance through increased social presence.

## References

1. Bodily, R., Verbert, K.: Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Trans. Learn. Technol.* **10**(4), 405–418 (2017)
2. Bodily, R., Verbert, K.: Trends and issues in student-facing learning analytics reporting systems research. In: *LAK 2017*, pp. 309–318 (2017)
3. Bull, S., Kay, J.: Open learner models as drivers for metacognitive processes. In: Azevedo, R., Alevan, V. (eds.) *International Handbook of Metacognition and Learning Technologies*. SIHE, vol. 28, pp. 349–365. Springer, New York (2013). [https://doi.org/10.1007/978-1-4419-5546-3\\_23](https://doi.org/10.1007/978-1-4419-5546-3_23)
4. Davis, D., Chen, G., Hauff, C., Houben, G.-J.: Activating learning at scale: a review of innovations in online learning strategies. *Comput. Educ.* **125**, 327–344 (2018). <https://doi.org/10.1016/j.compedu.2018.05.019>
5. Davis, D., Chen, G., van der Zee, T., Hauff, C., Houben, G.-J.: Retrieval practice and study planning in MOOCs: exploring classroom-based self-regulated learning strategies at scale. In: Verbert, K., Sharples, M., Kloboučar, T. (eds.) *EC-TEL 2016*. LNCS, vol. 9891, pp. 57–71. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45153-4\\_5](https://doi.org/10.1007/978-3-319-45153-4_5)
6. Davis, D., Jivet, I., Kizilcec, R.F., Chen, G., Hauff, C., Houben, G.J.: Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In: *LAK 2017*, pp. 454–463 (2017)
7. Gregori, E.B., Zhang, J., Galván-Fernández, C., de Asís Fernández-Navarro, F.: Learner support in moocs: identifying variables linked to completion. *Comput. Educ.* **122**, 153–168 (2018)
8. Guerra, J., Hosseini, R., Somyurek, S., Brusilovsky, P.: An intelligent interface for learning content: combining an open learner model and social comparison to support self-regulated learning and engagement. In: *IUI 2016*, pp. 152–163 (2016)
9. Hood, N., Littlejohn, A., Milligan, C.: Context counts: how learners’ contexts influence learning in a MOOC. *Comput. Educ.* **91**, 83–91 (2015)
10. Jivet, I., Scheffel, M., Drachler, H., Specht, M.: Awareness is not enough: pitfalls of learning analytics dashboards in the educational practice. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017*. LNCS, vol. 10474, pp. 82–96. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_7](https://doi.org/10.1007/978-3-319-66610-5_7)
11. Kizilcec, R.F., Davis, G.M., Cohen, G.L.: Towards equal opportunities in MOOCs: affirmation reduces gender & social-class achievement gaps in china. In: *L@S 2017*, pp. 121–130 (2017)
12. Kizilcec, R.F., Halawa, S.: Attrition and achievement gaps in online learning. In: *L@S 2015*, pp. 57–66 (2015)
13. Kizilcec, R.F., Pérez-Sanagustín, M., Maldonado, J.J.: Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Comput. Educ.* **104**, 18–33 (2017)
14. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: *LAK 2013*, pp. 170–179. ACM (2013)
15. Kizilcec, R.F., Schneider, E.: Motivation as a lens to understand online learners: toward data-driven design with the olei scale. *ACM Trans. Comput. Hum. Interact. (TOCHI)* **22**(2), 6 (2015)

16. Koedinger, K.R., McLaughlin, E.A., Jia, J.Z., Bier, N.L.: Is the doer effect a causal relationship?: How can we tell and why it's important. In: LAK 2016, pp. 388–397 (2016)
17. Koller, D., Ng, A., Do, C., Chen, Z.: Retention and intention in massive open online courses. *Educause Rev.* **48**(3), 62–63 (2013)
18. Kulkarni, C.E., Bernstein, M.S., Klemmer, S.R.: Peerstudio: rapid peer feedback emphasizes revision and improves performance. In: L@S 2015, pp. 75–84 (2015)
19. Littlejohn, A., Hood, N., Milligan, C., Mustain, P.: Learning in MOOCs: motivations and self-regulated learning in MOOCs. *Internet High. Educ.* **29**, 40–48 (2016)
20. Mahoney, M.J., Moore, B.S., Wade, T.C., Moura, N.G.: Effects of continuous and intermittent self-monitoring on academic behavior. *J. Consult. Clin. Psychol.* **41**(1), 65 (1973)
21. Nawrot, I., Doucet, A.: Building engagement for MOOC students: introducing support for time management on online learning platforms. In: WWW 2014, pp. 1077–1082 (2014)
22. Sagotsky, G., Patterson, C.J., Lepper, M.R.: Training children's self-control: a field experiment in self-monitoring and goal-setting in the classroom. *J. Exp. Child Psychol.* **25**(2), 242–253 (1978)
23. Schippers, M.C., Scheepers, A.W., Peterson, J.B.: A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. *Palgrave Commun.* **1**, 15014 (2015)
24. Yeomans, M., Reich, J.: Planning prompts increase and forecast course completion in massive open online courses. In: LAK 2017, pp. 464–473 (2017)
25. Zheng, S., Rosson, M.B., Shih, P.C., Carroll, J.M.: Understanding student motivation, behaviors and perceptions in MOOCs. In: CSCW 2015, pp. 1882–1895 (2015)
26. Zimmerman, B.J., Bandura, A., Martinez-Pons, M.: Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting. *Am. Educ. Res. J.* **29**(3), 663–676 (1992)
27. Zimmerman, B.J., et al.: A social cognitive view of self-regulated academic learning. *J. Educ. Psychol.* **81**(3), 329–339 (1989)



# Motivating Students to Enhance Their Knowledge Levels Through Personalized and Scrutable Visual Narratives

Bilal Yousuf<sup>(✉)</sup>, Athanasios Staikopoulos, and Owen Conlan

ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin,  
Dublin, Republic of Ireland

{byousuf, athanasios.staikopoulos, Owen.Conlan}@scss.tcd.ie

**Abstract.** Continuous learning and development have been shown to be directly impacted by poor engagement. With the issue of poor engagement of learners with their course content when using Online Learning Environments (OLEs) still at large, this research aims to analyze the influence that visual narratives could have on encouraging students to study and improve their knowledge levels, and thereby support their continuous learning and development. Interactive and explorable visualizations have been commonly used in OLEs to support students' continuous learning, development and engagement by highlighting their coverage of course content, presenting the tasks completed and their performance, displaying the students learning model and showing peer comparisons. However, personalized visual narratives that present student knowledge levels which can be scrutinized and challenged have not been used in OLEs to date. The research discussed in this paper shows how personalized and scrutable visual narratives encouraged students, enrolled into an adaptive OLE as part of their undergraduate degree program, to study their course content and subsequently improve their knowledge levels.

**Keywords:** Visualization techniques for learning · Personalized E-learning  
Interactive narrative

## 1 Introduction

Continuous learning and development is key to all forms of learning, whether it is classroom based or through Online Learning Environments (OLEs). The literature [1] has shown that students' engagement with course content is a key factor in their continuous learning and development, meaning that their knowledge levels may be affected by poor engagement. The usage of OLEs is continuing to rise [2] and supporting students to engage with such technologies has been an area of focus for many researchers in Technology Enhanced Learning [3–5]. Students' engagement with courses delivered through OLEs has been shown to decrease over time, when compared to traditional classroom settings [6, 7], thereby impacting their continuous learning and development. This research aims to address an ongoing challenge in Technology Enhanced Learning of supporting students continuous learning and development by describing and evaluating

an approach with the objective of encouraging students to enhance their knowledge levels. The research discussed in this paper focuses on using personalized and scrutable visual narratives in OLEs to guide students through their acquired knowledge and to encourage them to study their course material and thus to support their continuous learning and development.

Visualizations have been effectively used in supporting both the comprehension of a complex dataset and in allowing patterns to be detected [8]. OLEs that utilize visualizations to aid learning usually support visual interactions and visual explorations to enable students to scrutinize data and gain valuable insights [9–11]. However, the literature has shown that it is important to guide learners through their data to support them in understanding it [12]. Visual narratives (ordered sequences of steps consisting of visualizations and textual descriptions) utilize the benefits of visualizations, visual interactions and at times visual explorations to guide users through messages that are being communicated.

Visual narratives have been used in the Information Visualization domain [13–15], in online journalism [16, 17], and they have also been used in OLEs to support learner engagement [4, 18]. The results presented in the Information Visualization domain and in OLEs discussing the evaluation of visual narrative usage have been very encouraging [4, 13, 15]. However, to date, visual narratives have not been used to present student knowledge levels to learners using OLEs. The research discussed in this paper provides each learner a personalized and scrutable visual narrative (hereafter, referred to as visual narrative), which amongst other things communicates the knowledge level that the learning system has calculated per student per study topic. The visual narratives enable students to view and scrutinize the calculated knowledge levels (by the OLE) and then allows them to challenge these levels through controllable visualizations. The visual narratives were integrated into the AMAS adaptive OLE [19], which is used by second-year Computer Science and third-year Computer Engineering students to learn Database programming as part of *Information Management and Data Engineering*, which is a module from their undergraduate degree program in Trinity College Dublin.

The AMAS adaptive OLE consists of study topics (each with several sub-topics) followed by activities, which involve SQL programming and building databases. In the 2017–18 academic year, questions related to study topics were added to the AMAS OLE course (hereafter, referred to as the course) and students had the option to answer these questions enabling them to gauge how well they are mastering the course. In the same academic year, a visual narrative presenting the student's knowledge levels that could be visually challenged was made available to each student enrolled in the module. The visual narratives consisted of a beginning, middle, and end, with each section of the story, represented using interactive visualizations, textual descriptions, and exploration links to support data scrutinization. The visual narratives were presented in a web browser with each section of the story presented in tabs. Viewing and scrutinizing the visual narratives were also optional and were available through a link on the course.

In the 2017–18 academic year, 143 students enrolled in the course. 64 of the enrolled students attempted some or all the study topics related questions and 52 of these 64 learners also used their visual narratives to view and scrutinize their knowledge levels.

The evaluation discussed in this paper analyzes the extent to which the visual narratives encouraged students to study their course topics and to improve their knowledge levels.

The remainder of this paper is as follows: Sect. 2 discusses the related work. Section 3 describes the details of the visual narratives supported by this research and Sect. 4 discusses the research approach. Section 5 evaluates the influence of visual narratives in supporting students to study and increase their knowledge levels and Sect. 6 presents the conclusions.

## 2 Related Work

Visualizations have been frequently used in the Learning Analytics, Open Learner Modelling and Educational Data Mining domains within OLEs to present relevant information to both students and educators. This section analyzes the use of visualizations in OLEs, specifically in these three domains, to (1) communicate student data to learners and allow them to scrutinize it, and (2) to influence learners to enhance their knowledge levels.

### 2.1 Presenting and Scrutinizing Student Data

Dashboards with one or more visualizations have been a popular platform for the presentation of student data to learners using OLEs. Some of the dashboards communicate important information that can be interpreted at a glance. For example, Course Signals provides early warnings to students of potential problems by highlighting learner efforts and performance using traffic light indicators [20]. Competency Map presents learner competencies against course assignments using color-coded maps [22]. Other systems require students to interact with the dashboard to interpret some of the data. For example, LARAE visualizes learner actions to support progress and awareness and supports peer comparisons of forum posts [21]. Interacting with the dashboard allows modules to be selected. The CAM Dashboard consists of goal-oriented visualizations for students to reflect on the time spent on assigned activities, and view comparisons with fellow learners regarding progress made towards goals [23]. The StepUp dashboard provides visualizations of learner data to assist students in the learning process and to promote reflection [24]. ALAS-KA is a Khan Academy plug-in that processes raw learner data to extract information at a higher level through a set of metrics and presents it through visualizations to learners [25]. It uses five metrics (platform usage, progress, time distribution, gamification habits and exercise solving habits) to visually present student activity traces to learners to support self-reflection.

There has also been a focus to support students in scrutinizing the data presented through popular visual interaction techniques such as select, explore and elaborate (details-on-demand, drilldown views), filter, and coordinated views. Progressor, for example, visualizes the learner model and presents social comparisons supporting selection and filtering [26]. SAM allows students to apply filters to the data presented to view time spent and to drill down to view details behind the individual bars of the histogram [9]. Narcissus allows students to select components of the visualizations presented to

view the details behind the nodes [27]. The CAM Dashboard supports coordinated views where students can select an element on one visualization and see the details on another [23].

## 2.2 Presenting Competencies to Students

The focus of Open Learner Modelling is to present learner models to the individual students using visualizations to support reflection and allowing them to scrutinize and participate in the construction or modification of it. From the related work, OLMlets facilitates independent learning and assessment through visualizations by providing students with a skills meter, a ranked list and a textual summary of his/her knowledge level [28]. Competency Map and Next-Tell present the competencies that have been acquired by the learners [22, 29]. In addition to allowing students to explore their own model, Narcissus also provides visual representations of group models enabling students to view group progress [27].

In addition to viewing and exploring their learner models, a number of systems enable students to directly update their knowledge level by clicking an edit link and directly entering a score they believe their knowledge level should be at [10, 30, 31]. Other systems allow learners to influence their models by answering additional questions [32–35]. Flexi-OLM, for example, allows students to update their knowledge level in two ways, (1) by allowing them to click an edit link and directly enter a score they believe their knowledge level should be at, and (2) by persuasion, where the student can click a link and answer a series of questions, which updates the learner's knowledge level [33]. More recently, Bull proposed a negotiated learner modeling approach that relied on discussions between the system and the student using several categories, including skill level and competencies, statements and challenges, and the learner's understanding to reach a set of outcomes that would be used to update the model [36].

## 2.3 Advancing the State of the Art

The literature has highlighted the value that students have gained from using visualizations in OLEs [9, 26], however, there have also been evaluations which found that students at times had difficulty in understanding the data presented to them [24, 37, 38]. Data misinterpreted by students can impact motivation levels and lead to poor performance, and hence it is important that learners are guided through it [12]. The literature has highlighted visual presentations resembling narratives have been found to be quite useful to students with minimal misinterpretation of data [39]. To date, VisEN has been the only system to use visual narratives in OLEs and the evaluations of the framework have shown that visual narratives can support student engagement [4, 18]. The research presented in this paper aims to progress the state of the art by presenting a more complete and personalized story to the student about his/her engagement, time spent on activities and knowledge gained using the AMAS OLE, consisting of a start, middle, and end. The visual narrative first presents the study topics that the student has engaged with to date and this is followed by the time the learner has spent on these study topics and ends with the knowledge gained by the learner.



Similar to the related work [32–35], the research presented in this paper enables students to challenge their knowledge levels. However, it enhances it in two ways: it first guides students through their engagement and activity on the OLE and then presents their performance and knowledge levels, with the aim to allow the learners to gauge how actively they were engaged with their course and then to reflect on the knowledge level that the system calculated the students to have. Secondly, it allows the students to visually challenge the knowledge levels shown by selecting visual elements and dropping them in locations which they believe is a more accurate representation of their knowledge and then justify this by answering questions. The related work does not support visual control, instead, it provides students with a text box to enter a score.

### 3 Visual Narratives Supported by the AMAS OLE

Each student enrolled in the course was provided with a visual narrative which consisted of a personalized story to guide the learner through his/her interactions with the course. The visual narrative presented how the student had engaged with the course content to date, including a detailed breakdown of engagement per study topic. The story also presented the study topics completed by the learner and the time spent working on each. Finally, it presented the knowledge that AMAS calculated the user would have based on the level of engagement and on the results of study topic-related questions. Each part of the visual narrative allowed students to scrutinize the data presented by viewing data related to it and could view peer comparisons.

The engagement per student was calculated based on the time spent on study topics, pages viewed, resources downloaded and at what stage of the course the resources were accessed. The visual narratives guided the learners through their individual engagement at the course level and at a study topic and sub-topic level. The description in the visual narratives not only explained the data but informed the students of how well they were engaging with the course. The second part of the visual narrative guided students through the time they had spent on assigned study topics and resources used and allowed learners to view the time that peers had spent on completed study topics. The comparison allowed students to estimate how much time it could take them to complete a study topic if it had not been started.

The final part of the visual narrative presented the knowledge acquired by the student as calculated by the AMAS OLE. The aim of guiding students through their engagement and time spent on study topics and resources before they were presented the calculated knowledge was to aid them in understanding some of the data used in the calculation. Figure 1A presents the students' current knowledge level and the description (not shown in the figure) guided the student through this data by explaining how well the student has mastered the study topics. Figure 1B presents a student's knowledge level details for a study topic. The description (not shown in the figure) guided the learner through the data by explaining it and suggesting areas for improvement. The students' knowledge level was calculated based on (1) the degree to which learning resources were used, and (2) the results from the questions attempted by learners following the completion of each study topic. In the detailed view, students could challenge their knowledge level score

by dragging the bars up, if a student believed his/her score for the sub-topic should be higher or down if it should be lower. Dragging any of the bars up resulted in a set of random but relevant questions (to the sub-topic) being presented to the student to justify the knowledge gain and the answers provided by the student were used to recalculate the knowledge level. The students were limited to a single challenge per sub-topic to prevent an artificial knowledge gain that may result if students could challenge multiple times and guess the answers.

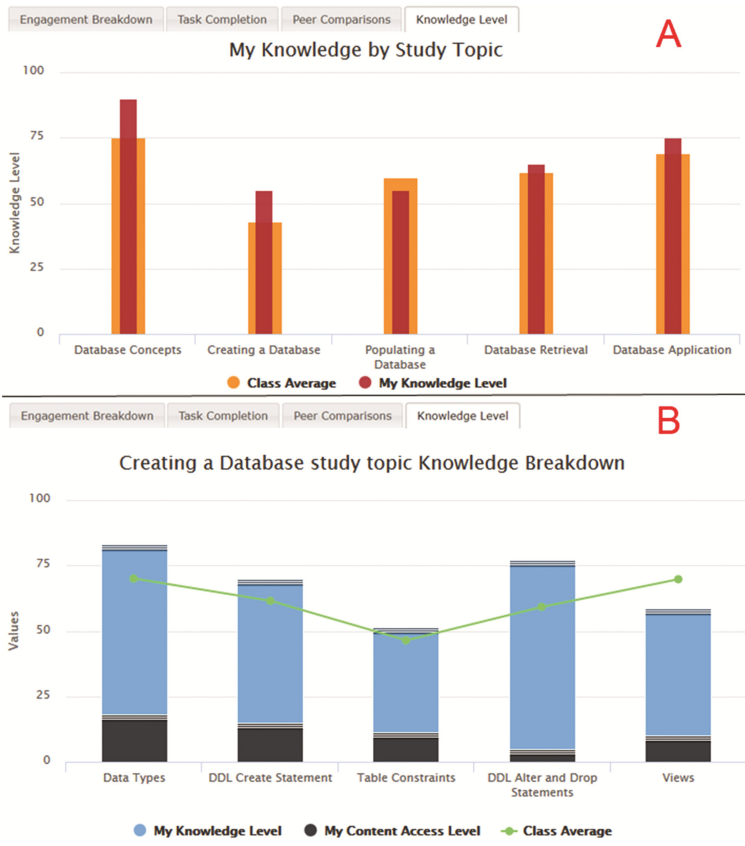


Fig. 1. Visualizations from the knowledge level part of the visual narrative

## 4 Research Approach

The aim of this research was to present visual narratives to students that could be analyzed and controlled to support their learning and development when enrolled in the Information Management and Data Engineering module using the AMAS OLE. The research also aimed to evaluate the impact that the visual narratives had on influencing the students' learning and development by engaging with the course material and

enhancing their knowledge levels after viewing their visual narratives. The visual narratives enabled students to analyze their engagement, the time spent on their learning activities and their performance. The learners could scrutinize their engagement by examining it at a task level and comparing it to class average and peer engagement. The visual narratives allowed the learners to explore the time they spent on tasks and estimates time to completion by scrutinizing other students' completion times. Similarly, students could examine how their knowledge level was calculated and challenge it. The process of scrutinizing the data required the students to click on elements within the visualization to load views which presented the related details visually. The aim of supporting visual scrutiny was to allow students to gain a better understanding of their own data and the message communicated through the visual narratives.

A study was conducted during and after the course had completed which analyzed (1) the impact (if any) that the visual narratives had student development, by examining the students' knowledge levels at various stages during the course, (2) the visual narrative usage patterns, and (3) the learners' perceptions towards the visual narratives. The research approach adopted by this study consisted of both quantitative and qualitative analyses. The data collected for this study consisted of student-logged data that was recorded by the AMAS OLE which included all the interactions the students had with their course material, with the questions attempted and with their visual narratives. The data collected also included the students' responses to a post-course questionnaire and their opinions regarding the impact that the visual narrative had on influencing them to develop their competencies. The quantitative analyses examined the impact that the visual narratives had on student knowledge levels using statistical measures. This analysis involved the examination of all the student interactions with the AMAS OLE during the course, which was over 87,000 interactions in the academic year. The quantitative analyses also examined the students' responses to the post-course questionnaire. The post-course questionnaire consisted of open-ended statements and students could provide their opinions. The qualitative analysis examined the learners' opinions towards the usefulness and the impact the visual narratives had on their development.

## 5 Evaluation

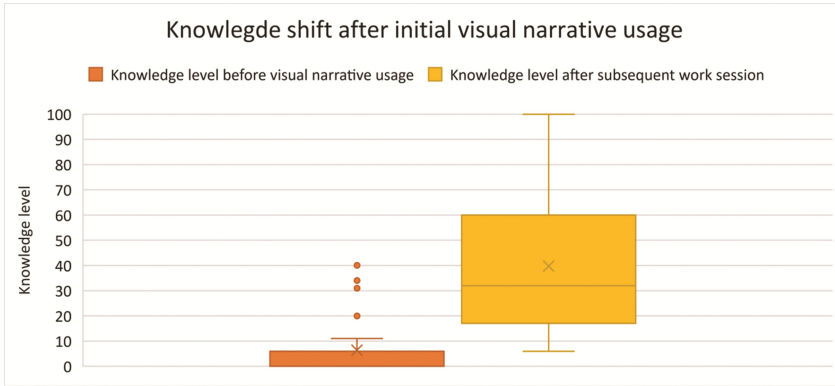
The AMAS OLE course ran during the first semester of the 2017–2018 academic year. The course consisted of five study topics which involved students reading course material, analyzing examples describing database schemas and studying SQL programming. The AMAS OLE provided an environment which allowed the students to study the material and work through the examples. Each of the five study topics ended with a set of questions which were optional but allowed students to gauge how well they had mastered the topics. Following the study topics, students were required to create SQL programs, and build and deliver a database using the AMAS OLE. The visual narratives were automatically updated as students studied and worked through their tasks, thereby providing them with a live reference point to understand how they were progressing through the course. Visual Narrative usage was optional, meaning that the students used them of their own volition.

The evaluation discussed in this paper focuses on the influence that the visual narrative had on students that increased their knowledge levels during the course. As mentioned in Sect. 1, 143 students enrolled in the course during the 2017–18 academic year. 64 students attempted some or all the study topic questions and 52 of these 64 had also used their visual narratives. The aim of the study discussed in this evaluation examines the visual narrative usage and knowledge gained by these 52 students to determine whether the visual narratives had any impact on knowledge gained. The first part of the study (Analysis 1) examines the student logged data to determine if the visual narrative usage influenced knowledge gain. It does this by analyzing students' knowledge levels before and after their visual narrative usage. It also examines the immediate response of the students after visiting their visual narratives and analyzes the visual narrative usage patterns. For the purpose of Analysis 1, a work session is defined, which involves all of a student's interactions with the AMAS OLE (studying course material, attempting questions and interacting with his/her visual narratives) until the learner logs out of the OLE. Part two of this study (Analysis 2) examines the students' perceptions towards the visual narratives, focusing on the usefulness of guiding learners through their knowledge levels by analyzing the relevant responses to the post-course questionnaire and their comments.

### 5.1 Student Knowledge Gain After Visual Narrative Usage

Analysis 1 focuses on the impact that the visual narratives had on student knowledge levels by analyzing their logged data. It examines student knowledge gains at various stages during the course and their visual narrative usage immediately prior to it.

Initially, Analysis 1 examines the student knowledge levels prior to their first visual narrative usage (by which point they may have attempted some questions) and then examines the knowledge level during the subsequent work session immediately after visual narrative usage. From the 52 students that attempted some or all the questions after the study topics, 63% showed a significant increase in knowledge in the subsequent work session after using their visual narratives for the first time. A shift of their knowledge levels from a mean of  $6.4 \pm 12.00$  to  $39.71 \pm 26.74$  was calculated to be significant at  $p < 0.05$ , with a  $t$ -value = 6.57. Figure 2 presents the knowledge levels of these students before their first usage of their visual narratives and their knowledge level after their subsequent work session. It is important to note that all the students that experienced this significant increase in knowledge levels had covered varying degrees of their study topics before they used their visual narratives for the first time, but their knowledge levels remained relatively low. It was only during the work session in which they used their visual narratives for the first time that they studied further and improved their knowledge. This highlights that the visual narratives had a positive impact on the knowledge levels of most of the students that used them. From the remaining students (of the 52 learners), 10% had a relatively high knowledge level (mean of  $73 \pm 13.92$ ) before their first visual narrative usage and hence did not have the same motivation to improve it. The other 27% did not show any knowledge level improvement in the subsequent work session.



**Fig. 2.** This highlights the shift in knowledge level for the students that experienced an enhancement in their knowledge levels after using their visual narratives for the first time.

To understand why 63% of the students experienced a boost in knowledge level and the other 27% did not (excluding the 10% with high knowledge levels prior to visual narrative usage), the second part of Analysis 1 examines the visual narrative usage patterns of both sets of students. A common usage pattern was apparent amongst the students that experienced a significant knowledge gain during the subsequent work session after their first usage of their visual narratives. It was found that they were involved in repeated visual narrative visits during that work session as they studied and attempted questions. Such a pattern was not evident amongst the other students (27%) who viewed their visual narrative fewer times than the former group.

Following this finding, Analysis 1 examines the correlation between visual narrative usage and knowledge level using the Pearson correlation coefficient for all 52 students that attempted the study topics and used their visual narratives. The correlation was found to be weak between visual narrative usage and knowledge levels,  $r(50) = .26$ ,  $p < .0005$  with the mean knowledge level equal to  $43.02 \pm 27.0$  and the mean visual narrative usage equal to  $6.1 \pm 5.44$ . This finding shows that continuous usage of the visual narrative in further work sessions did not have the same influence on knowledge enhancement as was the case with the work session immediately after the students' first visual narrative usage.

The final part of Analysis 1 examines the end-of-course knowledge levels of the 63% of the 52 students that experienced a significant increase in knowledge following their initial use of their visual narratives versus the 27% that did not experience this. The 10% of students with high knowledge levels were excluded as they would not have had the same motivation to improve their knowledge levels as it was already quite high. The average end of course scores for the 63% of students was  $44 \pm 27.69$ , with 9 of these students completing the course with a first-class honors grade. The average score of the 27% of students who did not experience a significant knowledge increase following their initial visual narrative usage was  $35.35 \pm 21.9$ , with none of these students completing the course with a first-class honors grade. This part of Analysis 1 also examines the end-of-course knowledge levels of the 12 students (from the original 64) that attempted some

or all the study topic questions but did not view their visual narratives. It was found that their average end-of-course scores were  $24.5 \pm 21$ , which shows that visual narrative usage supported students in enhancing their knowledge levels.

### 5.2 Student Perceptions Towards Their Visual Narratives

Following the course, students completed a questionnaire which included statements focusing on the visual narratives. The students were encouraged to provide comments after each response. Analysis 2 examines both the responses and student comments to some of the statements that focused on the visual narratives, namely those that covered the knowledge level section of the story. The responses from the students (52) who had attempted the study topic questions and had used their visual narratives were examined.

**Statement 1: The Visualizations Presenting my Knowledge Levels Motivated me to Improve my Knowledge of the Study Topics.** Figure 3 presents the student responses to statement 1, which shows that 64% of the students agreed or strongly agreed with the statement. 21% of students were undecided and 15% of the learners either disagreed or strongly disagreed with the statement.

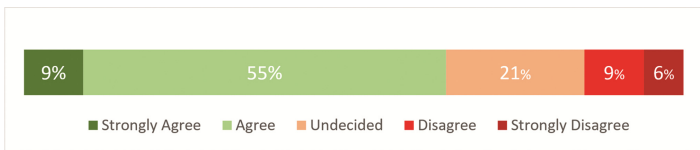


Fig. 3. Student responses to statement 1.

There were no comments provided by the students who did not improve their knowledge levels in the subsequent work session following their first visual narrative usage. From amongst those that did improve their knowledge levels in the subsequent work session, the feedback reflected their positive responses to the statement. For example, one student commented: *“Motivated me to improve my knowledge on study topics with smaller bar charts”*.

**Statement 2: I did not Always Agree with my Knowledge Level per Study Topic as Presented by the Learning Environment.** Figure 4 presents the student responses to statement 2, which shows that 57% of the students agreed or strongly agreed with the statement, 12% disagreed with it, and 31% of students were undecided.

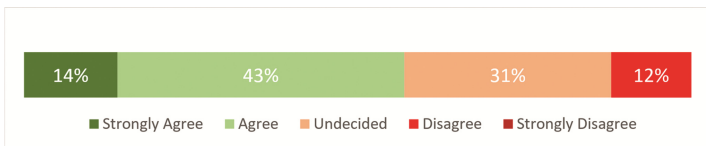
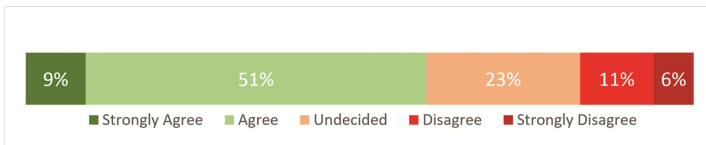


Fig. 4. Student responses to statement 2.

As mentioned in Sect. 3, the knowledge levels per study topic were formed using the time spent on learning material (their engagement) and the responses to questions. It is important to note that offline study (which may have included downloading the pdf and studying it) were not factored into the knowledge levels. The student comments focused on this part of the calculation and hence the majority agreed or were undecided about statement 2. For example, one student who agreed with the statement commented: *“I found myself extremely comfortable using the practice DB, yet only had 12% on the engagement level score”*.

**Statement 3: I Found it Useful to Visually Scrutinize a Breakdown of my Knowledge Level per Study Topic.** The visual narrative allowed students to visually scrutinize how their knowledge levels were calculated by study topic (as shown in Fig. 1B) and statement 3 focused on how useful the learners found such explorations. Figure 5 presents the student responses to statement 3, which shows that 60% of the students agreed or strongly agreed with the statement, with 23% undecided and 17% of the learners disagreed or strongly disagreed with it.



**Fig. 5.** Student responses to statement 3.

The student feedback from the learners that immediately improved their knowledge levels after visual narrative usage was reflective of their responses. For example, one student commented: *“Helps me to view more easily which areas I did better than others and where I needed to spend more time”*.

### 5.3 Findings from Analysis 1 and Analysis 2

Analysis 1 highlights an important finding that the visual narratives influenced the majority of students to study and enhance their knowledge levels. However, the students that experienced this enhancement (63%) failed to continue to improve their knowledge level to the same degree throughout the course as their knowledge levels only slightly improved from a mean of  $39.71 \pm 26.74$  to a mean of  $44 \pm 27.69$  by the end of the course. This shows that the visual narratives were very useful in providing an initial support to students to study and improve their knowledge levels. Further research is required to investigate how this improvement can be sustained over the duration of the course.

The questionnaire responses from the students (statements 1 and 3) showed that the visual narratives supported them in enhancing their knowledge levels and visually scrutinizing how the scores were calculated was useful as it informed the learners where they should improve. The responses to statement 2 highlighted that the majority of the students did not agree with how AMAS calculated their knowledge level, specifically how their engagement score was calculated, which accounted towards 20% of their

knowledge level. From Figs. 3, 4, and 5, it can be seen that over 20% of students were undecided regarding statements 1, 2, and 3 and in most cases, these learners did not provide comments with their statement responses. It may be possible that since the students disagreed with the engagement score calculation, they were undecided regarding the usefulness of the visual narratives. Hence the metric used to calculate the engagement score will be reviewed prior to the next deployment of the course.

## 6 Conclusions

This paper aimed to address the ongoing challenge faced by Technology Enhanced Learning, where students' continuous learning and development has been impacted by poor engagement with their course content when using OLEs. The research discussed in this paper introduced personalized and scrutable visual narratives to OLEs, specifically to the AMAS adaptive OLE. The visual narratives allow students to visually scrutinize the story in order to gain a better understanding of the message communicated and supported them in visually challenging their calculated knowledge levels.

The evaluation found that the majority of students (63%) that attempted their study topics questions benefitted from their visual narratives and this finding was further reinforced from the responses of the students to statement 1 of the post-course questionnaire. Further analysis in the first part of the study found that these students experienced a knowledge level boost after their initial visual narrative usage and further usage of their narrative in further work sessions did not influence their knowledge gain to the same degree. This was evident from the relatively small difference between their average knowledge level after their initial visual narrative usage and their knowledge levels at the end of the course. Visual narratives have been recently used in OLEs (4, 18) but this work progresses the state of the art by presenting a complete visual narrative, describing how each student was engaging with his/her course content, time spent of learning activities, resources used, and knowledge gained to date. Overall, this research found that personalized and scrutable visual narratives had a major impact in supporting students to enhance their knowledge levels.

Future work will investigate ways to maintain the initial boost provided by the visual narratives to the students for the duration of the course. In addition, alternative metrics will be used to calculate student engagement which was used as part of the students' knowledge levels, specifically to incorporate learners' offline activity. This will involve automated discussions between AMAS and the students to include coverage of their offline activity and their current competencies regarding the assigned study topics.

**Acknowledgments.** This research is supported by the Science Foundation Ireland through the CNGI program (Grant 12/CE/I2267) in the ADAPT Center ([www.adaptcentre.ie](http://www.adaptcentre.ie)) at Trinity College Dublin.



## References

1. Kuh, G.D.: Assessing what really matters to student learning inside the national survey of student engagement. *Change: Mag. High. Learn.* **33**(3), 10–17 (2001)
2. Berin, J.: Use Of MOOCs And Online Education Is Exploding: Here's Why. <https://www.forbes.com/sites/joshbersin/2016/01/05/use-of-moocs-and-online-education-is-exploding-heres-why/#799ccaa67649>. Accessed 02 Apr 2018
3. Jordan, K.: Initial trends in enrolment and completion of massive open online courses. *Int. Rev. Res. Open Distrib. Learn.* **15**(1), 33–160 (2014)
4. Yousuf, B., Conlan O.: Supporting student engagement through explorable visual narratives. *IEEE Trans. Learn. Technol.* **1** (2017). <https://doi.org/10.1109/TLT.2017.2722416>
5. Dixson, M.D.: Measuring Student engagement in the online course: the online student engagement scale (OSE). *Online Learn.* **19**(4) (2015)
6. Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., Woolf, B.P.: Repairing disengagement with non-invasive interventions. In: *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, AIED, vol. 2007, pp. 195–202 (2007)
7. Young, S., Duncan, H.E.: Online and face-to-face teaching: How do student ratings differ? *J. Online Learn. Teach.* **10**(1), 70 (2014)
8. Card, S.K., Mackinlay, J.D., Shneiderman, B.: *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, Burlington (1999)
9. Govaerts, S., Verbert, K., Duval, E., Pardo, A.: The student activity meter for awareness and self-reflection. In: *Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 869–884. ACM (2012)
10. Kump, B., Seifert, C., Beham, G., Lindstaedt, S.N., Ley, T.: Seeing what the system thinks you know: visualizing evidence in an open learner model. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 153–157 (2012)
11. Loboda, T.D., Guerra, J., Hosseini, R., Brusilovsky, P.: Mastery grids: an open source social educational progress visualization. In: Rensing, C., de Freitas, S., Ley, T., Muñoz-Merino, Pedro J. (eds.) *EC-TEL 2014*. LNCS, vol. 8719, pp. 235–248. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11200-8\\_18](https://doi.org/10.1007/978-3-319-11200-8_18)
12. Lonn, S., Aguilar, S.J., Teasley, S.D.: Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Comput. Hum. Behav.* **47**, 90–97 (2015)
13. Lee, B., Kazi, R.H., Smith, G.: SketchStory: telling more engaging stories with data through freeform sketching. *IEEE Trans. Vis. Comput. Graph.* **19**(12), 2416–2425 (2013)
14. Tableau (2018). <http://www.tableau.com>
15. Satyanarayan, A., Heer, J.: Authoring narrative visualizations with ellipsis. *Comput. Graph. Forum* **33**(3), 361–370 (2014)
16. New York times. <http://nyti.ms/sFYztk>. Accessed 08 June 2017
17. The Financial Times. <http://www.ft.com/cms/s/0/663b649e-b7e6-11de-8ca9-00144feab49a.html>. Accessed 19 Feb 2016
18. Yousuf, B., Conlan, O.: VisEN: motivating learner engagement through explorable visual narratives. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) *EC-TEL 2015*. LNCS, vol. 9307, pp. 367–380. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24258-3\\_27](https://doi.org/10.1007/978-3-319-24258-3_27)

19. Staikopoulos, A., O’Keeffe, I., Rafter, R., Walsh, E., Yousuf, B., Conlan, O., Wade, V.: AMASE: a framework for composing adaptive and personalised learning activities on the web. In: Popescu, E., Li, Q., Klamma, R., Leung, H., Specht, M. (eds.) ICWL 2012. LNCS, vol. 7558, pp. 190–199. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-33642-3\\_20](https://doi.org/10.1007/978-3-642-33642-3_20)
20. Arnold, K.E., Pistilli, M.D.: Course signals at Purdue. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 267–270. ACM (2012)
21. Charleer, S., Odriozola, S., Luis, J., Klerkx J., Duval, E.: LARAE: Learning analytics reflection & awareness environment. In: CEUR Workshop Proceedings, vol. 1238, pp. 85–87. CEUR-WS (2014)
22. Grann, J., Bushway, D.: Competency map: visualizing student learning to promote student success. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, pp. 168–172. ACM (2014)
23. Santos, J.L., Govaerts, S., Verbert, K., Duval, E.: Goal-oriented visualizations of activity tracking. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 143–152. ACM (2012)
24. Santos, J.L., Verbert, K., Govaerts, S., Duval, E.: Addressing learner issues with StepUp! In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 14–22. ACM (2013)
25. Ruiópez-Valiente, José A., Muñoz-Merino, P.J., Kloos, C.D.: A demonstration of ALAS-KA: a learning analytics tool for the khan academy platform. In: Rensing, C., de Freitas, S., Ley, T., Muñoz-Merino, P.J. (eds.) EC-TEL 2014. LNCS, vol. 8719, pp. 518–521. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11200-8\\_55](https://doi.org/10.1007/978-3-319-11200-8_55)
26. Hsiao, I.-H., Bakalov, F., Brusilovsky, P., König-Ries, B.: Progressor: social navigation support through open social student modeling. *New Rev. Hypermedia Multimed.* **19**(2), 112–131 (2013)
27. Upton, K., Kay, J.: Narcissus: group and individual models to support small group work. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 54–65. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-02247-0\\_8](https://doi.org/10.1007/978-3-642-02247-0_8)
28. Bull, S., Gardner, P., Ahmad, N., Ting, J., Clarke, B.: Use and trust of simple independent open learner models to support learning within and across courses. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 42–53. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-02247-0\\_7](https://doi.org/10.1007/978-3-642-02247-0_7)
29. Bull, S., Johnson, M.D., Masci, D., Biel, C.: Integrating and visualising diagnostic information for the benefit of learning. In: Reimann, P., Bull, S., Kickmeier-Rust, M., Vatrappu, R., Wasson, B. (eds.) *Measuring and Visualizing Learning in the Information-Rich Classroom*, p. 167. Routledge, Abingdon (2015)
30. Bull, S., Dong, X., Britland, M., Guo, Yu.: Can students edit their learner model appropriately? In: Woolf, Beverley P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 674–676. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69132-7\\_74](https://doi.org/10.1007/978-3-540-69132-7_74)
31. Czarkowski, M., Kay, J., Potts, S.: Web framework for scrutable adaptation. In: Kay, J., Lum, A., Zapata-Rivera, D. (eds.) *Proceedings of Learner Modelling for Reflection to Support Learner Control, Metacognition and Improved Communication, AIED workshop*, vol. 11, pp. 11–18 (2005)
32. Ginon, B., Boscolo, C., Johnson, M.D., Bull, S.: Persuading an open learner model in the context of a university course: an exploratory study. In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) ITS 2016. LNCS, vol. 9684, pp. 307–313. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-39583-8\\_34](https://doi.org/10.1007/978-3-319-39583-8_34)

33. Mabbott, A., Bull, S.: Student preferences for editing, persuading, and negotiating the open learner model. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 481–490. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_48](https://doi.org/10.1007/11774303_48)
34. Tchetagni, J., Nkambou, R., Bourdeau, J.: Explicit reflection in prolog tutor. *Int. J. Artif. Intell. Educ.* **17**(2), 169–215 (2007)
35. Thomson, D., Mitrovic, A.: Preliminary evaluation of a negotiable student model in a constraint-based ITS. *Res. Pract. Technol. Enhanced Learn.* **5**(1), 19–33 (2010)
36. Bull, S.: Negotiated learner modelling to maintain today's learner models. *Res. Pract. Technol. Enhanced Learn.* **11**(1), 10 (2016)
37. May, M., George, S., Prévôt, P.: TrAVis to enhance students' self-monitoring in online learning supported by computer-mediated communication tools. *Comput. Inf. Syst. Industr. Manag. Appl.* **3**, 623–634 (2011)
38. Sedrakyan, G., Leony, D., Muñoz-Merino, P.J., Kloos, C.D., Verbert, K.: Evaluating student-facing learning dashboards of affective states. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 224–237. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_17](https://doi.org/10.1007/978-3-319-66610-5_17)
39. Kuosa, K., Distanto, D., Tervakari, A., Cerulo, L., Fernández, A., Koro, J., Kailanto, M.: Interactive visualization tools to improve learning and teaching in online learning environments. *Int. J. Dist. Educ. Technol.* **14**(1), 1–21 (2016)



# Supporting the Adaptive Generation of Learning Game Scenarios with a Model-Driven Engineering Framework

Pierre Laforcade<sup>(✉)</sup> and Youness Laghouaouta

Computer Science Laboratory of Le Mans University, Le Mans, France  
{pierre.laforcade,youness.laghouaouta}@univ-lemans.fr

**Abstract.** Learning games are promising methods for autism therapy. In this context, our research project aims to propose an “escape-room” game for helping children with Autistic Syndrome Disorder (ASD) to learn visual performance skills. Given the specific needs of the intended players, the generation of learning scenarios has to be adaptive. For that, our proposal relies on Model Driven Engineering techniques to deal with dynamic scenarization instead of implementing fixed configurations of scenarios. Our approach proposes to express the game description components and child profiles as models from which adapted scenarios can be automatically generated by means of model transformations. In addition, an iterative co-design process based on rapid prototyping is introduced. It allows ASD experts to take part in the design activity and get fast feedback.

**Keywords:** Serious game · Autism · Learning scenarios  
Adaptation · Model Driven Engineering

## 1 Introduction

The use of serious games [3] in Autistic Syndrome Disorder (ASD) interventions has become increasingly popular during the last decade [4]. They are considered as effective new methods in the treatment of ASD and efficient means of transferring knowledge [4, 18, 19]. Computerized interventions for individuals with autism may be much more successful if motivation can be improved and learning can be personalized. In fact, game adaptivity (i.e. customize the game according to each learner individuality) is very important particularly for learner with specific needs.

This research work is conducted in the context of the *Escape it!* project. The objective is to develop a serious game to train visual skills of children with ASD. This serious game will borrow mechanics from “escape-room” games (i.e. the player has to solve a puzzle in order to open a locked door to escape the room). The current paper tackles the challenge of generating adapted learning sessions

to autistic children. For that, it was crucial to involve ASD experts in the first development stage. The aim is to guarantee that the proposed game fits to ASD characteristics while to be individually adaptive to each child.

We propose a model-driven design process that allows domain experts to take part in the design activity and guide the development of the game (i.e. the focus is on adaptation and the game scenes set-up). Hence, we provide experts with means to determine the game components and the way game sessions have to be constructed and adapted. Besides, our proposal includes a rapid prototyping support so that the experts can immediately test a playable version of the game and give relevant feedback about the adaptation and generation rules.

The remainder of this paper is organized as follows. In Sect. 2, we present the context of this research work. Section 3 provides a review of adaptation challenges and mechanisms for generating adapted scenarios. Then, Sect. 4 gives a global overview of our proposal followed by an application case in Sect. 5. Finally, Sect. 6 concludes this paper and presents future work.

## 2 The *Escape It!* Project

The project aims to develop a mobile *learning game* (i.e. a serious game with learning purposes) dedicated to children with ASD (Autistic Syndrome Disorder). The game intends to support the learning of visual skills derived from a curriculum guide [13]. It will be used both to reinforce and generalize the learning skills. These skills will be initiated by “classic” working sessions with tangible objects.

### 2.1 General Overview of the Serious Game

The serious game is based on a minimalist “escape-room” gameplay. The child (player) has to drag objects, sometimes hidden, to their correct locations in order to unlock the room’s door and get to the next level. The drag and drop gameplay for matching/sorting/categorizing pictures is already implemented in several mobile games targeting children with ASD. As for the “escape-room” orientation, it has been proposed by the autism experts involved in the project.

The involved experts consider that the proposed game can be an intermediate support for learning generalization between therapy structured setting and generalization in a child’s natural environment as fostered by the Pivotal Response Treatment (i.e. PRT is an intervention that focuses on the generalization of learned skills in the child’s natural environment [8]). The game propose to deal with “responding to multiple cues” and “self-management” which are among the four pivotal areas of PRT.

The game design relies on best practices founded in the literature [4, 19] and recommendations/requirements expressed by the ASD experts. The main concerns are listed below:

- Targeted skills: a subset of the visual performance skills derived from [13] that can be adapted for a mobile gameplay (e.g. matching an object to an identical

- object, sorting similar objects, categorizing objects with same functions or characteristics...).
- Variable game sessions: the game proposes from 3 to 6 levels at the convenience of the pairing adult or the child.
  - Scenes as meaningful living places grouped into themes: for example, the *bedroom*, *kitchen* and *living room* are related to the *home* theme. Whereas, *classroom* and *gymnasium* belong to the *school* theme.
  - Adapted difficulty: the difficulty level is set according to the current child's progress in the targeted skill. Basically, three successful activities for a same skill (along one or several game sessions) raise the difficulty level for this skill.
  - Generalizing the acquired skills: it is the process of taking a skill learned in one setting and applying it in other settings or different ways [9]. To this end, scenes have to be changed in accordance with previous difficulty levels. Hence, the game proposes non-identical challenges for the same skill. We quote variation examples: (i) changing the background and elements of a scene; (ii) adding background elements to disrupt visual reading; (iii) changing the objects to find and handle; (iv) adding other objects that are not useful for the resolution; (v) hiding objects behind or into others.

Figure 1 depicts an example of a scene which targets the B8 skill (i.e. sort non-identical items) in the 'Expert' difficulty level. Trucks and balls have to be found and moved into the appropriate storage boxes before the door opens. Interactive hiding places, like the closet and its drawer, can be opened showing hidden objects.



Fig. 1. An example of the *bedroom* scene

## 2.2 Components of a Game Scene and Design Issues

Whichever scenes are selected for the learning scenario, they share common features:

- A background image that depicts a familiar scene for children with recognizable objects.
- Several empty slots where objects to find can be placed.
- Additional decors to impair visual reading with respect to the difficulty level. Each one can:
  - Appear in different locations.
  - Create new slots for other game objects.
- Interactive hiding objects that provide new slots to hide objects and reveal them when touched.
- Solution objects where game objects have to be placed in/on. One or several places can be proposed to place a solution object or the different instances required to solve the level (e.g. for sorting objects two or more storage boxes can be used).

A game scenario is an ordered sequence of scenes with precise descriptions of their setups. All the related information (e.g. number of scenes, selected scenes, order, scenes components and locations...) has to be adapted to the child's profile when starting a new game session. There are various profile variables (e.g. current progress in learning skills, preferences/dislikes, difficulty level of each skill...) and a lot of combinations of elements to set-up a scene. It will be time-consuming and costly to design and develop all the combinations of settings. Therefore, we need to generate dynamically game sessions adapted to each child's profile. The following sections detail our proposal to address this issue.

## 3 Background and Positioning

The motivation for steering adaptivity in serious games is to improve the effectiveness of the knowledge transfer between the game and its players. Several studies tackled the adaptation issue in order to find a balance between the player's skills and the game challenge level. The learning goals to achieve are usually strongly coupled with the gradual personal improvement of a skill set. Generally, adaptive serious games have specialized *ad hoc* approaches where game components are adjusted in order to encourage training of a specific skill.

Research work dealing with adaptivity have different targets (game worlds and its objects, gameplay mechanics, nonplaying characters and AI, game narratives, game scenarios/quests...) [2, 7, 15]. Game scenarios are generally defined

as the global progression within a game level, its initial settings and the logical flow of events and actions that follow [5], whereas game worlds are the virtual environments within which gameplay occurs. In our context, we are focusing on learning game scenarios because each scene to achieve targets a specific skill. Besides, we disregard the flow of events or actions because our game will not embed script-oriented events. The resolution of a scene only requires that the learners find and move objects to their appropriate target locations. Our context partially maps the game world and its object definition in the way that the available objects of scenes can have zero or more instances according to the generation process.

Research work addressing game adaptivity also rely on various methods [2, 7, 15] (e.g. Bayesian networks, ontologies, neuronal networks, rules-based systems, procedural algorithms. . .). The model-driven approach we propose is not currently widespread. Nevertheless, it has been used in instructional design contexts to deal with learning scenarios specification and implementation issues [10].

Reaching beyond skill-driven adaptivity and integrating scenario with world adaptation/generation while the game is running remains a research challenge [11]. There are two approaches to tackle it: (1) during the loading stage of a game session by considering player-dependent information; and (2) in real-time during game playing. Our concern relies on the first approach.

In [1], the authors have proposed a system for generating content highlights the involvement of domain experts (i.e. teachers) to control the content generation. Teachers can select pre-created game objects, add new learning content to them and create relationships between objects. Knowledge about objects and their relationships seems a basis for solving and generating all the appropriate content. It could be a valuable contribution to control the generation of our learning game scenarios by using knowledge on the objects of each scene and their relationships. Such game knowledge should be specified at a high semantic level in order to involve domain experts.

Closer to our concerns, the work presented in [14] proposes a generic architecture for personalizing a serious game scenario according to learners' competencies and interaction traces [6]. The architecture has been evaluated with the objective to develop a serious game for evaluating and rehabilitating cognitive disorders. It is organized in three layers: domain concepts, pedagogical resources and serious game resources. In addition, this proposal allows the generation of three successive scenarios (conceptual, pedagogical and serious game scenarios) according to the three presented layers. As for the validation of the generated scenarios, the authors used an evaluation protocol. For that, experts were involved at first to validate the domain rules, *a priori* of the generator implementation, and then to produce scenarios for specific contexts. These scenarios are compared to the generated ones. Hence, experts guide the requirements specification and validation activities, but they are not directly involved in the generation process.



## 4 A Model-Driven Co-design Process for the Serious Game

Our general concern is the generation of learning scenarios adapted to children profiles while considering the game knowledge. More precisely, we have derived three related challenges:

1. How to make explicit and well defined the domain components (skills, game knowledge, learner model elements. . . ), as well as the mapping and generation rules.
2. How to use these information to drive the generation of adapted learning scenarios.
3. How to involve domain experts in the design and the validation of our serious game.

Model Driven Engineering (MDE) is a research domain promoting an active use of models throughout the software development process, leading to an automatic generation of the final solution. In our case, MDE allows expressing the game description, the learner profile, and the learning scenarios as active models. Hence, adapted scenarios can be dynamically generated by means of model transformations (challenges 1 and 2). These models are expressed in a high level of abstraction so that the participation of domain experts in the design/development activities does not require a technical background (challenge 3). Also, model transformations make it possible to automatically propagate changes of these models to the generated scenarios. Therefore, we can achieve quick feedback from domain experts (challenge 3).

### 4.1 A $3 \times 3$ (Meta-)modeling Architecture

We propose a  $3 \times 3$  metamodel-based architecture: 3-dimensions specification of domain elements to be managed, and 3-incremental perspectives on the resulting scenarios.

The generic domain concepts and relations, required for the generation of scenarios, are defined by three inter-related metamodels (see the top part of Fig. 2): *Learner* metamodel, *Game Description* metamodel and *Scenario* metamodel. The *Game Description* metamodel plays a central role because it describes static game knowledge and relations including those referencing the supported skills. Thus, the *Learner* and *Scenario* metamodels include references to it.

Different models that conform to the presented metamodels are managed (bottom part of Fig. 2). The game description model describes all the real game elements (skills, resources or exercisers, in-game objects. . . ). As for the profile model, it represents a player's (child's) profile. These models are transformed into three target scenarios (objective, structural and feature) that conform to the *Scenario* metamodel. Indeed, we have followed the generation principle from [14] where the final learning game scenario is built after three steps.

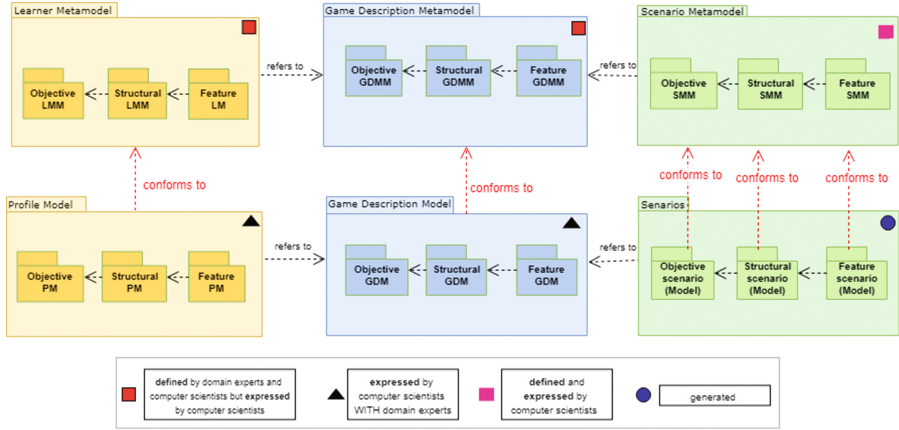


Fig. 2. The proposed 3 × 3 metamodel-based architecture

- *Objective scenario*: it refers to the selection of targeted learning objectives according to the user’s profile. In the *Escape it!* project, this is related to the elicitation of the visual performance skills in accordance with the number of levels to generate, the considered skills and the child’s progression.
- *Structural scenario*: it refers to the selection of learning game exercises or large game components. In our project, we focus on the various scenes where game levels will take place. This scenario specifies correspondences between the selected pedagogical large-grained resources (i.e. scenes) and their targeted skills.
- *Feature scenario*: it refers to the selection of additional inner-resources/fine-grained elements. In the *Escape it!* project, this concerns all objects of a scene. The feature scenario specifies the overall information required by a game engine to drive the set-up of a learning game session.

As illustrated in Fig. 2, each scenario’s perspective has been considered when defining the implied metamodels and expressing models. For example, the generation of an objective scenario considers a relevant subsets of the profile elements (e.g. skills and their levels for a specific child) and the game description elements (the ones representing the skills that are tackled by the game).

## 4.2 An MDE Based Process to Co-design the Serious Game

Figure 3 depicts the co-design process of the proposed serious game. This process involves domain experts and computer scientists to conjointly design and validate the domain elements and rules that are relevant for the generation of adapted scenarios. The *meta-modeling* and *transformation specification* activities are performed by computer scientists because of the required expertise. The remaining activities involve both ASD experts (i.e. with no technical background) and computer scientists.

- *Game analysis*: this activity aims at identifying and expressing the various domain elements, properties, relations and domain rules that are involved in the adaptive generation of scenarios. An application case and other explicit designs (e.g mock-ups, sound effects. . .) can also be expressed.
- *Meta-modeling*: this activity consists in specifying the metamodels that define the static domain elements according to the metamodeling architecture presented in Sect. 4.1.
- *Modeling profiles and game description*: domain experts and computer scientists express together the relevant models by using a dedicated editor.
- *Transformation specification*: this activity is related to the development of the model transformation(s) [12] that allows producing adapted scenarios from a profile and game description source models.
- *Scenarios generation*: this activity applies the aforementioned transformation to the profile and game description source models in order to generate the target inter-related scenarios (i.e. the objective, structural and feature scenarios). After that, the produced scenarios can be integrated into the execution engine with a view to producing a playable prototype of the game.
- *Test and validation*: during this activity, ASD experts and computer scientists make use of the generated prototype in order to verify the relevance, coherence and completeness of the generated scenarios. Besides, this activity deals with the validation of domain rules that drive the generation. Consequently, domain experts can approve these rules or suggest alterations.

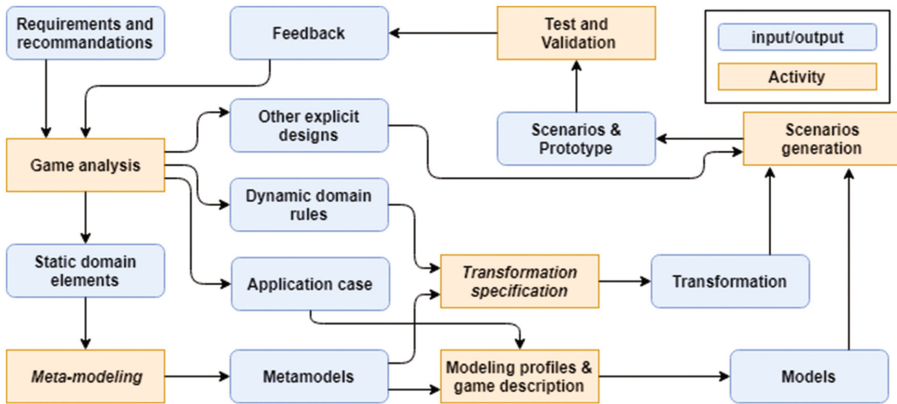


Fig. 3. The co-design process involving domain experts

These activities are part of an iterative process. One can consider at least three iterations focusing respectively on the three incremental scenarios: objective scenario, then structural scenario, and finally the feature scenario. Nevertheless, other iterations may be required for a same scenario’s perspective according to the feedback from “*Test and validation*” activity. Indeed, gaps between experts predictions and the generated scenarios can occur. Generally, the analysis of the

generated scenarios can highlight some misunderstandings within the interdisciplinary team, or some misconceptions about the generation rules. Therefore, re-engineering iterations have to be completed.

## 5 Application

In this section, we describe the application of the proposed co-design process to the presented serious game. This section is structured according to the aforementioned design activities and concerns one design iteration. It is worth noting that the focus here is on the global co-design process rather than on how the model transformations are implemented.

### 5.1 Game Analysis

Collaborative sessions with autism experts led us to identify the detailed description of each supported scene. This includes the various objects to place, hiding elements and solution objects. Furthermore, domain rules to apply when generating a scenario have been specified. Table 1 gives an overview of the main generation rules as well as the elements from the profile and game description models in relation with them.

**Table 1.** The different domain rules and relevant elements according to our  $3 \times 3$ -dimensions metamodeling architecture

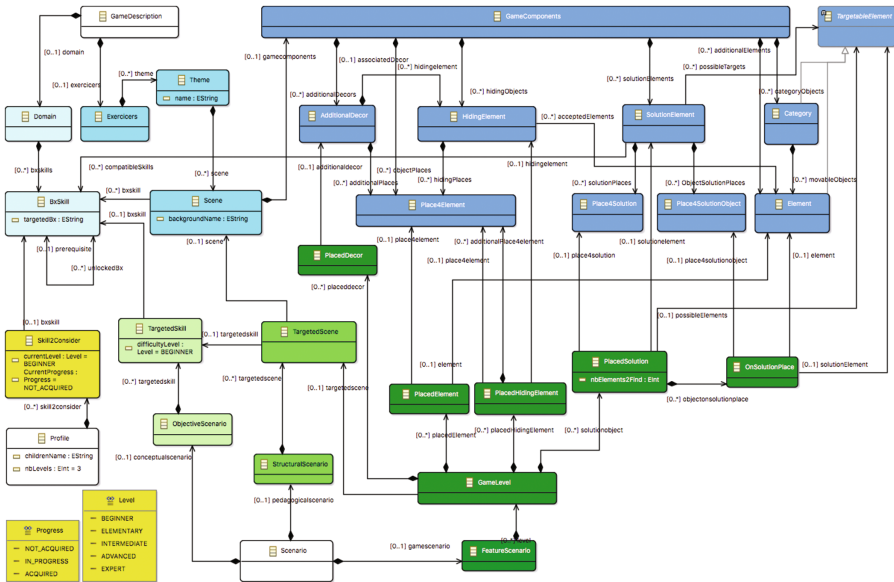
	Game description	User profile	Generation rules for scenarios
Objective scenario	–visual skills to acquire – <i>dependency</i> relations between skills	–acquired or in progress skills –their difficulty level –number of levels to generate	–only skills with <i>parents</i> at ‘ <i>Intermediate</i> ’ level or higher are eligible –80% of targeted skills with a difficulty level less than ‘ <i>Intermediate</i> ’
Structural scenario	–themes and associated scenes – skills targeted by each scene	– <b>themes/scenes to exclude/favor according to child’s preferences/dislikes</b> – <b>history of proposed scenes</b>	–generate different scenes from the same theme
Feature scenario	–background elements, hiding objects, available object places of each scene	– <b>scene objects to exclude/favor according to child’s preferences/dislikes</b> – <b>objects involved in previous sessions</b>	–mappings between each difficulty level and the objects to select and place into the scene

Some mapping rules have been established to guide scenes construction according to the difficulty level. Five difficulty levels have been defined (i.e. Beginner, Elementary, Intermediate, Advanced and Expert). For example, mappings for the ‘*Intermediate*’ level are given below:

- Background elements can appear.
- Hiding objects can appear with 0 or several hidden objects according to their available slots.
- All selectable objects are tied to the problem resolution (no objects for disturbing purposes).

### 5.2 Metamodeling

Recalling from Sect. 4.1, the domain elements and relations required for the adaptive generation of scenarios are structured according to three metamodels (i.e. the *Profile*, *Game Description*, and *Scenario* metamodels). By playing the role of computer scientists and relying on the identified static domain elements, we have used the EMF platform<sup>1</sup> to express the relevant metamodels (see Fig. 4). We have to notice that Fig. 4 depicts all related constructs as one metamodel for better comprehending the inter-metamodels references.



**Fig. 4.** Complete view of the metamodels with variations of colors to discern the different dimensions/perspectives

<sup>1</sup> <http://www.eclipse.org/modeling/emf/>.

A *Scenario* instance contains three inter-related elements: objective, structural and feature scenarios. By following the same decomposition approach, the *Game Description* constructs are decomposed into three subsets that match the scenario’s perspectives: the skills elements (visual skills), the exercises elements (scenes and themes) and the game components associated with a concrete exercise (background, objects, locations...). Some elements from *Exercises* and *Game Components* parts will refer to specific skills elements (e.g. scenes must specify which targeted skills they can deal with). As for the *Profile* constructs, they are limited to elements required for generating the objective scenario. The remaining perspectives are not yet handled by our proposal (they are highlighted with gray color in Table 1).

### 5.3 Modeling Profiles and Game Description

The game description is the first required input model. It has been expressed using a tree-based editor proposed by EMF tooling. Figure 5 shows three different extracts. The root element is a *Game Description* instance. The containment references are naturally represented within the tree-based representation, whereas properties and other references are detailed in the *Properties view* depending on the element being selected.

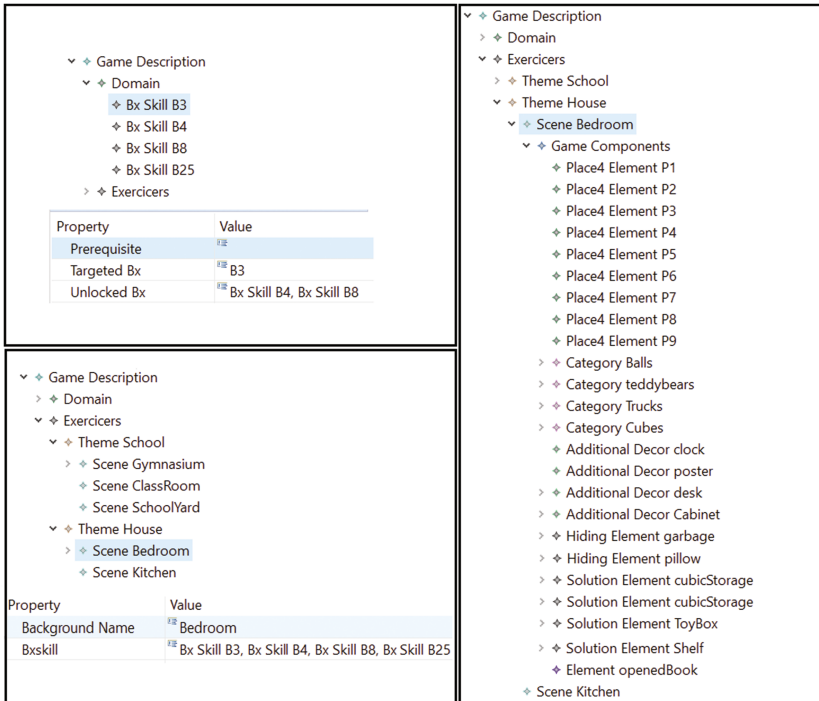


Fig. 5. Partial views of the game description input model

The top left part of Fig. 5 depicts four visual performance skills: B3, B4, B8 and B25 (respectively matching object to image, matching object to object, sorting categories of objects, making a seriation) and their dependency relations. For example, the B3 skill unlocks the B4 and B8 skills (i.e. completing B3 at its highest difficulty allows to progress independently with the learning of the B4 and B8 skills). The bottom left part depicts the description of the game scenes and their container themes. Finally, the right part details the elements involved in the *BedRoom* scene.

In opposition to a unique game description model, several child profiles have been expressed as input models. For that, ASD experts have proposed various fictive profiles but realistic according to them.

#### 5.4 Transformation Specification

The generation of scenarios adapted to child profiles is implemented as a model transformation written in Java/EMF [16]. This transformation is applied to the profile and game description models to allow the successive generation of the three perspectives of an adapted scenario. It is worth noting that the experts requirements related to dynamic domain rules are not easy to implement. In fact, the implemented model transformation uses an external constraints solving library to tackle some very specific generation steps.

By considering an existing procedural context generation taxonomy [17], our proposal to generate scenarios could be regarded as *online* (i.e. during the runtime), *necessary* (i.e. the content has to be correct), *parameterized* (i.e. it takes as an input the game description model), *stochastic* (i.e. randomness is used when several combinations are possible) and *constructive* (i.e. it never produces broken content).

#### 5.5 Transformation Execution

Employing the transformation presented above performs the generation of adapted scenarios. However, interpreting the generated models using basic EMF editors is not appropriate to perform domain rules validation. As a solution, we have implemented a support for integrating the generated scenarios in the Unity-based<sup>2</sup> game engine. This concerns the low level scenario (i.e. feature scenario) and makes it possible to play the related game session. By this mean, ASD experts can carry out effective tests of the game. It is worth noting that the scene depicted in Fig. 1 was generated using the proposed integration support.

#### 5.6 Test and Validation

We have conducted a collective validation session with two ASD experts. We have exploited the generated scenarios (each one corresponds to a specific profile) and then analyzed them by using the game prototype integration support. As a

<sup>2</sup> <https://unity3d.com/>.

feedback, the experts decided to disregard the 80/20 generation rule. This rule stipulates that 80% of the skills referenced by the generated scenario must be at a difficulty level less than ‘*Intermediate*’ against 20% at higher level. Indeed, the experts realized that this rule cannot be satisfied in all possible cases (basically for children not familiar with the game and those at an advanced stage).

On the other hand, the experts have proposed new rules concerning the selection of candidate scenes. The base principle is to diversify the scenes offered to the child while trying to use the same theme. Accordingly, the experts have expressed the rules below. They are cited in order of priority:

- All scenes must be different and belong to the same theme.
- All scenes must belong to the same theme. In addition, two successive scenes must be different.
- All scenes must be different (no constraints on themes).
- Two successive scenes must be different (no constraints on themes).

This design iteration confirms the need to involve experts in a co-design process ranging from requirements elicitation to test and validation. Indeed, relying on the expert’s knowledge is crucial for this type of project whose end users have specific needs. Moreover, the limitation to the requirements and recommendations of ASD experts cannot guarantee a good adequacy of the game with children. Indeed, the experts only become aware of the consistency of their choices through playing the game.

## 6 Conclusion

This paper focuses on the development of a serious game for helping young children with Autistic Syndrome Disorder to learn and generalize visual performance skills. It presents a co-design process that allows ASD experts and computer scientists to express and validate the domain elements and rules involved in the generation/adaptation of learning scenarios. Essentially, the proposed process is iterative and relies on MDE and rapid prototyping.

MDE provides support for adaptive generation of scenarios and allows varying situations proposed to domain experts without significant effort. Indeed, it is possible to express several profiles and apply the same transformation to automatically generate the consequent scenarios. As for rapid prototyping (based on the integration of scenarios in Unity), it allows simulating a real exploitation of the game under-development. Therefore, ASD experts can express more relevant feedback that can be considered in the following iteration.

A perspective of this work relies on change impact analysis to support a rapid generation of the new prototype related to the expressed feedback. This involves managing traceability links between the experts recommendations/requirements and the prototype generation mechanisms. In the same perspective, we intend to re-specify the model transformation responsible for generating adapted scenarios in a more structural and modular manner with a view to determining precisely the fragments impacted by an expressed change.



## References

1. Bieliková, M., Divéky, M., Jurnečka, P., Kajan, R., Omelina, L.: Automatic generation of adaptive, educational and multimedia computer games. *Signal Image Video Process.* **2**(4), 371–384 (2008)
2. Callies, S., Sola, N., Beaudry, E., Basque, J.: An empirical evaluation of a serious simulation game architecture for automatic adaptation. In: *Proceedings of the 9th European Conference on Games Based Learning*, pp. 107–116 (2015)
3. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*
4. Ern, A.M.: The use of gamification and serious games within interventions for children with autism spectrum disorder. B.S. thesis, University of Twente (2014). <http://essay.utwente.nl/64780/>
5. van Est, C., Bidarra, R.: High-level scenario editing for simulation games. In: *Proceedings of the 6th International Conference on Computer Graphics Theory and Applications-GRAPP*, vol. 2011 (2011)
6. Hussaan, A.M., Sehaba, K.: Consistency verification of learner profiles in adaptive serious games. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) *EC-TEL 2016*. LNCS, vol. 9891, pp. 384–389. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45153-4\\_31](https://doi.org/10.1007/978-3-319-45153-4_31)
7. Janssens, O., Samyny, K., Van de Walle, R., Van Hoecke, S.: Educational virtual game scenario generation for serious games. In: *Proceedings of the IEEE 3rd International Conference on Serious Games and Applications for Health (SeGAH 2014)*, pp. 1–8. IEEE (2014)
8. Koegel, L.K., Ashbaugh, K., Koegel, R.L.: Pivotal response treatment. In: Lang, R., et al. (eds.) *Early Intervention for Young Children with Autism Spectrum Disorder*, pp. 85–112. Springer, New York (2016). [https://doi.org/10.1007/978-3-319-30925-5\\_4](https://doi.org/10.1007/978-3-319-30925-5_4)
9. Leaf, R.B., McEachin, J.: *A Work in Progress: Behavior Management Strategies and a Curriculum for Intensive Behavioral Treatment of Autism*. Drl Books, New York (1999)
10. Loiseau, E., Laforcade, P., Iksal, S.: Abstraction of learning management systems instructional design semantics: a meta-modeling approach applied to the moodle case-study. In: Rensing, C., de Freitas, S., Ley, T., Muñoz-Merino, P.J. (eds.) *EC-TEL 2014*. LNCS, vol. 8719, pp. 249–262. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11200-8\\_19](https://doi.org/10.1007/978-3-319-11200-8_19)
11. Lopes, R., Bidarra, R.: Adaptivity challenges in games and simulations: a survey. *IEEE Trans. Comput. Intell. AI Games* **3**(2), 85–99 (2011). <https://doi.org/10.1109/TCIAIG.2011.2152841>
12. Mens, T., Gorp, P.V.: A taxonomy of model transformation. *Electron. Notes Theor. Comput. Sci.* **152**, 125–142 (2006)
13. Partington, J., Analysts, P.B.: *The Assessment of Basic Language and Learning Skills-revised (the ABLLS-R)*. Behavior Analysts (2010)
14. Sehaba, K., Hussaan, A.M.: Goals: generator of adaptive learning scenarios. *Int. J. Learn. Technol.* **8**(3), 224–245 (2013)
15. Sina, S., Rosenfeld, A., Kraus, S.: Generating content for scenario-based serious-games using crowdsourcing. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 522–529. AAAI Press (2014)

16. Steinberg, D., Budinsky, F., Paternostro, M., Merks, E.: *EMF: Eclipse Modeling Framework 2.0*, 2nd edn. Addison-Wesley Professional (2009)
17. Togelius, J., Yannakakis, G.N., Stanley, K.O., Browne, C.: Search-based procedural content generation: a taxonomy and survey. *IEEE Trans. Comput. Intell. AI Games* **3**(3), 172–186 (2011). <https://doi.org/10.1109/TCIAIG.2011.2148116>
18. Whyte, E.M., Smyth, J.M., Scherf, K.S.: Designing serious game interventions for individuals with autism. *J. Autism Dev. Disord.* **45**(12), 3820–3831 (2015). <https://doi.org/10.1007/s10803-014-2333-1>
19. Zakari, H.M., Ma, M., Simmons, D.: A review of serious games for children with Autism Spectrum Disorders (ASD). In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) *SGDA 2014*. LNCS, vol. 8778, pp. 93–106. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11623-5\\_9](https://doi.org/10.1007/978-3-319-11623-5_9)



# Student Drop-out Modelling Using Virtual Learning Environment Behaviour Data

Jakub Kuzilek<sup>1</sup>(✉), Jonas Vaclavek<sup>1</sup>, Viktor Fuglik<sup>1,2</sup>, and Zdenek Zdrahal<sup>1,3</sup>

<sup>1</sup> CTU in Prague, CIIRC, Jugoslavskych Partyzanu 1580/3, 160 00 Prague, Czech Republic  
jakub.kuzilek@cvut.cz

<sup>2</sup> Charles Univ, Fac Edu, Magdaleny Rettigove 4, 116 39 Prague, Czech Republic

<sup>3</sup> Open University, KMl, Walton Hall, Milton Keynes, MK7 6AA, UK

**Abstract.** With the rapid advancement of Virtual Learning Environments (VLE) in higher education, the amount of available student data grows. Universities collect the information about students, their demographics, their study results and their behaviour in the online environment. By applying modelling and predictive analysis methods it is possible to predict student outcome or detect bottlenecks in course design. Our work aims at statistical simulation of student behaviour in the VLE in order to identify behavioural patterns leading to drop-out or passive withdrawal i.e. the state when a student is not studying, but he has not actively withdrawn from studies. For that purpose, the method called Markov chain modelling has been used. Recorded student activities in VLE (VLE logs) has been used for constructing of probabilistic representation that students will perform some activity in the next week based on their activities in the current week. The result is an instance of the family of absorbing Markov chains, which can be analysed using the property called time to absorption. The preliminary results show that interesting patterns in student VLE behaviour can be uncovered, especially when combined with the information about submission of the first assessment. Our analysis has been performed using Open University Learning Analytics dataset (OULAD) and research notes are available online (<https://bit.ly/2JrY5zv>).

**Keywords:** Student Drop-out · Modelling · Virtual learning environment · Markov chains

## 1 Introduction

In the past decade, higher education experiences a massive boom of ICT based education. At present, educators and students extensively use Virtual Learning Environments such as Moodle platform [1]. The ICT based education is further boosted by the introduction of Massive Open Online Courses (MOOCs) platforms such as Coursera [2]. With all these platforms the amount of information about students grows. The possibilities of student data usage for improvement of the education have been investigated in over 200 studies in past years [3].

In 2014 Hlosta et al. [4] proposed two methods for activity analysis: General Unary Hypothesis Automaton and Markov chains. The first method produces set of rules that describe the data. The second generates state transition probabilities from state to state, which represents chances that student change behaviour based on his previous behaviour. The main disadvantage of both methods is the complexity of achieved results.

The idea of previously mentioned work is further extended by Okubo et al. [5]. The authors employed the Markov chain-based method using data from Kyushu University and provided the method as a Moodle analysis module.

Later on, Davis et al. [6] employed Markov chains in the analysis of MOOC data from edX and Coursera courses with over 100,000 students.

Our research focused on the exploration of student behaviour using VLE logs in order to uncover behaviour leading to withdrawal or passive withdrawal of the student. For that purpose, we employed Markov chain modelling [7] on behavioural data available in Open University Learning Analytics dataset (OULAD) [8], which contains the data from a Moodle-like system used at the Open University<sup>1</sup>. Furthermore, the previously used approach [4] has been simplified and the state space of student activities was reduced to 7 possible states, which will be further discussed in Sect. 3.

## 2 Data

The OULAD [8] contains information about 32,593 students visiting 22 Open University courses in years 2013 and 2014. The Open University is largest distance learning institution in the United Kingdom with more than 170,000 students. The typical course has one or more assignments, final exam and has the length of approximately 9 months. OU uses the Moodle-like platform (VLE) to deliver content to students. Usually, course VLE provides a plan of activities for the whole course and it is recommended for students to follow it. For more details see the original paper [8].

The dataset includes data about both students and courses. We focused on data from one course-presentation namely course *FFF* and presentation *2014J*. The course is focused on STEM subject more than 1/3 of the students withdrawn during the semester.

In the following text logs of student VLE activities, the information about first assessment submission and the date of de-registration of the student from the course will be used.

## 3 Methods

In this section, the process of Markov chain model construction will be presented. This can be divided into a transformation of log data to student state data and Markov chain construction itself.

---

<sup>1</sup> <http://www.open.ac.uk/>.

### 3.1 Transforming VLE Logs to States

At first, VLE logs were aggregated on a weekly basis. Next, by combining with course plan (available in OULAD dataset) the student state for every study week has been estimated as follows.

Each activity in VLE has been classified as planned or not based on the course plan. Next, summarization of the planned and non-planned activities for each student and each week has been computed. From the summarized data weekly states have been estimated. Student state in planned activities can fall into the three possible categories: student did nothing ( $O$ ), student did something from the plan ( $E$ ), and student did everything from the plan ( $A$ ). Similarly, unplanned activities can be categorized to: student did nothing ( $O$ ), and student did something out of the plan ( $E$ ). When combined 6 possible states emerged:  $OO, EO, AO, OE, EE, AE$ . For example, state  $OO$  means that student did nothing at all – nothing from a plan and nothing from other (not planned) activities.

Finally, state *Withdrawn*, which represents the fact that student has actively withdrawn from studies, has been added to the set of states resulting in seven possible states, in which every student can be in each week.

### 3.2 Markov Chains

For the construction of Markov chain, we will consider simplifications in order to reduce the problem to the most simple one: (1) the length of a course is infinite; (2) the probability of transition from state in one week to state in another week does not change over time (homogeneity condition of Markov chain); (3) student cannot return to a course when withdrawn; (4) the probability of changing the student state depends only on current week (this is called Markov property [7]). All above leads to the construction of so-called homogeneous absorbing Markov chain [7].

Markov chain is specified by the set of states  $S$ . In our case, these are defined by student states  $S = \{OO, EO, AO, OE, EE, AE, Withdrawn\}$ . From the set of states  $S$  and weekly student states, we can construct the state transition matrix  $P$ , where the entry in  $i$ -th row and  $j$ -th column represents the probability  $p_{ij}$  that a student moves from state  $s_i$  in current week to state  $s_j$  in following week. In addition, the computed transition matrix is reorganized in order to be in the canonical form [7].

Clearly, state *Withdrawn* is absorbing state, that means the student (the process) in this state cannot leave it. Since this state is of the interest we can analyse the resulting transition matrix of Markov chain by means of absorption time [7], which represents the average number of weeks needed to end up in the *Withdrawn* state for the student starting in state  $s_i$ .

## 4 Results

The Markov chain has been constructed for the three cases: (1) the whole cohort of students; (2) students who submitted the first assessment; (3) students who did not submit the first assessment. Following subsections present the results.

### 4.1 Markov Chain of the Whole Cohort

As depicted above, the transition matrix of the whole cohort of students has been constructed. Before the estimation of transition probabilities, the students with states containing a small number of samples (*E0* and *A0*) have been filtered out. The resulting model has 5 states and its transition matrix follows:

$$P_1 = \begin{matrix} & \begin{matrix} 00 & 0E & EE & AE & Withdrawn \end{matrix} \\ \begin{matrix} 00 \\ 0E \\ EE \\ AE \\ Withdrawn \end{matrix} & \begin{pmatrix} 0.66 & 0.29 & 0.02 & 0 & 0.02 \\ 0.13 & 0.75 & 0.09 & 0.01 & 0.01 \\ 0.05 & 0.45 & 0.37 & 0.11 & 0.01 \\ 0.03 & 0.24 & 0.63 & 0.09 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

Since the complexity of graphical representation is high, we decided to work with the transition matrix only. From the matrix  $P_1$  the vector of absorption times  $t_1$  is then computed:  $t_1 = (78 \ 81 \ 81 \ 82)^T$ .

### 4.2 Markov Chain of Submitting Students

Same as in case of the whole cohort the students with states containing a small number of samples (*E0* and *A0*) have been filtered out. Then the students who did submit the first assessment has been selected and the transition matrix was constructed:

$$P_2 = \begin{matrix} & \begin{matrix} 00 & 0E & EE & AE & Withdrawn \end{matrix} \\ \begin{matrix} 00 \\ 0E \\ EE \\ AE \\ Withdrawn \end{matrix} & \begin{pmatrix} 0.62 & 0.35 & 0.02 & 0 & 0.1 \\ 0.13 & 0.77 & 0.08 & 0.01 & 0 \\ 0.06 & 0.59 & 0.33 & 0.01 & 0 \\ 0.01 & 0.031 & 0.59 & 0.07 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

Based on the transition matrix the absorption times vector is computed:  $t_2 = (142 \ 145 \ 146 \ 146)^T$ .

### 4.3 Markov Chain of Non-submitting Students

Lastly the Markov chain for those who did not submit the first assessment has been computed. The students with states containing a small number of samples (*E0*, *A0* and *AE*) have been filtered out and the transition matrix has been constructed:

$$P_3 = \begin{matrix} & \begin{matrix} 00 & 0E & EE & Withdrawn \end{matrix} \\ \begin{matrix} 00 \\ 0E \\ EE \\ Withdrawn \end{matrix} & \begin{pmatrix} 0.95 & 0.03 & 0 & 0.02 \\ 0.51 & 0.41 & 0.03 & 0.05 \\ 0.38 & 0.38 & 0 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

From the matrix  $P_3$  the absorption times vector has been computed:  $t_3 = (50 \ 47 \ 37)^T$ .

## 5 Discussion of Results

When observing resulting transition matrix  $P_1$  of the whole student cohort, one can notice that the probability of student withdrawing from the studies is twice larger for students with no activity in VLE than for student with at least some activity in VLE.

Another interesting observation is that students with no planned activity tend to do nothing from the plan next week (states 00 and 0E) and those who did nothing will do nothing next week in 2/3s of cases. On the other hand, students doing everything from the plan do not tend to withdraw their studies and with high probability will do at least something from the plan next week. Also, they will interact with the VLE with probability 0.96. If we compare the average time to withdraw from the course (time to absorption) students starting in state 00 (doing nothing in the first week) has the lowest time to withdraw.

When we split the data to students who did submit and who did not submit the first assessment, which has been proven to be a good predictor of student success [9], we can observe dramatic changes in the structure of a Markov chain. First, students who submitted the first assignment (transition matrix  $P_2$ ) do not tend to withdraw from studies if they have at least minimal contact with VLE. Second, those who did everything planned tend to do at least something from a plan in the next week. Finally, only those who submitted the first assessment, but then did nothing in VLE have a small probability to withdraw.

What is much more interesting that students who did not submit the first assessment (transition matrix  $P_3$ ) but still interacted with the planned activities in the VLE, tend to withdraw from the studies with probability 0.25. Those, who did not submit the first assessment and did nothing in the VLE tends to do nothing next week (the probability is 0.95). They can be understood as passive withdrawal students– they do nothing, do not actively withdraw and fail the course at the end.

What is important is the fact of homogeneous Markov chains meaning transition probabilities are not changing over time. Of course, it is important to say that in real situation transition probabilities changes over time, but the model called non-homogeneous Markov chain is much harder to interpret. For that purpose, we stayed with the simple model, which can be further extended.

## 6 Conclusion

In this paper, we employed Markov chain modelling for the analysis of student behaviour in VLE and its influence on student drop-out from the course. For the purpose of reproducibility, we used OULAD dataset and all the results and codes are available at <https://bit.ly/2JrY5zv>. The preliminary results showed that we can uncover interesting patterns of behaviour, which might help tutors to uncover conditions leading to student withdrawal. Results also indicated a pattern for passive withdrawal students. Since this is

still work in progress we plan, for example, to include Monte Carlo simulation using computed Markov chains to simulate the behaviour of a single student.

**Acknowledgement.** This work was supported by junior research project by Czech Science Foundation GACR no. GJ18-04150Y.

## References

1. Moodle, H.Q.: Moodle statistics. Moodle HQ (2018). <https://moodle.org/stats/>. Accessed 25 Apr 2018
2. Coursera Inc., “Coursera,” Coursera Inc. (2012). <https://www.coursera.org/>. Accessed 10 Apr 2018
3. Papamitsiou, Z., Economides, A.A.: Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* **17**, 49–64 (2014)
4. Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., Wolff, A.: Modelling student online behaviour in a virtual learning environment. In: Proceedings of the 4th International Conference on Learning Analytics and Knowledge, Indianapolis (2014)
5. Okubo, F., Shimada, A., Taniguchi, Y., Konomi, S.: A visualization system for predicting learning activities using state transition graphs. In: Proceedings of 14th International Conference on Cognition and Exploratory Learning in Digital Age, Vilamoura (2017)
6. Davis, D., Chen, G., Hauff, C., Houben, G.-J.: Gauging MOOC learners’ adherence to the designed learning path. In: Proceedings of 9th International Conference on Educational Data Mining, Raleigh (2016)
7. Norris, J.R.: Markov Chains. Cambridge University Press, Cambridge (1997)
8. Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. *Sci. Data* **4**, 170171 (2017)
9. Wolff, A., Zdrahal, Z., Herrmannova, D., Kuzilek, J., Hlosta, M.: Developing predictive models for early detection of at-risk students on distance learning modules. In: Proceedings of the 4th International Conference on Learning Analytics and Knowledge, Indianapolis (2014)





# A Microservice Infrastructure for Distributed Communities of Practice

Peter de Lange<sup>1(✉)</sup>, Bernhard Göschlberger<sup>2,3</sup>, Tracie Farrell<sup>4</sup>,  
and Ralf Klamma<sup>1</sup>

<sup>1</sup> RWTH Aachen University, Aachen, Germany

{lange,klamma}@dbis.rwth-aachen.de

<sup>2</sup> Research Studios Austria FG, Salzburg, Austria

goeschlberger@researchstudio.at

<sup>3</sup> Johannes Kepler University Linz, Linz, Austria

<sup>4</sup> Open University, Milton Keynes, UK

tracie.farrell-frey@open.ac.uk

**Abstract.** Non-formal learning in Communities of Practice (CoPs) makes up a significant portion of today's knowledge gain. However, only little technological support is tailored specifically towards CoPs and their particular strengths and challenges. Even worse, CoPs often do not possess the resources to host or even develop a software ecosystem to support their activities. In this paper, we describe a distributed, microservice-based Web infrastructure for non-formal learning in CoPs. It mitigates the need for central infrastructures, coordination or facilitation and takes into account the constant change of these communities. As a real use case, we implement an inquiry-based learning application on-top of our infrastructure. Our evaluation results indicate the usefulness of this learning application, which shows promise for future work in the domain of community-hosted, microservice-based Web infrastructures for learning outside of formal settings.

**Keywords:** Learning infrastructures · Microservices  
Communities of Practice

## 1 Introduction

The vast majority of human learning happens outside of formal settings. Learning activities may be quite informal, as found in incidental learning, self-regulated learning and socialization [18]. Some learning may involve more structure or planning, which is generally referred to as non-formal learning [5]. A significant portion of this learning happens in Communities of Practice (CoPs) [20]. These communities are not bound together by an organization, but rather by sharing a common craft or profession, with the desire to learn from each other through knowledge sharing. While only few CoPs have the size and influence to get tools tailored to their needs, the long tail [1] of CoPs does not

possess the resources, such as central hosting infrastructures or shared budget. Consequently, they often adopt publicly available tools (e.g. social software) and re-purpose them according to their needs, mitigating the tools' technical shortcomings through socially enforced usage policies. Thereby, the CoP becomes dependent on the tool provider and also loses control over its data. Even if a CoP manages to establish a centralized infrastructure, this often results in dependencies on single, knowledgeable members or institutions and does not account for dynamic membership, a common characteristic of CoPs.

As a consequence, we claim that a suitable infrastructure for CoPs needs to be decentralized and managed by the community members themselves. It should be easily deployable, extensible and flexible in terms of scalability and accessibility from the outside. The microservice paradigm [14] with loosely coupled services bound together by lightweight protocols fits these demands perfectly. Combined with an underlying peer-to-peer (p2p) network of nodes managed by the CoPs themselves, the microservices should self-replicate through the network according to the community's current needs. Once deployed on the infrastructure, those services and development efforts should remain available, even after the contributing member has left the CoP. Like the ship in the Theseus paradox, a community should be able to persist, even though all of its members have changed over time, as long as there are people willing to engage. Serving as a *community's long term memory*, the infrastructure allows members to learn from their "ancestors", much like we can observe in scientific communities. Just like opening the water tap, using a certain learning environment should be available to every community member at all times. Thus, we propose a *Learning as a Utility* approach, which makes it possible for all community members to equally engage in development, hosting and using learning applications.

The contribution of this work is twofold. First, we describe a technical infrastructure that provides CoPs with an independent, sustainable and flexible way of developing, hosting and sharing their state-of-the-art learning applications on the Web. Second, we present a distributed version of a proven method for inquiry-based learning. Following a design science approach as proposed by Hefner [8], we start by presenting a real-world use case (Sect. 2). From this, we derive the functional requirements of the realized application and the technical design of both our infrastructure and application (Sect. 3). We evaluate our designed application in multiple iterations and discuss the implications (Sect. 4), before presenting related work (Sect. 5) and concluding this contribution (Sect. 6).

## 2 Use Case: Distributed Inquiry-Based Learning

In our use case, a community of young European youth workers are preparing for participation in a European-funded training course on "creative leadership". The participants are an international group, with different levels of experience, from multiple organizations and countries. The team must create learning content that appeals to this diverse group and meets their needs, which is a challenge given the complexity of both creativity and leadership as learning subjects.

In addition, the three trainers providing the course are distributed across different countries and organizations as well, with no possibility to meet beforehand. Since the whole CoP neither shares a geographic location, nor central infrastructure or budget, this use case stands exemplary for the needs and challenges of distributed communities of practice.

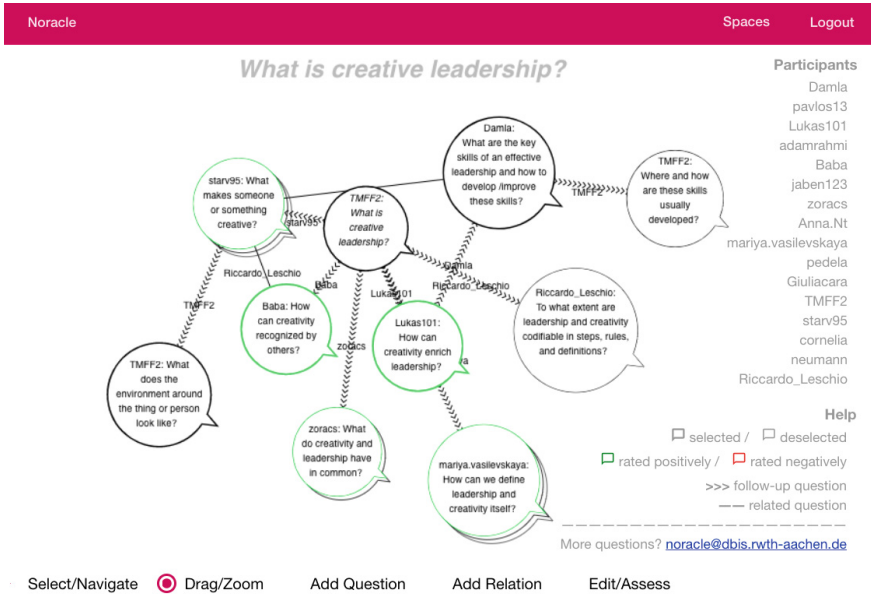
To help establish the boundaries of the participants' knowledge and identify common ground or potential conflicts, the trainers want to find out which questions the participants have about creative leadership and how those questions relate to one another. Specifically, the trainers implement a form of *Question-Based Dialog* called *Noracle* [6] before the training starts, to model and visually represent their common space of ignorance about creative leadership. This special form of inquiry-based learning starts with a central question raised by the trainers, which is then answered by the participants by raising follow-up questions. This way, the *Community Ignorance* becomes visible and the trainers gain insight about what the participants are interested in and their views on the subject. As participants create this *Problem Space*, they document the questions that they have about creative leadership, their assessments of the questions that others stated and any links they perceive between them. In its current form, this involves an on-scene session at the start of the training course, where the community has a limited time-frame to establish their community ignorance by writing down questions they have. A digital version of the concept could be applied already before the community meets. We state the following research questions:

R1: Does a digital version affect the community's perception of their ignorance?

R2: Can a decentralized learning infrastructure be managed by the community?

### 3 Realization of the Distributed Noracle

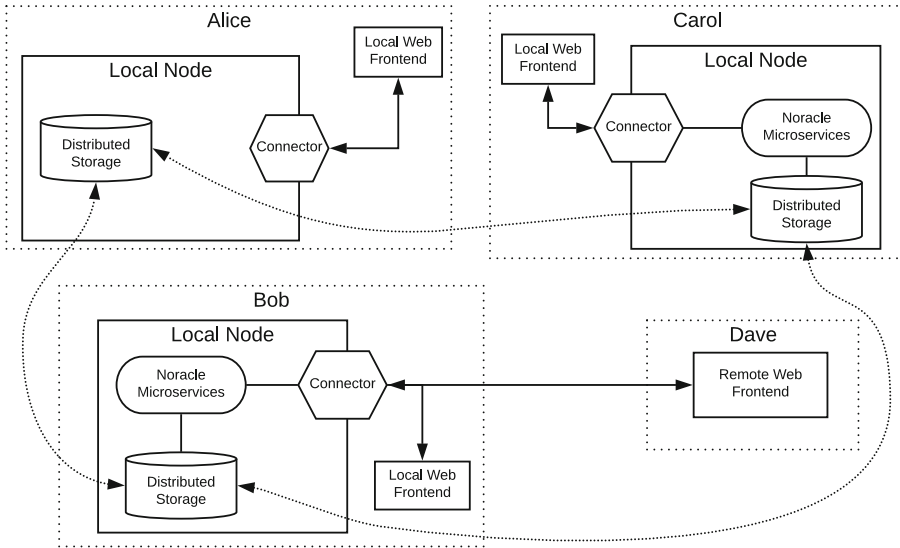
In this section, we describe the realization of a digital and distributed version of the *Noracle* method, an application which we first envisioned in [4]. It fulfills the use case described in the previous section and makes it possible to explore and map community ignorance through question-based dialog, asynchronously and without a formal infrastructure. A space is the main view of the application (shown in Fig. 1). Users can create a space and invite others to the space by sharing an invitation link. The user interface provides a list of subscribed spaces such that users can switch between spaces with two clicks. The space view consists of a canvas displaying the questions and their relations as a graph of speech bubbles. It also features a list of users subscribed to the space and a (collapsible) help section. Below the canvas, users can select their current interaction mode. The "Select/Navigate" mode allows users to define the portion of the graph that is displayed. Selected questions and direct neighbors of selected questions are displayed. If a displayed question that is not yet selected has neighbors that would be displayed upon selecting it, they are symbolically indicated as additional speech bubbles behind the question. In the "Drag and Zoom" mode,



**Fig. 1.** Screenshot of the Distributed Noracle application (Color figure online)

users can move questions around freely, as well as pan and zoom, to either view parts of the graph in detail or get a birds eye view. The “Add Question” and “Add Relation” mode allows users to add questions or relations by clicking on one question (add a question) or two questions (add a relation). Then, a dialog window opens that asks the user to enter the text of the question or the type of the relation. For relations, we allow for both *Follow Up* relations (depicted as small arrows indicating the direction), which is the default type of relation that is created between a new question and its parent question, as well as *Link* relations (depicted as straight lines) that display a certain connection of similar questions, although they are not in a direct *Follow Up* relationship. Finally, the “Edit/Assess” mode enables users to either modify their own questions and relations or to assess the value of questions or relations of others. We use a coloring mechanism that displays the entity according to its overall rated usefulness in a specific color, ranging from green to red.

Figure 2 shows an exemplary usage scenario of a Distributed Noracle session. While *Bob*’s node features the set of microservices that realize the application, *Alice* has decided to start an empty node without any services running on it. This can have several reasons, also including the lack of resources, both in terms of computing power or, especially in mobile settings, energy. *Carol*’s node also contains a set of Noracle microservices, whilst *Dave* has not started a node at all and uses *Bob*’s node to access the remote Web frontend for participating in the collaborative session. As this scenario demonstrates, our framework provides flexible access to the application with several possibilities to join a session.



**Fig. 2.** Exemplary usage scenario of the Distributed Noracle

Depending on the currently available resources of a community member, our framework allows to flexibly start and stop (parts of) applications on a node. Because a central infrastructure is unavailable, this usage scenario does not feature any centralized component, like a master node or a central URL for the Web frontend. Rather, the whole infrastructure is distributed among the community. In the following, we first present a short overview of our technical infrastructure, before we describe the realization of the Distributed Noracle in more detail.

### 3.1 A Distributed Microservice Infrastructure

The technical basis we use for this work is called *las2peer* [10], an open source p2p framework for implementing and hosting Java microservices. Every *las2peer* node in our distributed community learning infrastructure consists of at least two components. The first is the *Distributed Storage*. This storage is partitioned and partly duplicated throughout the network, allowing for a shared, yet synchronized data store. Technically, we base our storage and inter-node communication mechanisms on the *FreePastry* library<sup>1</sup>, a p2p overlay network that provides both a messaging system as well as a *DHT* (Distributed Hash Table) storage system. To ensure privacy, security and data protection, we added end-to-end encryption in form of an *Envelope* system on top of it, ensuring each message and all data stored via the system is encrypted. The second component a node has to integrate is the so called *RESTful Web Connector*. It realizes the communication

<sup>1</sup> <http://www.freepastry.org>.

to the outside, with the capability of routing RESTful calls to an application's (Gateway) interface.

Our framework is capable of load balancing requests to microservices in the entire network, may it be because the service simply does not exist on the local node, or the node is currently overloaded with requests and offloads the task to other nodes in the network. Upstarting services register themselves to the network by calling a specific routine of the node, which then manages their location in the shared storage for all nodes to look-up. This *Sidecar Pattern*-like service registration and discovery ensures that a connector will find the nearest service that currently is flagged as being capable of taking requests. The communication between microservices is realized using a *Message Oriented Middleware* (MOM) that is based on a *Publish & Subscribe Pattern*. Each node registers all running services as subscribers to their corresponding "Service Topic". If a service wants to call another service, it performs a remote method invocation that is sent throughout the network. A node hosting a corresponding service that receives this request will route it to the service, which will handle it. The answer is then sent again in the same way throughout the network. Several timeout mechanisms and an acknowledgment system prevent messages with missing receiver to be forwarded endlessly or messages being answered by multiple services. By using the p2p network to enforce an *Event-driven Architecture* (EDA) of microservice-based applications, we target the needs of fast-changing topologies in CoPs, where complete knowledge of the network might both not be available or even desirable. Nodes can join and leave the network at any time, and the network keeps a persistent shared storage with *Eventual Consistency* (following the BASE model of modern cloud computing architectures [16]), regardless of the current topology. Besides this, it is of course possible for a microservice to implement and maintain its own database, separately of the distributed storage.

### 3.2 Building the Distributed Noracle

The Distributed Noracle application consists of a set of five microservices. A *Space Service* handles the creation of spaces and their members. The *Question Service* takes care of creating and updating questions, while the *Relation Service* does the same for relations. The *Vote Service* handles both votes for questions and relations. Finally, the *Agent Metadata Service* is responsible for storing additional metadata (such as the name) for the members of the CoP. Additionally to these five services, the *Noracle Service* serves as the *Gateway Service* of the application. It differentiates itself from the other microservices that make up the application by providing a RESTful API to the outside. Apart from this, it is implemented as any other microservice in the network, the difference is in terms of semantics (e.g. it does not access the distributed storage facilities). Being called by the connector, it distributes the requests to the set of microservices we just described.

To give a concrete example of inter-microservice communication of the Distributed Noracle application, consider an incoming request for creating a question. This RESTful request would be transferred from the *RESTful Web*

*Connector* to the *Noracle (Gateway) Service*, which would send a request to the *Question Service*. This service in turn would invoke the corresponding *Space Service* for further details, for example if the user is allowed to create a question in this particular space. Upon receiving the answer from the *Space Service*, the *Question Service* would create a new *Question* object in the distributed storage and call the *Relation Service* for creating the corresponding relation between the newly created question and its parent. Finally, the *Question Service* would answer to the *Noracle (Gateway) Service* so that it can forward the HTTP Response to the *Web Frontend*, whether the question has been successfully created. This particular scenario is not necessarily limited to a single node, the microservices can be situated anywhere in the network and it is also neither needed nor desired that a particular microservice knows which instance of the called microservice did handle the request. In the exemplary usage scenario depicted in Fig. 2, if Alice's node receives such a request, it would be distributed throughout the network, because Alice's node does not host any of the application's microservices. Depending on their current load, the request would be processed by the node of either Carol or Bob, and their *Noracle (Gateway) Service* would possibly distribute the just described sub-request again to microservices on other nodes. The flexible scalability of the infrastructure also allows several instances of the same microservice residing at a node, spawning automatically according to the current need. The infrastructure is designed for failure in a way, that non-responding microservices are automatically shut-down and replaced by new instances.

The frontend of our application is based on the Angular 4 framework and it is part of the node, served from the distributed storage. Therefore, we developed a *File Service* that provides a RESTful interface for storing and serving Web frontends directly from the network, removing the need for an additional Web server. Authentication is done using the OpenId Connect *Single Sign-on* (SSO) standard. To provide CoP members with the software needed to start their own node, we created a *Node Package*. It is a small folder that contains an empty node preconfigured to connect to a network via a (configurable) *Seed Node*. It then replicates the microservices of the application via the p2p network and starts them locally. The application and its underlying framework are released as open source software<sup>2</sup>.

## 4 Evaluation

We evaluated our application in four iterations, including one preliminary evaluation, with different types of learning communities. Each evaluation had a certain focus that led to a gradual improvement of the tool. In the following, we describe each of these evaluations in more detail.

---

<sup>2</sup> <https://distributed-noracle.github.io>.

## 4.1 Preliminary Evaluation

In the preliminary evaluation, a Web science research group at a university used a paper mock-up of the Distributed Noracle for questioning current priorities in their research field. The purpose of this evaluation was to determine whether the method could be transferred to a digital space and which features would be required. This community was appropriate because of the shared interest in a topic, diverse levels of experience, and a loose collaborative structure.

**Participants and Procedure:** 8 members of the community took part in the trial. Half of the participants were more experienced members of the team, as determined by whether or not they were supervising PhD students. The other half were PhD candidates or post-doctoral researchers. To represent a shared digital space, the participants worked asynchronously on a large poster in the lab. A general reflection question was posed as the central question in the Distributed Noracle mock-up: “What is the most relevant, open question for social semantics?” Each participant received a differently colored marker to represent her contributions to the poster. As participants added questions, they were also asked to circle questions they supported and draw links between questions to show their relationship. Participants also starred those contributions they thought were most helpful. The evaluation lasted for three days.

**Analysis and Outcomes:** After concluding the exercise, the participants completed a short evaluation on the insights they could draw from looking at the question graph. They also expressed thoughts about the overall value of the proposed artifact. The main outcome of this evaluation was that the tool could help to *structure dialog more efficiently and encourage users to consider broader or new perspectives*, but that *participants need assistance in interpreting the graph*. The need to transfer the process of question-based dialog to a digital space to increase its value was established through this evaluation.

## 4.2 Interface Evaluation

The first evaluation of the digital tool was conducted with participants on an “on arrival” training for participation in the European Voluntary Service (EVS) program. The participants used the Distributed Noracle to consider the future of European youth work in the context of a project planning session. This community was appropriate because of the ill-defined nature of the topics that participants were exploring and the lack of shared infrastructure between them.

**Participants and Procedure:** 7 participants between the age of 20–25 from different European and Erasmus+ partner countries took part in the study. The participants had similar levels of experience in the area of youth work (1–2 years). In this evaluation, the participants worked synchronously. All participants used a given link to access the single-node deployment of the Distributed Noracle. After a project planning session in their face-to-face seminar, the participants joined the space and continued their reflections online. They had a set period of time to explore the application with the general reflection question posed to



them: “What is the future of European Youth Work?” As participants added questions, they were also asked to assess questions they found helpful and create links between different questions to show their relationship. The exercise lasted for approximately 30 min.

**Analysis and Outcomes:** The addition of some analytic features helped users to get a sense for a question’s *importance, quality and validity*. Examples for this are the marking of questions where conflicts are present in red, or darkening the circle that surrounds the topic as more and more contributors agree that the question is relevant. Users made suggestions primarily for improvements related to the interface, as some participants found the layout and animations slightly disorientating. This was mainly due to the prototypical nature of the first iteration and we improved the overall look and feel for the next evaluations.

### 4.3 Technical Evaluation

The second evaluation was conducted with workshop participants of the Joint European Summer School on Technology Enhanced Learning (JTELSS). The purpose of this evaluation was to test the technical features of the tool, in particular the distributed architecture. The community was considered appropriate for a technical evaluation of the learning tool because of their experience with educational software.

**Participants and Procedure:** Approximately 20 people participated in the workshop. First, the participants were given a short introduction to the method of question-based dialog and to the application. As part of this introduction, participants were guided on how to start their own node and join the network. Participants used their own technical devices to launch their nodes. We provided a local seed node the participants could connect to. The participants were then given about 20 min of time to explore the tool. We provided a general starting question in a sample space. Participants were also asked to assess questions they found helpful and create links between different questions to show their relationship. In addition, they were invited to create their own space and invite other participants to join.

**Analysis and Outcomes:** Despite some technical problems, mainly related to firewall restrictions of the local WiFi network, most of the participants were eventually able to connect their node to our on-scene network. Participants not able to start their own node used other participant’s nodes to join the problem space, and thus were able to participate as well. This proved the capability of starting ad-hoc Distributed Oracle networks within a community. The data we received from this evaluation was afterwards used to improve the application, leading to a more stable version used in our pedagogical evaluation.

### 4.4 Real-World Pedagogical Usage Evaluation

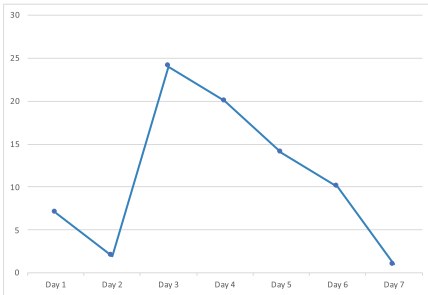
The third evaluation was conducted with the community described in Sect. 2. Participants of an European training course on creative leadership were invited

to participate in an experiment using the Digital Oracle to help prepare for the course and get a sense of the participants' existing knowledge gaps. The purpose of this evaluation was to test the application in a real asynchronous and distributed setting, adding monitoring data to the qualitative verbal and written data.

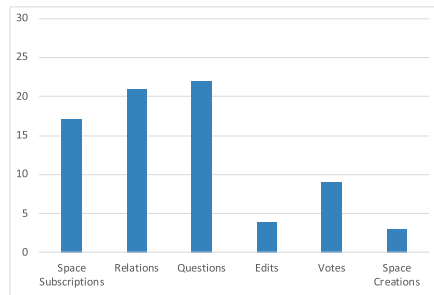
**Participants and Procedure:** 34 participants took part in the evaluation. The participant group was diverse, with different nationalities, levels of experience and knowledge about the subject of the training course, *Creative Leadership*. One week before the training course, participants were notified via email that an "experiment" would be taking place, using a beta version of an application to help prepare for the training. They were informed that their participation in the experiment was completely voluntary, but that it would help to establish what participants found most confusing or difficult about the concept of creative leadership. They received information on how to join the Distributed Oracle and were invited to contribute questions to a specific reflection question related to the training course. Since the participants were locally distributed with prior contact only via email, we created an artificial distributed setting by creating a network of nodes at a university. We provided a URL to the participants that automatically distributed them to their specific node. This created a scenario where each participant had her own node, without the actual need for a technical setup procedure that would have been unfeasible for this particular evaluation, especially regarding the evaluation of the results. After the first 48 hours, participants were asked via email to review the questions that other participants had posted so far once again and evaluate how important or useful they are to the over-all discussion. Once the participants arrived at the training course, the entire trainer team and the trainees participated in an analysis of the question graph and an evaluation of the tool's features. The evaluation included three items: *What insights can you draw from the graph? What features or functions might improve the value of this tool for you? In which situations could you imagine to use it?* Each individual had five minutes to review the graph and to take some notes. Then, the facilitator gathered the insights in a plenary session, during which the participants' statements were also clustered according to their shared theme.

**Analysis and Outcomes:** With regard to the insights that could be drawn from the graph, the group found it quite easy to see what is important, such as focusing on the development of creative skills. They noticed that many questions related to this topic in some way. There was a considerable agreement about the importance of these types of questions (as indicated by the green color). They also realized that they had taken a *very individualistic perspective on creativity and leadership*, with very few questions having to do with the social aspect of creative development. This type of reflection can be mainly contributed to the graph-like structure of the problem space, with its highlighting of importance capabilities. The way that questions were formulated allowed the participants to differentiate between questions related to defining creativity and questions related to the process of developing or improving creativity. Features that

participants felt were important to develop had to do with analytic features to help uncover other types of insights or consequences. For example, only one trainee had noticed that similar questions were repeated several times in the graph. In addition, a third of the participants said that they would find it helpful if there was a way of knowing exactly how many people or a percentage of people found a question useful. All of the participants and the trainer team felt that the tool would be improved by having a way of visualizing what insights or consequences could be drawn. The trainees agreed that the tool helped establishing the interests of a group in advance, which is useful in a variety of settings. The training team remarked, that instructions were extremely important in helping the participants to know how to use the application. Especially with new users, facilitation could be very useful in helping to maintain the quality of the space by demonstrating question-asking and some of the application’s additional features. The training expressed the usefulness of the application as a preparatory exercise for a training course, workshop or seminar.



**Fig. 3.** Activities over time



**Fig. 4.** Activities by type

Additionally to our previous evaluations, we monitored the complete network for user activities [17]. Figure 3 shows the relevant activities monitored during the one week period we had the network running for this evaluation, while Fig. 4 shows the complete number of (selected) monitored activities per type. We started the monitoring the day we sent out the invitation mail, while we asked the participants to start their 48h collaboration phase on the beginning of day three. As one can see, activity is high between day 3 and 5, while it declines afterwards. Still, the number of recorded activity before and after this “official” trial phase shows the intrinsic motivation participants had to visit the problem space, an important factor for learning activities in self-regulated learning scenarios. Another interesting observation we made during analyzing the monitoring data was, that the average question depth was 1.9, meaning that on average a question was about two questions away from the seed question. We perceive this as another indicator of the usefulness of the graph-based visualization, since most questions did not connect directly to the seed question, but

to follow-up questions, demonstrating the evolving awareness of the community ignorance, represented by the growth of the graph.

## 4.5 Discussion

Improvements proposed by users mostly dealt with the interface and analytic features, such as additional ways of visualizing other aspects of the dialog by making nodes larger or smaller, allowing for certain questions to be marked as “resolved” and additional ways of linking questions. Most of the users in all three evaluations said that such a tool can be useful in the planning stages of a project and at the beginning of any complex task or assignment to gain orientation. In addition, participants saw affordances for structuring group- and teamwork in schools.

The trainer team of the real-world pedagogical usage evaluation stated they were able to save considerable time in gathering important information on the trainees’ expectations and knowledge. In a typical training scenario, a half day would have been spent on these types of abstract questions about the program. In this case, it only took 45 min of analyzing the resulting question-graph to achieve an even better result. In addition, starting the process in advance seemed to have the effect that the group took the exercise more seriously, which led to these better results. Possible reasons for this mentioned by the trainers were that when the method is used in face-to-face settings, the participants are naturally distracted by the person they have in front of them. The tendency to move towards providing answers or advice makes it more difficult to keep them on task. Working asynchronously with the participants appeared to have resolved this as it was not necessary to always repeat that the participants should only ask questions.

From the technical point of view, due to their prototypical nature, the evaluations showed potential weak points of our application, such as the stability and ease of starting a node. While we were able to solve many technical challenges during and after the technical and pedagogical evaluation, we are still working on improving both points. Nevertheless, all three different evaluation scenarios proved that our prototype is already applicable in real-world usage scenarios.

## 5 Related Work

Question asking is seen as one of most important skills for innovation, since it contributes to lateral thinking and thus better problem solving [19]. Question-based dialog is viewed as a specific type of a sense-making tool that is also *representation-centric* [11]. To help structure discourse analysis, computational linguistics has offered frameworks to examine collaborative sense-making in virtual environments [9]. For example, argumentation platforms offer a representation-centric approach to collaboration. Contributions are visually represented, categorized as issues, claims, premises and evidence, with modifying functions to

support or refute other constituents of the argument. Cohesion graphs of discussion threads, which represent contributions as nodes at different levels, can examine lexical chains in discourse analysis to understand influence on conversation and identify key issues in conversation. Related works in this domain mostly deal with the issue of how face-to-face scenarios differ from online discussions and how to aggregate community knowledge [12]. Instead of representing *knowledge* in the form of arguments, the Distributed Oracle examines the *gaps* in community knowledge in the form of questions.

The question of system maturity, flexibility and also interoperability is still an active research area [15]. The idea of using p2p-based systems for sharing of educational resources came up first with the creation of EDUTELLA [13], a network for exchanging information about learning objects. Recent development in this area is the InterPlanetary File System [3] project, which describes itself as a *peer-to-peer hypermedia protocol* and shares the concern for *increasing consolidation of control [on the Web]*. Related development approaches have been characterized as p2p cloud computing [2] and edge-centric computing [7]. Despite the high research activity in this domain, we did not find any recent approaches that focus on supporting CoPs with self-managed, decentralized infrastructure. Forums, blogs and wikis are still the most commonly adopted tools for CoPs that need to accommodate geographically distributed participants at scale. However, they do not preserve a representation of contributions that can be elaborated or amended as the community changes, making them harder to sustain for CoPs.

## 6 Conclusion and Future Work

In this paper, we presented both a microservice-based Web infrastructure for distributed learning communities and an application of it in form of an inquiry-based learning tool for CoPs. We followed a design science approach and incrementally tailored our application to the needs of the community, according to the outcome of each evaluation. Our approach concentrated on taking into account the specific attributes of CoPs, like temporal and spatial dynamics. By consequently addressing these attributes, we support CoPs in their efforts to share and acquire knowledge. As information remains available throughout the communities' existence and services evolve continuously at the same time, our infrastructure ensures sustainability and adaptability, aptitudes we reckon to be crucial in the development of a more democratic and egalitarian Web.

In future work, we want to improve our distributed monitoring by ways of providing this information to the community. One particular approach we are working on is the introduction of social learning bots that guide the users through the problem space, tailoring themselves to the user by analyzing the previously monitored usage data. Furthermore, a feature for checking similar questions and also tracking how often they arise could be useful. We are also working on a way how to visualize if a question has been resolved. Finally, we are investigating ways of improving the underlying framework to be even more easily manageable by CoPs. In particular, the switch from the microservice paradigm to a "serverless", Function as a Service (FaaS) supporting platform seems worth investigating.

**Acknowledgments.** This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 687669 (WEKIT) and from the European Union's Erasmus Plus programme, grant agreement 2017-1-NO01-KA203-034192.

## References

1. Anderson, C.: The long tail: why the future of business is selling less of more. Hyperion (2006)
2. Babaoğlu, O., Marzolla, M.: The people's cloud. *IEEE Spectr.* **51**(10), 50–55 (2014)
3. Benet, J.: IPFS - content addressed, versioned, p2p file system. arXiv preprint [arXiv:1407.3561](https://arxiv.org/abs/1407.3561) (2014)
4. de Lange, P., Farrell-Frey, T., Göschlberger, B., Klamma, R.: Transferring a question-based dialog framework to a distributed architecture. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 549–552. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_60](https://doi.org/10.1007/978-3-319-66610-5_60)
5. Eshach, H.: Bridging in-school and out-of-school learning: formal, non-formal, and informal education. *J. Sci. Educ. Technol.* **16**(2), 171–190 (2007)
6. Farrell-Frey, T., Gkotsis, G., Mikroyannidis, A.: Are you thinking what i'm thinking? representing metacognition with question-based dialogue. In: ARTEL 2016, pp. 51–58 (2016). <http://ceur-ws.org/Vol-1736/>
7. Garcia Lopez, P., et al.: Edge-centric computing: vision and challenges. *SIGCOMM Comput. Commun. Rev.* **45**(5), 37–42 (2015)
8. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. *MIS Q.* **28**(1), 75–105 (2004)
9. Iandoli, L., Quinto, I., De Liddo, A., Buckingham Shum, S.: On online collaboration and construction of shared knowledge: assessing mediation capability in computer supported argument visualization tools. *JASIST* **67**(5), 1052–1067 (2016)
10. Klamma, R., Renzel, D., de Lange, P., Janßen, H.: las2peer - a primer. ResearchGate (2016). <https://doi.org/10.13140/RG.2.2.31456.48645>
11. McLoughlin, C., Patel, K., O'Callaghan, T., Reeves, S.: The use of virtual communities of practice to improve interprofessional collaboration and education: findings from an integrated review. *J. Interprof. Care* **32**(2), 136–142 (2018)
12. Meyer, K.A.: Face-to-face versus threaded discussions: the role of time and higher-order thinking. *JALN* **7**(3), 55–65 (2003)
13. Nejdli, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., Risch, T.: EDUTELLA: A p2p networking infrastructure based on RDF. In: WWW 2002, pp. 604–615. ACM (2002)
14. Newman, S.: Building Microservices: Designing Fine-Grained Systems. O'Reilly Media Inc, USA (2015)
15. Ochoa, X., Ternier, S.: Technical learning infrastructure, interoperability and standards. In: Duval, E., Sharples, M., Sutherland, R. (eds.) Technology Enhanced Learning, pp. 145–155. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-02600-8\\_14](https://doi.org/10.1007/978-3-319-02600-8_14)
16. Pritchett, D.: BASE: an acid alternative. *Queue* **6**(3), 48–55 (2008)
17. Renzel, D., Klamma, R., Jarke, M.: IS success awareness in community-oriented design science research. In: Donnellan, B., Helfert, M., Kenneally, J., VanderMeer, D., Rothenberger, M., Winter, R. (eds.) DESRIST 2015. LNCS, vol. 9073, pp. 413–420. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-18714-3\\_33](https://doi.org/10.1007/978-3-319-18714-3_33)

18. Schugurensky, D.: The forms of informal learning: Towards a conceptualization of the field. Technical report, Centre for the Study of Education and Work (2000)
19. Sloane, P.: The Leader's Guide to Lateral Thinking Skills: Unlock the Creativity and Innovation in You and Your Team. Kogan Page (2017)
20. Wenger, E.: Communities of Practice: Learning, Meaning, and Identity. Learning in Doing. Cambridge University Press, Cambridge (1998)



# Multimodal Analytics for Real-Time Feedback in Co-located Collaboration

Sambit Praharaj<sup>1</sup>(✉) , Maren Scheffel<sup>1</sup> , Hendrik Drachslers<sup>1,2,3</sup> ,  
and Marcus Specht<sup>1</sup> 

<sup>1</sup> Open Universiteit, Valkenburgerweg 177, 6419AT Heerlen, Netherlands

<sup>2</sup> DIPF, Schloßstr. 29, 60486 Frankfurt am Main, Germany

<sup>3</sup> Goethe Universität, Robert-Mayer-Str. 11-15, 60629 Frankfurt am Main, Germany  
{sambit.praharaj,maren.scheffel,hendrik.drachslers,marcus.specht}@ou.nl

**Abstract.** Collaboration is an important 21st century skill; it can take place in a remote or co-located setting. Co-located collaboration (CC) is a very complex process which involves subtle human interactions that can be described with multimodal indicators (MI) like gaze, speech and social skills. In this paper, we first give an overview of related work that has identified indicators during CC. Then, we look into the state-of-the-art studies on feedback during CC which also make use of MI. Finally, we describe a Wizard of Oz (WOz) study where we design a privacy-preserving research prototype with the aim to facilitate real-time collaboration in-the-wild during three co-located group PhD meetings (of 3–7 members). Here, human observers stationed in another room act as a substitute for sensors to track different speech-based cues (like speaking time and turn taking); this drives a real-time visualization dashboard on a public shared display. With this research prototype, we want to pave way for design-based research to track other multimodal indicators of CC by extending this prototype design using both humans and sensors.

**Keywords:** Collaboration · Feedback · CSCL  
Intervention · Multimodal indicators · Multimodal learning analytics

## 1 Introduction

Collaboration is an important skill in the 21st century [10]. It can take place in different settings and for different purposes: collaborative meetings [17, 36, 38], collaborative problem solving [34], collaborative project work [7, 8], collaborative programming [15] and collaborative brainstorming [37]. Some are in co-located and some in remote settings. “The requirement of successful collaboration is *complex, multimodal, subtle*, and learned over a lifetime. It involves *discourse, gesture, gaze, cognition, social skills, tacit practices*, etc.” [emphasis added] [35]. Moreover, in each context, the indicators of collaboration vary. For instance, in collaborative programming pointing to the screen, grabbing the mouse from the partner and synchrony in body posture are relevant indicators for



good collaboration [15]; whereas in collaborative meetings gaze direction, body posture, speaking time of group members are more relevant indicators for good collaboration quality [17, 36, 38]. Thus, it is essential to understand what the different types of collaboration and their purpose are and what are the relevant indicators. These indicators help to formulate the intervention or feedback mechanism to facilitate collaboration [2, 5, 30]. Moreover, engaging in a collaborative task does not essentially build collaborative skills [12]; rather on-time feedback encourages self-reflection [23]. The type of feedback is also dependent on the goal of the task which can be to evaluate collaboration as a process [2] or collaboration as an outcome (indicated by learning gain) [30] or both [30]. To understand this in-depth, we have formulated two research questions:

**RQ 1:** What *collaboration indicators* can be observed and are relevant for the *quality* of collaboration during CC?

**RQ 2:** What are the state-of-the-art *feedback* mechanisms that are used during CC?

There has been a dearth of studies on automated multimodal analysis in non-computer supported environments [40]. Considering the time and effort required to build a sensor-based automated system which can also give real-time feedback, we chose to create a WOz research prototype which can integrate human observers and existing sensor technology. This enables us to study different CC settings with a variety of multi-source multimodal indicators coming from automated sensors as well as human observers.

The remainder of the paper is structured as follows: in the related work (Sect. 2) section we answer RQ 1 and RQ 2; it is followed by an explanation of our prototype design based on the WOz study (Sect. 3); this is followed by a discussion (Sect. 4) of the answers to our research questions; finally, a conclusion (Sect. 5) is drawn and we throw some light on future work and open questions to be answered.

## 2 Related Work

In this section, we will first analyze related work according to the different indicators used during CC from multiple modalities; and secondly review the different feedback mechanisms used during CC.

### 2.1 Multimodal Indicators During Co-located Collaboration

Different categories of verbal and non-verbal indicators have been used in the literature to measure collaboration quality ranging from tangible interaction, different speech-based cues, to gaze and eye interaction. Schneider and Blikstein [27] used Tangible User Interface (TUI) for pairs of students to predict learning gains by analyzing data from multimodal learning environments. They tracked the gesture and posture using a Kinect Sensor<sup>1</sup> (Version 1) which can

---

<sup>1</sup> An integrated sensor tracking simultaneously infrared, depth, audio and video.

track the posture and gesture of a maximum of four students at a time based on their skeletal movements. They found that the *hand movements* and *posture movements* (coded as active, semi-active and passive) are correlated with learning gains. The more active a student is, the higher is the learning gain. Even the number of transitions between these three phases was a strong predictor of learning. Students who used both hands showed higher learning gains. Some of the activities that were logged by the TUI, like the frequency of opening the information box in the TUI can be correlated with learning gain. All these features were fed into a supervised machine learning framework to predict learning gain. Similarly, Martinez-Maldonado et al. [21] used TUI indicators for group work based on the log data generated and the gesture and posture of group members around the TUI.

Other works detected non-verbal cues during collaboration without a TUI. Stiefelhagen and Zhu [36] tried to detect the impact of *head orientation* on the gaze direction in a group round table meeting with four members. They found that on an average 68.9 % of the time head orientation can estimate gaze direction. Moreover, attention focus of group members can be easily predicted 88.7 % of the time using head orientation as the only input. Similarly, Cukurova et al. [7] performed an experiment on 18 members in six groups of three members each to detect non-verbal cues of collaboration using human observation. *Hand position* (HP) and *head direction* (HD) was a good predictor of competencies in Collaborative Problem Solving (CPS). They extended this work and formed the NISPI framework [8] using HP and HD as non-verbal indicators. These indicators were obtained during a prototype design by students (11–20 years old) using the Arduino toolkit. Then, they were coded for each student as: 2 (*active*) if a student is interacting with the object for problem solving, 1 (*semi-active*) if the head of the student is directed towards an active peer and 0 (*passive*) for all other situations. Using this coding, different collaboration dimensions like *synchrony*, *individual accountability* (IA), *equality* and *intra-individual variability* (IIV) were formed. High competencies of CPS was detected if high levels of synchrony, IA and equality is detected in the groups.

Speech-based cues are an integral part of any collaborative task. Lubold and Pon-Barry [19] found that *proximity*, *convergence* and *synchrony* are different types of coordination cues obtained from the speech features (like intensity, pitch and jitter) of the pair of students collaborating. It helped them to detect rapport between group members. It was observed from correlation analysis that proximity, convergence and synchrony measured using pitch can be a good predictor of rapport between the group members during collaboration. Students also self-reported rapport which was compared and collaboration levels were determined. Bassiou et al. [4] assessed collaboration among students solving math problems automatically. They used *non-lexical speech* features; thereby, preserving the privacy. They used a combination of manual annotation and Support Vector Machine (SVM) to predict the collaboration quality of the group. Types of collaboration marked are: Good (all 3 members are working together and contributing to the discussion), Cold (only two members are working together),

Follow (one leader is not integrating the whole group) and Not (everyone is working independently). This coding was based on two types of engagement: simple (talking and paying attention) and intellectual (actively engaged in the conversation). They found that the combination of the speech-activity features (i.e., *solo duration*, *overlap duration of two persons*, *overlap duration of all three persons*) and speaker-based features (i.e., *spectral*, *temporal*, *prosodic* and *tonal* features of speech) are good predictors of collaboration. Simple indicators like the speaking time of each member can also be a good indicator of collaboration [2, 5]. Even a mixture of verbal and non-verbal indicators along-with *physiological signals* like skin temperature [24] can be a good collaboration indicator [18, 20].

Besides, *eye gaze* can be an indicator of collaboration quality. Some researchers [16, 25, 28] while using eye gaze analysis found that (JVA) *Joint Visual Attention* (i.e., the proportion of times gazes of individuals are aligned by focusing on the same area in the shared object or screen) is a good predictor of the quality of collaboration of a group which is reflected by the groups performance. Moreover, Schneider and Pea [28] showed that JVA can be used as a reflection mechanism in remote settings to show each student their partners gaze patterns in real-time to improve collaboration. Schneider et al. [30] got the same results by replicating the experiment in a co-located setting. The work by Schneider and Pea [29] used JVA, network analysis and machine learning to determine different dimensions of a good collaboration like *mutual understanding*, *dialogue management*, *division of task*, *signs of coordination* as outlined by Meier et al. [22].

Moving on to the different purposes in which collaboration has been studied, Spikol et al. [33, 34] studied collaborative learning specifically in the context of Collaborative Problem Solving (CPS). They tracked the distance between *hand movements* and *faces* of group members. Later the recorded video streams were coded by experts with 0 (for passive), 1 (for semi-active) and 2 (for active) based on different combinations of head and hand positions for training the machine learning classifier for predicting the quality of collaboration. Recent work by Chikersal et al. [6] dives deep into the deep structure of collaboration in dyads. They found that synchrony in facial expressions correlated with collective intelligence of the group but not significantly correlated with the synchrony of electrodermal activity of members. Another work by Grover et al. [15] studied CPS in a pair programming context based on a pilot study. They captured data from different modalities (i.e., video, audio, clickstream and screen capture) unobtrusively using Kinect. For initial training of the classifiers using machine learning, experts coded the video recordings with three annotations (i.e., High, Medium and Low) when they found evidences of collaboration between the dyads. These evidences include *pointing to the screen*, *grabbing the mouse* from the partner and *synchrony in body position*. Later this classifier could predict the level of collaboration.

Moreover, post-hoc coding with the help of human coders has been an effective method followed for a long time to detect different indicators of collaboration. Davidsen and Ryberg [9] videotaped the work of pairs making a collaborative

**Table 1.** Overview of studies on co-located collaboration.

References	Indicators	Goal
[27]	Hand movements, posture & TUI logs	Post-hoc analysis of indicators on learning
[30]	Joint Visual Attention (JVA)	JVA indicates learning
[34]	Distance between hands & faces	Extraction of multimodal features during collaboration
[15]	Pointing, body position & grabbing mouse	Post-hoc classification of collaboration
[2]	Total speaking time	LED display to regulate audio participation in real-time
[37]	Number of ideas	Real-time metaphorical feedback to support CB
[5]	Total speaking time	Conversation clock will regulate the equity of conversation in real-time
[8]	Hand position and head direction	Build a non-verbal indicator framework for collaboration
[19]	Intensity & pitch of sound, self reports	Detect collaboration levels based on rapport obtained from audio cues & self-reports
[4]	Speech overlap duration, no overlap duration, spectral, temporal, prosodic & tonal speech features	Predict collaboration quality from audio cues
[9]	Dialogue, gesture, posture & gaze	Detect indicators of collaboration from videotaped recordings of collaboration tasks
[26]	Eye contact, posture & amplitude of voice	Detect indicators of collaboration from videotaped recordings of collaboration tasks

discussion around a touch screen measuring “The size of one meter”. The pair was trying to translate the design from graph paper to the touch screen to measure one meter. They found that *body movements*, *language* and *gestures* can be helpful to discover different facets of collaboration. Similarly, Scherr and Hammer [26] observed videotaped groups and identified four clusters based on the collaborative behaviour from both verbal and non-verbal indicators (like *eye contact with peers*, *straight posture*, *clear and loud voice*, etc.). Besides, some works [32, 37] considered *epistemological* aspects of collaboration during brainstorming where the number of ideas generated by each member was the indicator of quality of collaboration. Detecting individual attention levels in classroom from the responses to questions (i.e. epistemological) is also common [39].

In summary, collaboration indicators can vary from non-verbal, verbal, physiological to log files obtained from shared objects like TUI or computers. It depends on the context. Table 1 shows the overview of the multimodal indicators detected. We can find two types of co-located collaboration indicators, i.e., *social* (verbal, non-verbal and physiological) and *epistemological* (logs, ideas).

## 2.2 Feedback During Co-located Collaboration

Using these multimodal indicators, different feedback mechanisms have been developed in the past to facilitate CC. Kulyk et al. [18] designed a mechanism to give real-time feedback to participants in group meetings (with 4 members) by analyzing their speaking time and gaze behaviour. The feedback was in the form of different coloured circles representing attention from other speakers measured by eye gaze, speaking time and attention from listeners. This feedback was projected on the table in-front of where each participant was sitting using a top-down projector. They performed both quantitative and qualitative evaluation to evaluate the effect of the feedback: the feedback was accepted as a positive measure by most group members; use of feedback had a positive impact on the behaviour of group members as they had a balanced participation and improved eye gaze. Terken and Strum [38] used a similar setting and feedback mechanism; they discovered that the feedback on speech increased the equity of participation in the group. But, surprisingly feedback on gaze behaviour had little effect on the interaction pattern of group members. Similarly, Madan et al. [20] used sensors to capture nodding, speech features and galvanic skin response of dyads and built a real-time group interest index. This group interest index helped them to drive a real-time feedback. This feedback showed some group characteristics in different modes: individual PDA feedback, personal audio feedback, haptic feedback in the shoulder and public shared projected display. They studied these group characteristics in different contexts like speed dating and brainstorming sessions.

Some simpler versions of feedback which leverage the audio cues (like speaking time) during collaboration have proved effective in the past. For instance, Bachour et al. [2] performed an experiment to measure audio participation where each group (with 3–4 members) performed a task around a smart table. It gave them real-time feedback during the task by glowing different coloured LED lights for each member. The number of LED lights that glowed for each colour denoted the total speaking time for that member. They found that a real-time feedback helped to maintain the equity of audio participation among the members. Another similar approach was used by Bergstrom and Karahalios [5] with the help of a conversation clock. In this clock, different coloured concentric rings represented spoken participation of each member in the 4 member group. The bars and the dots in the ring denoted the length of conversation and periods of silence respectively.

Moving on to the epistemological aspect of collaboration, Tausch et al. [37] used an intuitive metaphorical feedback moderated by human observers during *collaborative brainstorming*. Three members in each group performed the task. The group members were supposed to discuss a certain topic and their collaboration was measured by the number of ideas generated. A comparison metric for collaboration such as a baseline was calculated as the average number of ideas generated by all members. Using this baseline, each group member was marked as below average or above average depending on the number of ideas generated by each member. Then the human observers controlled the public shared display

which showed a *metaphorical garden*. Each group member was represented by a flower and the group was displayed as a tree with leaves, flower and fruit. The growth of the flower and the tree symbolized the participation (measured by the contribution of ideas) of the individual and the group respectively. More balanced participation was shown by a well grown tree with leaves, fruits and flower. If a group was having unbalanced participation for a long time then lightning flashes were shown in the group garden. Another example of feedback during collaborative brainstorming was implemented by Shih et al. [32]. It supports collaborative conceptual mapping to discuss a topic and organize the ideas.

Besides the use of visual and haptic feedback was effective in some collaboration tasks around a TUI. Anastasiou and Ras [1] gave real-time textual and haptic feedback to each group consisting of 3 members working around a TUI. The group members were needed to use different objects and find the desired power consumption using the TUI. At the end, they used a questionnaire and found that most participants of the experiment favoured the use of both visual and haptic feedback over audio feedback. Martinez-Maldonado et al. [21] used a TUI and gave real-time feedback on group performance for the teachers in tablets so that they can intervene when needed and can also make a post-hoc reflection after the task is over.

Use of external sensing devices to facilitate collaboration during meetings has proved its worth before. Kim et al. [17] used a sociometric badge<sup>2</sup> which acted as a meeting mediator to capture audio and postures during meetings of 4 members in one group. This badge bridged the gap of dominance and increased the equity of participation among the group members using a real-time feedback on their personal mobile phones. This feedback showed a circle in the middle of a screen connected by four lines to small squares in each corner of the screen representing the individual group members. The colour and position of the circle denoted the interactivity of the group. When the group had a balanced participation then the circle was darker in colour and in the centre of the screen. The thickness of lines connecting the circle represented the speaking time of each group member. Apart from the personal mobile display to give feedback, Balaam et al. [3] used an ambient display showing a coloured circle visualization based on the non-verbal indicator of synchrony during a collaborative task of calendar planning. DiMicco et al. [13] used a shared group display to influence the speaking participation of each group member during a group activity.

In summary, most of these studies were in controlled conditions with small groups consisting of dyads and triads only. Table 2 shows the overview of feedback mechanisms used during co-located collaboration. Some real-time feedback mechanisms acted as a mere reflection for the group to self-regulate instead of an actionable feedback; while others used a post-hoc analysis for the teachers (or facilitators) to reflect on the group activity. The mode of display varied from a public display to smart phone display.

---

<sup>2</sup> An electronic sensing device worn around the neck that can collect and analyze social dynamics.

**Table 2.** Overview of studies on co-located collaboration feedback.

References	Indicators	Feedback
[17]	Total speaking time & body posture	Graphical with coloured shape and lines using personal mobile screens
[37]	Number of ideas	Metaphorical as a groupgarden using public shared display
[18]	Speaking time & eye gaze	Graphical with coloured concentric circles using public table-top private projection
[38]	Speaking time & eye gaze	Graphical with coloured concentric circles using public table-top private projection
[20]	Nodding, speech features & galvanic skin response	Graphical group characteristics using audio, haptic, PDA and public shared display
[2]	Total speaking time	Coloured LED light using public shared table top LED display
[5]	Total speaking time	Coloured concentric rings with lines and dots using public shared table top display
[3]	Pointing	Coloured circle visualization using ambient display
[21]	Log data about different actions performed with the TUI	Pie chart and other statistical charts using private tablet for teachers
[1]	Log data about content knowledge from TUI	Textual and haptic using public TUI display
[13]	Total speaking time	Coloured bar charts using public shared display

In a nutshell, most of the studies in *related work* are in controlled conditions and using specialized furniture, TUI and badges. These settings can be suitable for adhoc CC which can be difficult to adapt in a dynamic setting. They also do not cater to the privacy and fairness of individuals. Most of these studies employ human observers as post-hoc annotators for coding videos to detect traces of collaboration. To tackle these issues, we devise a human-based prototype where privacy, in-the-wild setting and dynamic design is at the centre of our WOz study.

### 3 A WOz Study: Designing the Research Prototype

Based on our analysis, we aimed for creating a flexible research infrastructure that allows us to study feedback in CC making use of different indicators and combining them in different feedback instruments and media. We followed a design-based approach focusing on a specific type of meeting and evaluated different types of indicators, human-observer interfaces, as well as feedback mechanisms. The main components of our research prototype are a defined set of



**Fig. 1.** Meeting room



**Fig. 2.** Annotator room



**Fig. 3.** Public display

indicators and sensors, a user interface for CC observation managed by human observers, as well as a set of feedback components.

### 3.1 Experimental Context

We performed the experiments during three PhD meetings with 3–7 members in each meeting in the room as shown in Fig. 1. Due to the frequent availability of these meetings and ease of not designing the task per se, we chose them. Our main focus was to execute the study *in-the-wild* and preserving the *privacy*. Thus, we used a human annotator who was present in the adjacent room separated by a one-sided transparent wall as shown in Fig. 2. Although it is difficult to see in the picture, the visibility through the wall from the side of the annotator was transparent; while the visibility from the meeting room was opaque. A microphone was used to listen to the conversation in the other room but audio was not recorded. The real-time feedback was shown on a big shared public display in the meeting room (as depicted in Fig. 3) which was managed by the annotator. The real-time feedback visualization could make use of observation data from the human observer and also visualize raw-data, e.g. the audio volume of the group work. The collaborators got a virtual sense of being tracked by a microphone automatically when they saw the changing real-time feedback of their speaking participation on the screen.

### 3.2 Data Logging

For the sake of clarity in data logging, we have segregated the multimodal channel annotation into verbal and non-verbal (i.e., gestures and postures) channels and identified different non-verbal indicators as: looking at laptop or peers; looking down; looking at the feedback; typing with laptop; and making different hand gestures. The verbal indicators are: occurrence, pauses, overlaps, interruptions in speech; affirmatives in speech; and asking questions. But, to ease the logging process for the human annotator, we chose to only focus on the simpler observable audio cues which is the speaking time and turn taking of each group member in a first study. The speech-based cues are ubiquitous in any collaboration and non-verbal cues may be difficult to monitor for one annotator in a large group setting. The annotator was seeing the annotation interface embedded in a



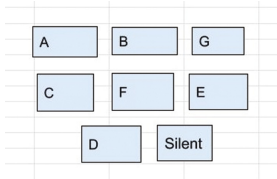


Fig. 4. Annotation interface

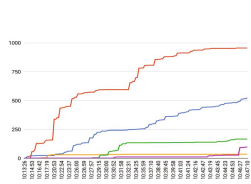


Fig. 5. Mid feedback

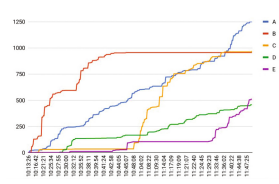


Fig. 6. End feedback

Google sheet as shown in Fig. 4. To preserve the *privacy*, we gave the annotator a coding sheet where each collaborating member was given an alias name from the English alphabet. Moreover, each participating member signed a consent form. Whenever a person starts speaking, the annotator pressed the corresponding button in the interface which automatically creates a cell in the Google sheet with the start time and name of that person. Whenever the annotator presses another person’s button, the end time of the previous person is registered in the sheet. This was possible as the buttons were coupled with a JavaScript to perform the operation. To ensure the reliability of the coding scheme, we had a provision to include multiple annotators but did not use it for our experiments as it involved only simple clicking of a button.

### 3.3 Modeling Participation During Collaboration

The sheet interface was connected to a chart embedded in Google Slides which was updated in real-time when a value is entered by pressing a button. The other columns in Google sheet were automatically populated based on the defined formula which calculates the cumulative speaking time of each member from the beginning of the meeting. Figure 5 shows the group dynamics after the first 30 min during a meeting using a line chart as displayed during the meeting on the big public shared display in the room. The times shown on the horizontal axis is the plot time obtained from the end time of speaking of a member. The value in vertical axis is the total speaking time in seconds from the beginning of the meeting. Figure 6 shows the status of the line chart at the end of the meeting. Here, the speaking time and turn taking represented the participation of each group member. We also collected oral feedback from both the annotators and the collaborators during the iterative design phase.

### 3.4 Results

From our first three iterations in the PhD meetings, we developed a first prototype for analyzing turn-taking and speaking time feedback. Our results showed that we need a higher level annotation interface. Thus, we supported human observers in that they only need to press a button when a new person starts talking. For the visualization on the public shared display, we experimented with different visualizations of the speaking time. Based on participants’ feedback we

altered the display format from an original pie chart to a line chart for displaying the development of the conversation over time. An example of the feedback at different times of a meeting can be seen in Figs. 5 and 6. We can observe that speaker B, who is a second year PhD student, dominates the conversation in the first 30 min; it was his turn to speak regarding his PhD project at that time. But, from that time on-wards he stops to participate in the meeting; indicated by the line parallel to horizontal axis in Fig. 6. We can also observe at the end of the meeting that speaker A, who is the promotor, has spoken the most and changed turns very often to intervene during the meeting; the turn-taking was evident from the frequent change of the shape of the line indicated by small or large spikes.

## 4 Discussion

**RQ1: On the multimodal indicators during CC indicating collaboration quality** — Based on the literature study, we discovered different multimodal indicators during CC in multiple contexts. They can be grouped into *social* (i.e., verbal, non-verbal and physiological) and *epistemological* (i.e., ideas and data logs) indicators. For detecting the social indicators, sensors have been used in past works. But, for detecting the epistemological indicators human help was required as it is difficult for sensors to automatically detect the number of ideas generated from speech by understanding the semantics.

**RQ2: On the feedback during CC** — Feedback during CC is either real-time (for reflection or guiding) or post-hoc (for the purpose of reflection). This brings into the picture two *stakeholders*: the teachers (or facilitators) and the group members. We need this distinction as it will help in designing the feedback. Some works used TUI and other electronic mediums like Interactive White Boards (IWB) and tablets during collaboration which requires a lot of preparation before a collaborative task. Therefore, it is difficult to use it in real-world dynamic settings. Besides, there is a trade-off between personalization for the group and privacy. More personalized feedback meant for the whole group is less privacy preserving. Thus, there should be a decision on the level (i.e., group, individual or both) of feedback to be shown depending on the circumstances at hand.

**On the research prototype to give real-time feedback** — We take a step in building an initial prototype design with the aim to facilitate real-time collaboration during meetings. We were successful in building a click-based interface for the annotator which also reduces memory overhead. This helps us to create a hybrid setup without building an actual automated sensor-based system to experiment with different types of real-time feedback mechanisms during CC. We can later use these insights to build the sensor-based or hybrid setup. Here, we can build individual components in a modular fashion to track other indicators of collaboration quality; and integrate them to a single dashboard.

## 5 Conclusions and Future Work

Collaboration being an important skill and ubiquitously present in our day to day activities, we try to look into the different collaboration indicators in various contexts in the literature. We find different types of indicators like gaze, speaking time, posture, gesture, number of ideas generated, etc. Then we look into the impact of feedback during collaboration and find that visual real-time feedback has some impact on the collaboration like improving the equity of audio participation. This feedback can range from private displays (like PDA, mobile phones) to a more public one (like TUI, shared display).

Based on this overview, we took a step further and built a real-time feedback prototype during collaboration based on a privacy-preserving WOz study in-the-wild. Here, we study collaboration during co-located PhD meetings using human observers acting as a proxy for sensors. We find that the human observers could easily track ‘who spoke when and for how much time’ by pressing a button.

As future work suggestions, we need to define the *goal* and *outcome* of the collaboration task and make it clear in the evaluation criteria as to whether we measure collaboration as a process, outcome or both. Then, we can focus on the feedback mechanisms for facilitating collaboration. We can also borrow some insights from the mapping of multimodal data to feedback in an individual learning context [11]. The feedback can be: human based, sensor based or a hybrid of both. We need to decide the type (number of pointing gestures, speaking time, number of interruptions, number of eye contact with peers, etc.), modelling (i.e., individual, group or both) and display of feedback (i.e., personal, public or both) based on *action-based research* [14] where we need to take the preliminary feedback of different stakeholders like teachers (or facilitators) and the group members. Our long term goal is to do action-based research and build a sensor-based automated (or hybrid) feedback system during CC using the currently built research prototype. Here, we can include different feedback components to identify multiple indicators of collaboration and proceed towards an automated system using deep neural networks to integrate data from multiple sensors [31].

## References

1. Anastasiou, D., Ras, E.: A questionnaire-based case study on feedback by a tangible interface. In: Proceedings of the 2017 ACM WS on Intelligent Interfaces for Ubiquitous and Smart Learning, pp. 39–42. ACM (2017)
2. Bachour, K., Kaplan, F., Dillenbourg, P.: An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Trans. Learn. Technol.* **3**(3), 203–213 (2010)
3. Balaam, M., Fitzpatrick, G., Good, J., Harris, E.: Enhancing interactional synchrony with an ambient display. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 867–876. ACM (2011)

4. Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., D'Angelo, C., Alozie, N.: Privacy-preserving speech analytics for automatic assessment of student collaboration. In: INTERSPEECH, pp. 888–892 (2016)
5. Bergstrom, T., Karahalios, K.: Conversation clock: visualizing audio patterns in co-located groups. In: 40th Annual Hawaii International Conference on System Sciences, p. 78. IEEE (2007)
6. Chikersal, P., Tomprou, M., Kim, Y.J., Woolley, A.W., Dabbish, L.: Deep structures of collaboration: physiological correlates of collective intelligence and group satisfaction. In: CSCW, pp. 873–888 (2017)
7. Cukurova, M., Luckin, R., Mavrikis, M., Millán, E.: Machine and human observable differences in groups' collaborative problem-solving behaviours. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 17–29. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_2](https://doi.org/10.1007/978-3-319-66610-5_2)
8. Cukurova, M., Luckin, R., Millán, E., Mavrikis, M.: The nispi framework: analysing collaborative problem-solving from students' physical interactions. *Comput. Educ.* **116**, 93–109 (2018)
9. Davidsen, J., Ryberg, T.: This is the size of one meter: childrens bodily-material collaboration. *Int. J. CSCL* **12**(1), 65–90 (2017)
10. Dede, C.: Comparing frameworks for 21st century skills. In: 21st Century Skills: Rethinking How Students Learn, vol. 20, pp. 51–76 (2010)
11. Di Mitri, D., Schneider, J., Drachler, H., Specht, M.: From signals to knowledge. A conceptual model for multimodal learning analytics. *J. Comput. Assist. Learn.* **34**(4), 338–349 (2018)
12. Dillenbourg, P.: What do you mean by collaborative learning? (1999)
13. DiMicco, J.M., Pandolfo, A., Bender, W.: Influencing group participation with a shared display. In: Proceedings of the 2004 ACM Conference on CSCW, pp. 614–623. ACM (2004)
14. Dyckhoff, A.L.: Action research and learning analytics in higher education. *eled* **10**(1) (2014)
15. Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., Divakaran, A.: Multimodal analytics to study collaborative problem solving in pair programming. In: Proceedings of the 6th International Conference on LAK, pp. 516–517. ACM (2016)
16. Jermann, P., Mullins, D., Nüssli, M.A., Dillenbourg, P.: Collaborative gaze footprints: correlates of interaction quality. In: CSCL 2011 Conference Proceedings, vol. 1, pp. 184–191. International Society of the Learning Sciences (2011)
17. Kim, T., Chang, A., Holland, L., Pentland, A.S.: Meeting mediator: enhancing group collaboration using sociometric feedback. In: Proceedings of the 2008 ACM Conference on CSCW, pp. 457–466. ACM (2008)
18. Kulyk, O., Wang, J., Terken, J.: Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 150–161. Springer, Heidelberg (2006). [https://doi.org/10.1007/11677482\\_13](https://doi.org/10.1007/11677482_13)
19. Lubold, N., Pon-Barry, H.: Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In: Proceedings of the 2014 ACM WS on Multimodal Learning Analytics Workshop and Grand Challenge, pp. 5–12. ACM (2014)
20. Madan, A., Caneel, R., Pentland, A.S.: Groupmedia: distributed multi-modal interfaces. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 309–316. ACM (2004)

21. Martinez-Maldonado, R., Clayphan, A., Yacef, K., Kay, J.: Mfeedback: providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Trans. Learn. Technol.* **8**(2), 187–200 (2015)
22. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. *Int. J. CSCL* **2**(1), 63–86 (2007)
23. O'Donnell, A.M.: *The role of peers and group learning* (2006)
24. Pijera-Daz, H.J., Drachsler, H., Kirschner, P.A., Järvelä, S.: Profiling sympathetic arousal in a physics course: how active are students? *J. Comput. Assist. Learn.* **34**(4), 397–408 (2018)
25. Richardson, D.C., Dale, R.: Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* **29**(6), 1045–1060 (2005)
26. Scherr, R.E., Hammer, D.: Student behavior and epistemological framing: examples from collaborative active-learning activities in physics. *Cogn. Instr.* **27**(2), 147–174 (2009)
27. Schneider, B., Blikstein, P.: Unraveling students interaction around a tangible interface using multimodal learning analytics. *J. Educ. Data Min.* **7**(3), 89–116 (2015)
28. Schneider, B., Pea, R.: Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *Int. J. CSCL* **8**(4), 375–397 (2013)
29. Schneider, B., Pea, R.: Toward collaboration sensing. *Int. J. CSCL* **9**(4), 371–395 (2014)
30. Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., Pea, R.D.: 3d tangibles facilitate joint visual attention in dyads. *International Society of the Learning Sciences, Inc. [ISLS]* (2015)
31. Schneider, J., Di Mitri, D., Limbu, B., Drachsler, H.: Multimodal learning hub: a tool for capturing customizable multimodal learning experiences. In: Drachsler, H., et al. (eds.) *EC-TEL 2018. LNCS*, vol. 11082, pp. 45–58. Springer, AG (2018)
32. Shih, P.C., Nguyen, D.H., Hirano, S.H., Redmiles, D.F., Hayes, G.R.: Groupmind: supporting idea generation through a collaborative mind-mapping tool. In: *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pp. 139–148. ACM (2009)
33. 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)
34. Spikol, D., Ruffaldi, E., Landolfi, L., Cukurova, M.: Estimation of success in collaborative learning based on multimodal learning analytics features. In: *Proceedings of the 17th ICALT*, pp. 269–273. IEEE (2017)
35. Stahl, G., Law, N., Hesse, F.: Reigniting CSCL flash themes. *Int. J. CSCL* **8**(4), 369–374 (2013)
36. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: *CHI 2002 Extended Abstracts on Human Factors in Computing Systems*, pp. 858–859. ACM (2002)
37. Tausch, S., Hausen, D., Kosan, I., Raltchev, A., Hussmann, H.: Groupgarden: supporting brainstorming through a metaphorical group mirror on table or wall. In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction*, pp. 541–550. ACM (2014)
38. Terken, J., Sturm, J.: Multimodal support for social dynamics in co-located meetings. *Pers. Ubiquitous Comput.* **14**(8), 703–714 (2010)

39. Triglianos, V., Prahara, S., Pautasso, C., Bozzon, A., Hauff, C.: Measuring student behaviour dynamics in a large interactive classroom setting. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 212–220. ACM (2017)
40. Worsley, M., Blikstein, P.: Leveraging multimodal learning analytics to differentiate student learning strategies. In: Proceedings of the 5th International Conference on LAK, pp. 360–367. ACM (2015)



# Towards Personalized Learning Objectives in MOOCs

Tobias Rohloff<sup>(✉)</sup> and Christoph Meinel

Hasso Plattner Institute, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany  
{tobias.rohloff,christoph.meinel}@hpi.de

**Abstract.** Instead of measuring success in Massive Open Online Courses (MOOCs) based on certification and completion-rates, researchers started to define success with alternative metrics recently, for example by evaluating the intention-behavior gap and goal achievement. Especially self-regulated and goal-oriented learning have been identified as critical skills to be successful in online learning environments with low guidance like MOOCs, but technical support is rare. Therefore, this paper examines the current technical capabilities and limitations of goal-oriented learning in MOOCs. An observational study to explore how well learners in five MOOCs achieved their initial learning objectives was conducted, and the results are compared with similar studies. Afterwards, a concept with a focus on technical feasibility and automation outlines how personalized learning objectives can be supported and implemented on a MOOC platform.

**Keywords:** Learning objectives · MOOCs  
Goal-oriented learning · Self-regulated learning  
Learning analytics · E-learning

## 1 Introduction

Massive Open Online Courses (MOOCs) offer the opportunity of free education for everyone who has access to the Internet. Since the first evaluations of such online courses, a main criticism is the low completion rate ranging from 5 to 10%, which has been discussed frequently [3, 11]. This certification-centered focus is reasonable from the perspective of a MOOC platform provider or teaching team since these stakeholders are interested in the success of their courses. Nevertheless, it turned out that a lot of learners dropped out of courses for different reasons, mostly due to poor time management or course difficulty [13]. The initial assumption that MOOCs will largely attract less-educated people and students had to be adjusted. Lifelong learners, especially well-educated professionals, form a large part of the learning community and not necessarily all of them are interested in gaining a certificate [4]. Therefore, the meaning of success in MOOCs was discussed again since a dropout can also mean that a learner got all the knowledge it needed at this time [17]. Thus, alternative measurements

were proposed. For example, Renz, Schwerer, and Meinel [20] introduced the concept of a learning material consumption rate, next to the completion rate, to determine success. From the learner’s perspective, the meaning of success is connected to their motivation and goals, and lifelong learners have varying learning objectives. Therefore, researchers started to define success based on the intention-behavior gap [7] to measure achievement based on students’ individual reported goals. Unfortunately, courses with self-reported learning goals based on learners’ intention are rarely implemented and conducted. In terms of personalization the preparation of alternative learning paths, either by varying topics or proficiency levels, requires additional resources. This results mostly in increased production time and cost. Modularization can confuse students more than it supports them [12]. Instead, goal-oriented learning – as part of a broader self-regulated learning strategy – has been identified as a valuable skillset in online learning environments [13, 28]. Nevertheless, technical support for personalized learning objectives in MOOCs is limited.

Thus, this paper provides two contributions to the field of technology enhanced learning. To examine the current technical capabilities and limitations of goal-oriented learning, an observational study is presented to explore how well learners in MOOCs achieved their initially specified learning objectives, based on five courses ( $N = 25,801$ ). The results are compared with similar studies, to examine their general validity and emphasize the importance of such work. Secondly, a concept is outlined how personalized learning objectives can be supported and implemented on a MOOC platform. Thereby, the focus is set on technical feasibility and a high level of automation, which is a critical issue for the success of goal setting and self-evaluation in such a high-scalable online learning environment.

## 2 Pedagogical Rationale

Mayes and De Freitas [18] described learning outcomes of e-learning environments in higher and further education. They extended Goodyear’s [6] three kinds of learning in higher education – which are *academic*, *generic competence* and *individual reflexivity* – by *skill*-based outcomes to fully encompass further education. They presented design principles of learning environments, whereas many researchers recommend to apply constructivism in distance education [9]. They summarized the following principles:

- The learner actively constructs knowledge, through achieving understanding
- Learning depends on what we already know, or what we can already do
- Learning is self-regulated
- Learning is goal-oriented
- Learning is cumulative

The authors outlined two main aspects for activities to construct understanding: interactions with material systems and concepts in the domain, and interactions where learners discuss their developing understanding and competence. In the



research literature they recognized an increasing focus on the design of learner-centered methods and environments, whereby the ultimate goal of educational technology is the achievement of individualized instruction. Nevertheless, personalization at scale comes with many instructional and technical hurdles. Thereby, goal setting is a first step to understand learners' intention and motivation.

Also, self-regulated and goal-oriented learning have been identified as important topics in educational psychology due to their influence on learners' achievement [5, 15]. Especially in large-scale online learning environments with little support and guidance like MOOCs, self-direction is a critical skill for learners' goal achievement [13, 28], but many learners have difficulties in applying self-regulation [16]. A lot of models and frameworks for self-regulated learning have been proposed. This work focuses on the following metacognitive strategies, which were especially developed to support goal-oriented learning [5, 15, 29]:

**Goal setting** to agree on the effort required to achieve objectives on different learning content granularity.

**Strategic planning** to determine the sequence, schedule and completion of activities to accomplish learning goals.

**Self-evaluation** to monitor the learning progress and outcome in relation to the defined learning goals.

### 3 The Status Quo of Learning Objectives in MOOCs

For the support of self-regulated learning in MOOCs certain approaches have been researched, for example a time planner to schedule the next learning session [22], recommendations of learning strategies [14] or personalized feedback with dashboards [2]. Yet, no approach is applied largely. Additionally, few related work is available which examines goal-oriented learning in MOOCs. This work aims to fill this gap by better supporting the strategies of goal-oriented and self-regulated learning in MOOCs. Therefore, this section investigates the current capabilities and limitations of goal setting and self-evaluation on a state-of-the-art MOOC platform before comparing the results with similar studies.

#### 3.1 Evaluated Courses

To investigate the targeted and accomplished learning objectives of MOOC participants, five courses have been examined in this study (Table 1). These courses were conducted on openHPI<sup>1</sup>, the MOOC platform of Hasso Plattner Institute. The taught topics are all based on the field of information technology and computer science and the required proficiency levels range from beginner to academic and professionals. In total, 25,801 learners had been enrolled at *course middle*. The *middle* is a course-specific date, which marks the last reasonable point to enroll for a course with the possibility to still gain a *Record of Achievement*. A *Record of Achievement* is issued to those who have earned more than 50% of the

<sup>1</sup> <https://open.hpi.de/>.

maximum number of points for the sum of all graded assignments. A *Confirmation of Participation* is issued to those who have completed at least 50% of the course material.

The first course, *Object-Oriented Programming in Java* (javaEinstieg2017), was a four weeks course for beginners running from March 27, 2017 through May 14, 2017. Every week introduced different Java language features and object-oriented programming concepts with video lectures, followed by self tests and online programming exercises. Most of the programming exercises were graded for the final certificate. Additionally, an optional team peer assessment was conducted, where learners had the chance to gain bonus points. A total number of 9,242 enrollments were taken at course middle. The next course was a two weeks workshop with the topic *Introduction into a Java IDE* (javawork2017). This course was held from May 01, 2017 through May 15, 2017 and built upon the taught concepts of the javaEinstieg2017 course. Thus, a basic knowledge about the Java programming language was recommended. The first two weeks showed practical knowledge with lecture videos, followed by ungraded self tests. At the end a graded peer assessment was conducted, which was the requirement to gain a certificate. 4,112 learners enrolled at course middle. The third course was a two week course as well, and addressed the question *How does a search engine work?* (searchengine2017) from May 29, 2017 through June 20, 2017. The course was designed to be an introduction of the topic for persons outside the discipline, but also as a starting point for professionals and academic people who want to get a first overview. The course structure followed the typical MOOC approach with consecutive videos and self tests. At the end a graded exam was performed and 4,145 participants had been enrolled at course middle. The fourth course about *Mainframes* (mainframes2017) was held from June 05, 2017 through July 27, 2017. This six weeks course provided an in-depth perspective on mainframe architectures, application development, databases, security and storage management. Thus, this courses mainly targeted academic and professional people. Next to the video lectures and self tests, a weekly graded assignment was conducted, as well as a graded exam at the end of the course. At course middle 3,026 learners had been enrolled. The *In-Memory Data Management* (imdb2017) course

**Table 1.** Evaluated courses

Course	Enrollments		No-Shows		Weeks	Language
	Middle	End	Middle	End		
javaEinstieg2017	9242	10402	2632	2387	4	German
javawork2017	4112	4336	2631	2241	2	German
searchengine2017	4145	4484	2443	1824	2	German
mainframes2017	3026	3396	1356	1281	6	German
imdb2017	5276	5825	2874	2697	6	English
Total	25801	28443	11936	10430	-	-

dealt with the management of enterprise data in column-oriented in-memory databases and their inner mechanics. The course was running for six weeks from September 18, 2017 through November 18, 2017 and 5,276 learners enrolled in it. Due to the specific technical focus, the target groups were academics and professionals. This course was graded by a weekly assignment and a final exam.

In summary, the evaluated courses provide a well-balanced data basis with different course lengths, target groups and proficiency levels, as well as different theoretical and practical examination modalities. All of them offered the two introduced certificate types: a *Record of Achievement* and a *Confirmation of Participation*. Table 1 also displays the number of enrollments and *no-shows*. Based on Hill's [8] definition of *no-shows* (learners who enrolled for a course but never viewed any content), an overall *show rate* of 53.78% at course middle was reached. Additionally, following the definitions of Renz, Schwerer, and Meinel [20] a total *completion rate* of 29.02% and *consumption rate* of 52.30% were measured. When comparing the *show rate* and *consumption rate*, it can be seen that almost all active learners that enrolled before course middle visited more than 50% of all learning content and therefore gained a *Confirmation of Participation*.

### 3.2 Methodology

When accessing one of the courses for the first time, a welcome text is presented to the learner with general information about the course. The following item is an optional pre-course survey, which asks the learner about its primary goal for the enrollment into this course amongst other general questions. Based on the platform's feature set and available certificates, four mutually exclusive objectives are provided:

- Objective 1 – I would like to receive a record of achievement in the end and learn the course content.
- Objective 2 – I am mainly interested in learning the course content. The record of achievement is not important to me.
- Objective 3 – I am only interested in selected learning units.
- Objective 4 – I just want to look around.

An overview of all criteria to achieve and to exceed the learning objectives is shown in Table 2. The achievement of objective 1 and 2 can be traced by course completion if a certain certificate was gained. To accomplish objective 1, a *Record of Achievement* needs to be reached. For objective 2 the assumption was made, that if a learner consumed the majority of learning content (50%), a *Confirmation of Participation* was achieved.

For the accomplishment of objective 3 and 4 a behavioral analysis based on user interaction events was conducted. To achieve objective 3, the user needs to watch at least 1 video lecture. This is the base unit to measure if the user consumed and interacted with any learning content since there is no platform feature available that enables the user to select the specific learning content she

**Table 2.** Criteria for learning objective achievement

Objective	Criteria to achieve objective	Criteria to exceed objective
Objective 1	Accomplish record of achievement	n/a
Objective 2	Accomplish confirmation of part	Accomplish objective 1
Objective 3	Watch at least 1 video	Accomplish objective 1 or 2
Objective 4	Visit at least 3 items	Accomplish objective 1 or 2 or 3

is interested in. For objective 4, the visit of at least 3 items is defined as the criteria to achieve the learning goal. This specific number was chosen because the first visited item is the welcome text when entering the course, the second is the survey itself, and the third item visit is the proof that at least one learning item was visited. These assumptions already show limitations of the platform regarding goal setting and evaluation.

By following this approach, no post-course survey was necessary to determine goal achievement of all students that responded to the pre-course survey. All measurements are based on platform data, which should reduce the influence of the survivorship bias. Therefore, it was not required that learners finished the course or sending a post-course survey via email to all participants.

### 3.3 Pre-course Survey

The results of the pre-course survey for every course can be seen in Table 3. A total amount of 9,698 users provided their learning objective. In relation to the total number of *shows at course middle*<sup>2</sup> (13,865) a response rate of 69.95% was reached. Between 22.52% and 36.03% stated, that they want to receive a *Record of Achievement* (objective 1), with a total result of 26.63%. The majority of users (61.54%) are mainly interested in learning the course content, without the need to gain a *Record of Achievement*, and therefore chose objective 2, ranging from 54.41% to 65.80%. Between 3.62% and 5.41% selected objective 3, since they are only interested in selected learning units, with a total result of 4.45%. At last, 7.37% stated that they only want to look around (objective 4), with a range from 5.94% to 10.74%.

### 3.4 Goal Achievement Analysis

When assessing the results of the pre-course survey, it is notable that only about one quarter of the users are interesting in a graded performance appraisal and considerably more than half of the users are mainly interested in the content itself without the need of a *Record of Achievement*. This reflects the varying learning objectives of lifelong learners since especially well-educated professionals

<sup>2</sup> Based on the *total enrollments at course middle* minus the *total number of no-shows at course middle* from Table 1.

form a large part of the learning community and not all of them are necessarily interested in gaining a certificate [4].

**Table 3.** Pre-course survey: what is your primary goal for the enrollment into this course?

Course	Objective 1	Objective 2	Objective 3	Objective 4
javaEinstieg2017	1006 (22.52%)	2940 (65.80%)	191 (04.27%)	331 (07.41%)
javawork2017	342 (23.73%)	927 (64.33%)	78 (05.41%)	94 (06.52%)
searchengine2017	528 (32.18%)	924 (56.31%)	78 (04.75%)	111 (06.76%)
mainframes2017	319 (29.79%)	591 (55.18%)	46 (04.30%)	115 (10.74%)
imdb2017	388 (36.03%)	586 (54.41%)	39 (03.62%)	64 (05.94%)
Total	2583 (26.63%)	5968 (61.54%)	432 (04.45%)	715 (07.37%)

**Table 4.** Achieved learning objectives of all courses

Objective	Satisfied	Exceeded	Satisfied or exce.	Missed
Objective 1	1099 (42.55%)	n/a	1099 (42.55%)	1484 (57.45%)
Objective 2	1176 (19.71%)	1558 (26.11%)	2734 (45.81%)	3234 (54.19%)
Objective 3	223 (51.62%)	165 (38.19%)	388 (89.81%)	44 (10.19%)
Objective 4	77 (10.77%)	636 (88.95%)	713 (99.72%)	2 (00.28%)
Total	2386 (25.09%)	2359 (24.81%)	4745 (49.90%)	4764 (50.10%)

Few users stated that they are only interested in selected learning units or only want to look around. This may be related to the fact that at course start only the first week was available, and the remaining content followed week by week. This is a typical approach in MOOCs to foster discussions in the forum and support the mastery learning approach. Nevertheless, this reveals the shortcoming that at course beginn it is hard to get an overview of all content and topics that will be taught in the following weeks.

In Table 4 the overall goal achievement is displayed. At first, it can be seen that nearly half of the users achieved or exceeded their goals and the other half missed their objective. Also the total satisfied and exceeded achievements are almost equally distributed. From this insight it can be derived that there is a large user group that either changes their goal during course runtime or drop out due to course difficulty, poor time management, illness or other issues. In both cases it shows the limitation that learning objectives cannot be set in a proper way which allows the user to also adjust them at a later point of time. Nevertheless, the results show a big range when comparing the different learning objectives with each other since the objectives with the highest achievement rate required much less course activity and vice versa.

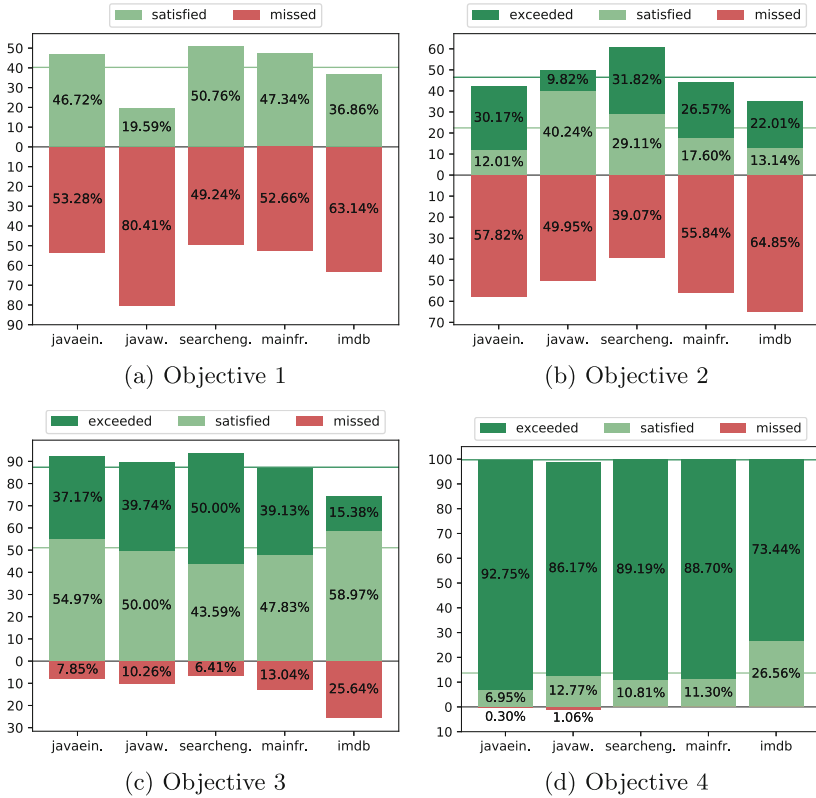


Fig. 1. Achieved learning objectives per course

Figure 1 displays the individual achievement rates for all courses, grouped by the defined objectives. These results are centered around a zero line in order to allow an easy comparison of the achieved learning objectives. Satisfying and exceeding a goal are stacked upwards, whereas missing a learning goal is stacked downwards. Additionally, the first horizontal line in the upper space marks the average mean of satisfying a goal, and the second line the average mean of satisfying and exceeding a goal combined. The specific mean values can be seen in Table 4. Compared with a standard deviation of 0.1155 for satisfying objective 1, it is notable that only the *javawork2017* course showed a greater deviation. This can be attributed to the fact that this course was only graded by a peer assessment, which required much more effort than a typical multiple choice examination. The highest achievement rate was reached by the *searchengine2017* course. This course was only graded by a single final exam without any weekly assignments, which reduced the required effort. The other three courses were graded by weekly assignments and a final exam. The achievement rates of objective 2 show a much higher variation, and objective 3 and 4 show overall high achievement rates, since these goals require less engagement. All in all, the

individual achievement rates across the different courses point to the fact that goal achievement strongly depends on the course design, examination and difficulty of different goals.

### 3.5 Related Research

Obviously, a sample size of five courses does not allow to draw general statements about goal achievement rates in MOOCs. Therefore, related and similar studies are presented in this section. A case study by Wilkowski, Deutsch, and Russell [25] about one course showed that 52.5% of their participants ( $N = 20,977$ ) intended to complete their evaluated course with a (free of charge) certificate, from which 27% met or exceeded this goal at the end. The other learners preferred to learn new skills or explore the course content. Combined with these students who targeted smaller learning goals, a total number of 42.4% met or exceeded their goals at the end. The authors recommended to offer more personalized course designs based on students' goals, to move beyond the one-size-fits-all approach in MOOCs.

Another study with 37,880 enrollments across six courses by Staubitz and Meinel [24] showed that only a few learners (0.64–1.24%) are interested in gaining a (charged) verified certificate to earn credits for their degree, on-the-job training or job applications. From the participants who booked this certificate option, between 63.3% and 92.0% gained a certificate at the end, whereby the paid fee increased the motivation. Henderikx, Kreijns, and Kalz [7] examined the success of two MOOCs based on the intention-behavior gap. In the first course 59% of their participants achieved or achieved more than initially intended ( $N_1 = 65$ ). An even higher success rate of 70% was found in the second course ( $N_2 = 101$ ). These results are based on a subset of learners who responded to the post-survey which leads to survival bias. Nevertheless, they “underline the importance of individual perspectives” and recommend to consider that “individual goal achievement does not necessarily matches goal achievement from the institutional perspective.” Other studies, which measured certificate achievement based on students' self-reported intention to complete a course, found completion rates between 22 and 29% [19,26] or around 9% [15].

### 3.6 Discussion

To summarize regardless of the variation in the reported goal achievement rates, a substantial percentage of students both meet or exceed, or miss their goals in MOOCs. The specific ratio is course-specific and probably depends on the course design and difficulty. Nevertheless, this and related studies show the importance to better support the presented strategies for self-regulated and goal-oriented learning in MOOC environments. Thereby, different shortcomings have been identified.

Currently, goal setting is mostly done with a pre-course survey. This maybe helps the teaching team to get a broad insight into the overall motivation of their learning community. However, the learners have mostly neither a possibility to

self-evaluate their learning process and outcome regarding their stated learning objective, nor be able to adjust their objective during the course runtime. Learner dashboards mostly focus on overall course completion [10], which does not reflect the objective of a large amount of learners, as the analysis has shown.

Also, the measurement of goal achievement is mostly done manually since the survey responses cannot be processed automatically. Sub-goals like the completion of a certain topic section or week are only provided if the teaching team prepares such survey answers. Generic answers like “I am only interested in selected learning units” as in this study include a certain bias since the learner is not aware of which selected learning units are available at all. Furthermore, some studies about strategic planning were briefly presented [14, 22], but these were not a focus topic of this paper’s analysis. However, strategic planning must be considered in a concept to better support personalized learning objectives, next to goal setting and self-evaluation.

## 4 A Concept to Support Personalized Learning Objectives in MOOC Environments

This section outlines a concept to support goal setting, strategic planning and self-evaluation, to implement goal-oriented learning as personalized learning objectives in MOOCs. It builds on top of the previously identified capabilities and shortcomings of MOOC platforms in general but with a technical focus on feasibility and automation in the context of the openHPI platform. Nevertheless, the introduced features should be realizable on any other MOOC platform as well.

### 4.1 Goal Setting

Currently, goal setting is mostly done with pre-course surveys in many MOOC platforms. This should be implemented as a course-independent platform feature, which offers the available learning objectives in a clear way. It needs to be studied if this should be a mandatory step, e.g. as part of the course enrollment process, or as an optional advice, which can be shown to the user while browsing through the course. Therefore, a multivariate experiment can be used to examine if this is accepted and used by all learners or only by a sub-group. Also, it should be possible to change the targeted objective at any given time. By implementing such a feature, goal setting does not need to be maintained by the teaching team as a survey anymore. Also, it is finally possible to evaluate the learning objectives inside the platform itself to further monitor the learning progress based on them.

In order to offer course-specific learning objectives, the learning content needs to be categorized and labeled first. Typically, knowledge transfer in MOOCs is based on video lectures and assessed with quizzes. Video segmentation is a well researched field, e.g. by visual transition detection [27], and can be further improved with outline extraction through analyzing the presentation slides [1]. Related quiz questions could be identified with natural language processing



techniques. Also the course structure itself supports the categorization, since it already offers an order and titles for each learning item and section.

The biggest challenge could be a practical one: the availability of content. Quite often course content is provided and uploaded during the course runtime when users already started to learn. This is problematic with regard to the selection of learning objectives. It could be solved by either offering new goals as soon as they are available or by supporting the teaching team to implement a structured course outline before course start without the content. A course builder tool could enable to plan the weeks of a course ahead and help to enrich them with goal metadata. The requirements for such a tool should be developed in cooperation of real world teaching teams. Interviews are necessary to understand their production processes, dependencies and deadlines. However, these processes vary strongly between organizations and machine-based automations always come with a certain error-rate. Therefore it must be ensured that labeled content can be corrected and improved by human, either teaching teams or learners.

## 4.2 Strategic Planning

Strategic planning methods were identified as positive predictors of goal achievement [15]. Especially regarding learning objectives technical support to plan time management and effort regulation come in handy. Features like custom reminders, priorities and due dates for certain learning items or goals are straightforward to implement and well testable with control groups. Some first work was already done in this field [22] but needs to be carried out in-depth. To further increase learning efficiency, mobile learning can be used to integrate learning activities into daily routines, sending push notification as reminders or to parallelize learning tasks with second screen companion applications [21].

## 4.3 Self-evaluation

Learner dashboards are a common practice to monitor learning progress and goal achievement. The design and evaluation of such visualization tools can be done on different levels like metacognitive, cognitive, behavioral, emotional, self-regulative or tool usability. However, a strong mismatch between a dashboard's goal and its evaluation was identified in a literature review of 26 papers, for which reason Jivet et al. [10] proposed certain design recommendations. They emphasize dashboards as pedagogical tools designed on educational concepts, whereas the comparison with peers should be used with caution. Also, only a subgroup of learners will benefit at large from such tools and it should be integrated into the regular learning activities. To examine the overall tool, also Scheffel et al. [23] proposed an evaluation framework for learners and teachers. Nevertheless, goal monitoring and achievement was not considered in these studies.

A central course dashboard also provides the opportunity to become a personal assistant which helps to navigate through the course content. Next to such a central element, smaller widgets attached to the learning content could

provide instant feedback about it and the individual performance. Additionally, when achieving a smaller learning objective a greater one could be promoted to further increase motivation and engagement. The technical foundation for such tools are advanced learning analytics capabilities, as presented in [20].

## 5 Conclusion

This paper introduced the potential of personalized learning objectives in Massive Open Online Course to shift the focus from completion-centered success rates based on gained certificates to individual course goals which better accomplish the needs of lifelong learners. Therefore, the current status quo of learning objectives in MOOCs was examined with an observational study of five courses how well learners in MOOCs achieved their initially intended learning objectives. The results and the comparison with similar studies show that goal achievement rates are course-specific and likely depend on course design, examination modalities and difficulty. In total, almost 70% of all active learners at course middle provided a course objective ( $N = 13,865$ ). 49,90% of learners achieved or exceeded their goals, but also the effort required for a specific goal heavily affected the achievement rates. Nevertheless, technical support for personalized learning objectives is rare. Most studies rely on self-reported data from user surveys, which does not allow to provide feedback based on the selected goals and also the teaching team cannot draw any further conclusions about progress and success afterwards.

From a pedagogical perspective, self-regulated and goal-oriented learning were identified as critical skills for learner achievement, especially in online learning environments with low guidance and support like MOOCs. Therefore, the strategies goal setting, strategic planning and self-evaluation were outlined with possible implementations in a concept to support personalized learning objectives in MOOCs. Thereby, the focus was set on technical feasibility and automation to provide such functionality on a platform level instead of individual course designs by different teaching teams. This should pave the way for further research in this field and support the transition from a one-size-fits-all approach in online learning at scale to a more individual learning experience tailored for the needs of lifelong learners.

## References


1. Che, X., Yang, H., Meinel, C.: Adaptive E-Lecture video outline extraction based on slides analysis. In: Li, F.W.B., Klamma, R., Laanpere, M., Zhang, J., Manjón, B.F., Lau, R.W.H. (eds.) ICWL 2015. LNCS, vol. 9412, pp. 59–68. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-25515-6\\_6](https://doi.org/10.1007/978-3-319-25515-6_6)
2. Davis, D., Chen, G., Jivet, I., Hauff, C., Houben, G.: Encouraging metacognition & self-regulation in MOOCs through increased learner feedback. In: Proceedings of the LAK 2016 Workshop on Learning Analytics for Learners, pp. 17–22 (2016). <http://ceur-ws.org/Vol-1596/paper3.pdf>

3. Davis, D., Jivet, I., Kizilcec, R.F., Chen, G., Hauff, C., Houben, G.-J.: Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, LAK 2017, pp. 454–463. ACM (2017). <https://doi.org/10.1145/3027385.3027411>
4. Emanuel, E.J.: Online education: MOOCs taken by educated few. *Nature* **503**(7476), 342 (2013). <https://www.nature.com/articles/503342a>
5. Farsani, M.A., Beikmohammadi, M., Mohebbi, A.: Self-regulated learning, goal-oriented learning, and academic writing performance of undergraduate Iranian EFL learners. In: TESL-EJ, vol. 18(2) (2014). <https://eric.ed.gov/?id=EJ1045129>
6. Goodyear, P.: Psychological foundations for networked learning. In: Steeples, C., Jones, C. (eds.) *Networked Learning: Perspectives and Issues*. Computer Supported Cooperative Work, pp. 49–75. Springer, London (2002). [https://doi.org/10.1007/978-1-4471-0181-9\\_4](https://doi.org/10.1007/978-1-4471-0181-9_4)
7. Henderikx, M.A., Kreijns, K., Kalz, M.: Refining success and dropout in massive open online courses based on the intention-behavior gap. *Distance Educ.* **38**(3), 353–368 (2017). <https://doi.org/10.1080/01587919.2017.1369006>
8. Hill, P.: Emerging student patterns in MOOCs: a (revised) graphical view (2013). <https://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view/>
9. Huang, H.-M.: Toward constructivism for adult learners in online learning environments. *Br. J. Educ. Technol.* **33**(1), 27–37 (2002). <https://doi.org/10.1111/1467-8535.00236>
10. Jivet, I., Scheffel, M., Specht, M., Drachsler, H.: License to evaluate: preparing learning analytics dashboards for educational practice. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, LAK 2018, pp. 31–40. ACM (2018). <https://doi.org/10.1145/3170358.3170421>
11. Jordan, K.: Initial trends in enrolment and completion of massive open online courses. *Int. Rev. Res. Open Distrib. Learn.* **15**(1) (2014). <https://doi.org/10.19173/irrodl.v15i1.1651>
12. Joyner, D.A.: Congruency, adaptivity, modularity, and personalization: four experiments in teaching introduction to computing. In: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, L@S 2017, pp. 307–310. ACM (2017). <https://doi.org/10.1145/3051457.3054011>
13. Kizilcec, R.F., Halawa, S.: Attrition and achievement gaps in online learning. In: Proceedings of the Second (2015) ACM Conference on Learning @ Scale, L@S 2015, pp. 57–66. ACM (2015). <https://doi.org/10.1145/2724660.2724680>
14. Kizilcec, R.F., Pérez-Sanagustín, M., Maldonado, J.J.: Recommending self-regulated learning strategies does not improve performance in a MOOC. In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale, L@S 2016, pp. 101–104. ACM (2016). <https://doi.org/10.1145/2876034.2893378>
15. Kizilcec, R.F., Pérez-Sanagustín, M., Maldonado, J.J.: Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Comput. Educ.* **104**, 18–33 (2017). <https://doi.org/10.1016/j.compedu.2016.10.001>
16. Lajoie, S.P., Azevedo, R.: Teaching and learning in technology-rich environments. In: *Handbook of Educational Psychology*. Routledge (2006). <https://doi.org/10.4324/9780203874790.ch35>
17. Liyanagunawardena, T.R., Parslow, P., Williams, S.A.: Dropout: MOOC participants' perspective. In: *The Second MOOC European Stakeholders Summit, EMOOCs 2014*, pp. 95–100 (2014). <http://centaur.reading.ac.uk/36002/>

18. Mayes, T., De Freitas, S.: Review of e-learning theories, frameworks and models. In: JISC E-Learning Models Desk Study (2004). <http://www.jisc.ac.uk/whatwedo/programmes/elearningpedagogy/outcomes.aspx>
19. Reich, J.: MOOC Completion and Retention in the Context of Student Intent (2014). <https://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent/>
20. Renz, J., Schwerer, F., Meinel, C.: openSAP: evaluating xMOOC usage and challenges for scalable and open enterprise education. In: Proceedings of the 8th International Conference on E-Learning in the Workplace (2016). ISBN 978-0-9827670-6-1
21. Rohloff, T., Renz, J., Bothe, M., Meinel, C.: Supporting multi-device e-learning patterns with second screen mobile applications. In: Proceedings of the 16th World Conference on Mobile and Contextual Learning, pp. 25:1–25:8. ACM (2017). <https://doi.org/10.1145/3136907.3136931>
22. Rzepka, S.: Lifelong learning in context - from local labor markets to the world wide web. Ph.D. thesis. Ruhr-Universität Bochum, Universitätsbibliothek (2018)
23. Scheffel, M., Drachler, H., Toisoul, C., Ternier, S., Specht, M.: The proof of the pudding: examining validity and reliability of the evaluation framework for learning analytics. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 194–208. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_15](https://doi.org/10.1007/978-3-319-66610-5_15)
24. Staubitz, T., Meinel, C.: Automated online proctoring of MOOCs - first results. In: Envisioning Report on Quality and Recognition of MOOCs. EADTU (2018, submitted)
25. Wilkowski, J., Deutsch, A., Russell, D.M.: Student skill and goal achievement in the mapping with Google MOOC. In: Proceedings of the First ACM Conference on Learning @ Scale Conference, L@S 2014, pp. 3–10. ACM (2014). <https://doi.org/10.1145/2556325.2566240>
26. Yeomans, M., Reich, J.: Planning prompts increase and forecast course completion in massive open online courses. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, LAK 2017, pp. 464–473. ACM (2017). <https://doi.org/10.1145/3027385.3027416>
27. Zhang, X., Li, C., Li, S.W., Zue, V.: Automated segmentation of MOOC lectures towards customized learning. In: IEEE 16th International Conference on Advanced Learning Technologies (ICALT 2016), pp. 20–22 (2016). <https://doi.org/10.1109/ICALT.2016.25>
28. Zheng, S., Rosson, M.B., Shih, P.C., Carroll, J.M.: Understanding student motivation, behaviors and perceptions in MOOCs. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, pp. 1882–1895. ACM (2015). <https://doi.org/10.1145/2675133.2675217>
29. Zimmerman, B.J.: Models of self-regulated learning and academic achievement. In: Zimmerman, B.J., Schunk, D.H. (eds.) Self-Regulated Learning and Academic Achievement: Theory, Research, and Practice, pp. 1–25. Springer, New York (1989). [https://doi.org/10.1007/978-1-4612-3618-4\\_1](https://doi.org/10.1007/978-1-4612-3618-4_1)



# PsychOut! a Mobile App to Support Mental Status Assessment Training

Carrie DEMMANS EPP<sup>1,2</sup> , Joe Horne<sup>2</sup>, Britney B. Scolieri<sup>2</sup>, Irene Kane<sup>2</sup>,  
and Amy S. Bowser<sup>2</sup>

<sup>1</sup> EdTeKLA Research Group, University of Alberta, Edmonton, Canada  
cdemmansepp@ualberta.ca

<sup>2</sup> University of Pittsburgh, Pittsburgh, USA  
{jhorne, irkl, amy.bowser}@pitt.edu

**Abstract.** Learning about how to assess a patient's mental health status is traditionally performed through lectures. While this conveys information, it does not situate the material in real contexts which promotes knowledge transfer to clinical settings. As a result, a mobile app was developed to help nurse trainees learn about mental status assessment. This app was integrated into a nurse-training program and evaluated for its influence on student learning experiences. Two deployment studies were conducted. Nurse trainee interactions with the app and their perceptions of its appropriateness across these studies indicate this novel technology supported appropriate learning experiences. We discuss the implications our findings have for integrating mobile apps into formal learning settings.

**Keywords:** Mobile learning · Situated learning · Healthcare  
Ill-defined domains

## 1 Introduction

Nurses are expected to assess the mental status of a patient so mental-status assessment training is incorporated into their education programs. This training is typically delivered through lectures [1, 2] that are known for their ability to deliver information rather than their ability to enable the student to apply knowledge in vivo. Consequently, interest in situated learning approaches is growing [1, 2] because it trains learners in the application and use of knowledge.

Given concerns about patient safety [3], it would be inappropriate for nurses to start learning about mental status assessment in a clinical setting. Other approaches to learning how to assess mental status are needed. One approach is to create real-life simulations where students interact with actors playing a role [1]. The cost of this approach is inhibitive. Another training approach, which has yet to be studied widely, is the use of mobile applications (apps) to situate this learning because using highly-contextualized activities promotes knowledge transfer to real-world settings [4]. This class of approaches could enable the transfer of knowledge from classroom settings to clinical settings while providing a safe and low-cost learning environment.

The use of situated learning within apps has supported student learning in some domains (e.g., [5]). However, this approach to supporting nurse training has yet to be studied even though live role-playing activities are beginning to be used within nursing programs [1]. Part of the reason there has been little study of the use of apps to simulate diagnosis activities is because there are so few apps targeting either this task or the training of nurses and others who must perform medical diagnoses.

We discuss how using mobile apps fits with current nurse-training practices and builds on prior computer-assisted learning research before describing a novel tablet app, PsychOut! This app allows students to learn about mental status assessment by performing those assessments. We report on the results of integrating this app within a nurse education program where we conducted two studies of its use. The first study included 60 students, and the second study saw 85 students use the app to support their learning. Study results indicate the potential of PsychOut for supporting nurse education and suggest potential avenues for improving the integration of this app into classroom settings. Based on these findings we contribute to the discussion of how mobile tools can be effectively used in learning environments where students are expected to learn how to navigate and manage complex situations.

## 1.1 Nurse Training Practices

Mental status assessment is initially taught via lectures [1, 2] and supported with readings. This traditional approach to learning is then built upon when nurse trainees enter clinical settings, where they interact with real patients. Some schools additionally use role-play or simulations to help nurse-trainees learn how to assess a patient's mental status [6]. However, a systematic review of teaching practices in nursing found there was poor evidence for the effectiveness of any of the approaches being used [7].

More recently, we have seen an increase in different types of situated learning because its use of highly-contextualized activities supports meaningful learning and promotes the transfer of knowledge to real-world settings [4]. Simulations and role-playing are examples of situated-learning practices that are being adopted within nurse-training programs [1] as are game-based simulations [8]. Some studies have shown that the use these approaches have led to increases in student critical thinking skills [2], knowledge [1], and comfort [1].

## 1.2 Computer-Assisted Situated Learning

The use of situated learning within computer-based applications has been studied for many years [5, 9–11], with guidelines for system development being created [9] and a variety of approaches being taken. In the case of Umka [11], students were meant to discuss different ethical dilemmas and increase their understanding of the issues at hand. Like Umka, those who played Conundrum [10], faced ethical dilemmas that one might encounter in professional contexts. However, these students played through a scenario rather than discussing their options with fellow learners.

These exemplar systems supported student learning within the targeted domains: language learning [12], professional training for programmers [10], and medical

awareness for lay women [5]. However, the context-heavy nature of situated learning [13] means using the same approach across domains may not result in knowledge transfer. For students to benefit from situated learning, it must occur in the context for which they are training [14]. In our case, this is the nursing community of practice.

While this approach has been proposed [15, 16] and occasionally studied [5] in medical domains, it has yet to be studied in nurse-training. Moreover, there has been little study of mobile situated-learning outside of a language-learning context. The one exception to this is the LiveBook system [16], which aims to help doctors learn how to diagnose different physical ailments. However, its effect on learner knowledge has yet to be studied. This means that we do not yet know how to best integrate these types of e-learning systems into medical professionals' training. The limited study of mobile, situated learning and nursing's move towards situated learning make the study of mental status assessment through a mobile, situated-learning app well timed.

## 2 PsychOut!

PsychOut is an iOS tablet app that was designed to complement existing teaching practices and situate learning in realistic scenarios so nursing students could learn to assess a patients' mental health status. To meet this goal, representative clinical situations were selected and a case-based approach to learning was taken. These cases were developed over a year following learner-centred design practices where feedback was sought at multiple stages [17].

Each case aims to teach the learner about a different mental status. The cases cover a range of topics that include depression, delirium, and alcoholism. Learners are given situational information about each patient (Fig. 1 – B and C). This information is provided through a variety of multimedia resources (video, images, text, and audio) and is meant to simulate how a situation might unfold in a clinical setting. In a manner that is consistent with a choose-your-own-adventure story, each case is broken into stages and learners decide how they should proceed (Fig. 1 – D). Learners choose their response from a list of options, proceed to the next stage within a case, and are awarded points. This process continues until the learner has finished the final stage in a case. At which point, the learner receives explicit feedback about how well s/he has handled that mental status assessment situation. Explicit feedback (Fig. 1 – E and F) is provided through scores and the explanations that accompany each scenario option. Implicit feedback takes the form of patient responses to learner-selected actions.



**Fig. 1.** The case selection screen (A), a patient’s case file (B), and the introductory video (C) for the selected case. The screen showing how learner’s select their path through a case (D); the feedback screen showing their score, professionalism, and empathy towards the patient (E); and a detailed review of the choices they made and why those choices were (sub)optimal (F).

### 3 Methodology

The app and its integration into classroom settings followed design-based research practices [18]. We report on the initial (Study 1) and subsequent (Study 2) deployment of PsychOut within a course at a top-twenty nursing school.



### 3.1 Study Procedures

For each study, the app was integrated into four sections of the same course. The course lasted approximately three hours and took place at the end of students’ day: approximately half of the time was dedicated to app use and the other half to their regular lecture. As is typical of these students’ clinical training environments, they had been in class for over 3 hours when the deployment began. Students were given an introduction to PsychOut and shown all of its features. Students were then asked to proceed through four cases in the sequence listed: Jake (teen suicide), Ava (intimate partner violence), Mrs. Peabody (elderly alcohol abuse), and Mr. Smith (delirium).

A basic cross-over design was used (Fig. 2). Students completed a pre-test. They were then divided into two groups: App First and Lecture First. The lecture first students left the room to attend a lecture that was designed to support the same learning objectives as the app. This is the same lecture that is typically used within students’ training program. The App First group stayed in the room. Halfway through the session, students switched learning conditions. Before switching conditions, they completed a second test. At the end of the class period, students completed a post-test and questionnaire. Independent of condition, student behaviors were observed.



Fig. 2. Study activity sequence

### 3.2 Study Instruments and Data Collection

In keeping with a design-based research tradition [18], several small adjustments were made between Study 1 and Study 2. While no major changes were made to the app between studies, several bugs were fixed. Adjustments to study procedures and instruments are detailed within the appropriate sub-sections below.

**Training.** All students received training in app usage before they were asked to complete the assigned cases. This training was modified from Study 1 to Study 2.

*Study 1.* Application features were demonstrated and explained by a member of the research staff (the second author). He did this while proceeding through one of the cases that had not been assigned. Students were given an opportunity to ask questions.

*Study 2.* To increase the consistency in student training, we created a video to provide the same training as in Study 1. Students watched the video and were given an opportunity to ask questions. In addition to increasing consistency, this change should enable instructors to more easily incorporate the app into their classes.

**Questionnaire: Student Perceptions.** Study questionnaires elicited two forms of feedback. The first was open-ended responses that focused on student learning experiences, app usability, and qualities of the cases studied. The second was closed responses

that focused on their perceptions of specific aspects of the app and their learning experience.

*Study 1.* Open-ended questions focused on what students liked or disliked about using the app and the number of cases they completed. Likert-type items (Strongly Agree - 5; Agree - 4; Neutral - 3; Disagree - 2; Strongly Disagree - 1) asked students to rate aspects of usability, the cases they completed, and their confidence in their ability to conduct mental status assessments. The confidence self-assessment was conducted both before and after using PsychOut. Open-ended questions asked students how they navigated the app and how using the app made them feel about their learning.

*Study 2.* Based on students' open-ended responses from Study 1, additional Likert-type items were added to capture information about specific aspects of their app usage and learning experiences. Students were also asked to directly compare their experiences in both the lecture and app conditions, and they were asked to explain some of their ratings for different aspects of the app. The confidence assessment was removed; this aspect of student experience was instead captured through the skipping of questions or guessing of answers on the test as a proxy for student confidence.

**Tests.** The same test form was used for the pre-, mid-, and post-test so that changes in student knowledge could be measured. It was a multiple-choice test. Items were designed so that they were consistent with the type of questions students would encounter on their nurse licensing exam, while keeping the learning objectives of the app in mind. Questions were modified from Study 1 to Study 2 to increase the reliability of the test since about half of the test items from Study 1 exhibited ceiling effects. New questions were developed to replace these items. These new questions were pilot tested and refined using standard item-development procedures [19, 20]. One other change was made to capture student confidence at a more fine-grained level. Students were given the option to indicate that their answer to a test item was a guess.

**Learner Observation.** Multiple (2–3) observers moved around the room, which contained between 10 and 20 students. Observations of different students' interactions were logged over time. These behavioral observations included notes about student body language (e.g., posture and fidgeting), progress through scenarios, physical interactions with the app (e.g., how they touched the screen, whether they skimmed text-based content or replayed sections of videos), their facial expressions, on and off-task behaviors (e.g., texting), and any problems they seemed to experience (e.g., app crashes or questions about the content of a case).

### 3.3 Data Analysis

**Qualitative Analysis Procedures.** Open-ended responses and observations were analyzed following the procedures described in Charmaz's constructive grounded theory [21]. Observational data was triangulated across observers and students to identify consistent behaviors that characterize students' responses to using the app.

**Quantitative Analysis Procedures.** Responses to Likert-type items were analyzed using standard statistical procedures. The paper-based instruments allowed students to abstain from responding, which means responses do not always sum to 85 for Study 2 only. Since all data were not normally distributed, median (Mdn) and inter-quartile range (IQR) are reported. Independent samples t-tests were used to assess differences between groups when the measure met the necessary assumptions. In this case, Cohen's *d* is used to characterize the size of the difference. Otherwise, Mann-Whitney U-tests were used to test differences between groups and *r* is used to characterize the size of the difference. The proportional gain score [22] is used to characterize how student confidence changed following app usage.

**Pre/Post-Test.** The same quantitative analysis procedures were used, and tests were scored by two domain experts. We do not report information about the test results from Study 1 since it was unreliable. The same problems were not observed during Study 2 (i.e., student performance varied) so we report student learning from Study 2.

### 3.4 Participants

Students were about to begin their clinical training in hospital psychiatric units. They were given the option to allow us to use their data but were not allowed to opt out of classroom activities. All students consented to their data being used. Their data was associated with a number and we do not know the student to number mapping.

Study 1 had 60 nurse trainees. Study 1 consent did not include demographic information. Study 2 had 85 nurse trainees; 77 of whom were female. These participants were 22 years old on average ( $SD = 3.36$ ).

## 4 Findings and Discussion

### 4.1 Study 1

There was one student who completed only one case. All others completed at least three cases, with 73% ( $n = 44$ ) of students completing all four cases.

**App Usage.** Student body language during the deployment was indicative of engagement, especially near the beginning of class. Students were leaning over the tablets and had focused facial expressions. Some students also demonstrated emotional responses to specific cases (e.g., furrowed brows, watery eyes, and red faces). These responses were most noted during the teen suicide case and one student commented on how the app was "helpful and insightful because a lot of the correct answers surprised" that student. These responses align with one of the goals of the app, which is to help prepare students for emotionally charged situations since they are likely to encounter them when they enter clinical settings.

Beyond students' affective responses to the app, nurse-trainees exhibited four interaction patterns when selecting options: (1) some carefully reviewed all options before selecting how they would proceed, (2) others quickly skimmed the options and more

deeply read those that seemed reasonable before proceeding, (3) some would read carefully until they found the first response they felt was reasonable and then select that response without viewing all options, and (4) some skimmed and selected without fully reading any one option. These interaction patterns are indicative of broader approaches to information seeking that may require the addition of monitoring or adaptive features to encourage the appropriate use of information.

As to the perceived amount of time spent using the app, some students indicated the “scenarios were too long” and were “time consuming” while others “would have liked to do more cases”. This desire to complete additional cases was observed, with some of the faster students starting to explore a fifth case near the end of class. However, student behaviors more broadly showed signs of fatigue that suggests the app should be used for shorter periods. Fidgeting (e.g., foot tapping and self-grooming behaviors) seemed to noticeably increase around the 45-minute mark. Nearer the end of the class, students were seen rubbing their eyes and yawning. In addition to fatigue, software bugs may have contributed to increases in off-task behavior (e.g., texting or playing on their mobile phones under the desk) as the session progressed.

These behaviors indicate the current method of integrating this app could be improved so that it better meets learners’ attentional needs. These behaviors require instructors pay additional attention to classroom orchestration when integrating software for individual use. While some students quickly recovered from software bugs or crashes, others requested explicit support, while yet others quit interacting with the app, suggesting a need for additional support.

**Perceived Usability.** Students indicated which app elements could be improved. Among them were system stability, the time required to perform situated learning tasks, the immediacy and visibility of feedback, increased difficulty in the scenarios, and a desire to interact with people.

While students felt the app “was very engaging. The scenarios were interesting. The game interface was simple and easy to work with”, there was evidence of the app not fully supporting their needs because the design of certain features was inconsistent with their feedback preferences or resulted in some students not finding a feature. For example, students wanted “to know the reasoning behind why the wrong choices were incorrect”. While the app had a feature that provided this feedback, some students “did not come across ... [the] explanation for the correct answers”. Others did. For example, one learner said, “I liked the use of video scenarios to teach and then that I could go over rationales of questions & keep reviewing”, while another stated that she “enjoyed choosing [her] responses during each round and getting specific ‘empathy’ or ‘professionalism’ points”.

Among those who found the feedback feature, some felt that the provisioning of feedback could have been better timed: “I didn’t like that it didn’t immediately explain why you got an answer wrong and allow you to correct your response”. These learners did not understand how the app was meant to simulate reality where you cannot travel back in time to prevent a mistake; you can only do your best to deal with the consequences of an error and prevent similar errors in the future. It is worth noting that learners can replay any one scenario to explore different diagnosis options. This re-play

functionality can support the exploratory and self-correction needs this learner had, but it requires that a situation be relived in its entirety. Alternatively, introducing a group-level reflection of how their cases proceeded may help these students to better understand the implicit and explicit feedback provided within the app.

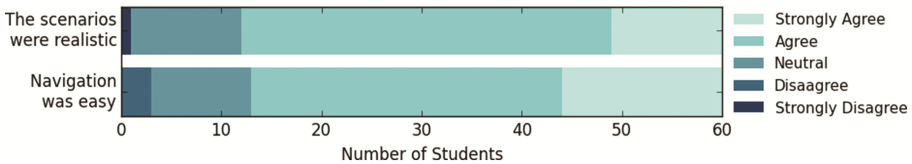
This inconsistency in learner experience and student reports that “some features weren’t discovered until later” or the “controls/directions were a tad vague” imply the app’s interface and case study flow or training need adjustment to ensure students receive the feedback that they desire and need.

**Learner Experience.** Student responses to the questionnaire indicate that they liked using the app because it was interactive, realistic, allowed them to apply their knowledge, and helped them to appreciate the consequences of different choices.

Participants generally found the cases realistic (Fig. 3) and felt the game was easy to navigate. As students said, “it was fun to be able to play through scenarios that can actually happen on the unit” which gave them a “fun and interactive way to learn about mental status assessment and how to communicate with psych patients”. The cases were “very realistic [with] real people” with “answer choices that all seemed legitimate” “in that incorrect answers had unfavorable responses from the patients”.

This realism forced students “to critically think through the different interactions”. The app also gave students “the opportunity to approach situations like a nurse” and “helped [them] visualize what talking to a patient would be like”. Moreover, seeing how “the answers [they] picked were played out and the patient’s response” was perceived to make the app “clear and more engaging/realistic than a typical lecture”. When students found that the cases “seemed unrealistic/fake”, it was because they were “too predictable” or because they could “choose a response from a list”.

Student confidence in their ability to conduct mental status assessments improved by a median of 25.0% (IQR = 50.0) from before using the app (*Mdn* = 3, *IQR* = 2, *Min* = 1, *Max* = 5) to after using it (*Mdn* = 4, *IQR* = 1, *Min* = 1, *Max* = 5), where 1 is not at all confident and 5 is completely confident. This significant gain in student confidence was large ( $U = -4.60, p < .001, r = .59$ ). It was also consistent with their perceptions that the game helped them to see how patients might respond to different actions that they could take as a psych nurse.



**Fig. 3.** Student perceptions of the realism of the cases (top), and an aspect of app usability (bottom).

Some students disliked the app because they were “not a computer learner” and they held the belief that assessment “is better learned by a one-on-one interaction and critique of responses”. As both quotes indicate, students’ pre-existing beliefs about how one

learns influenced their willingness to learn via a mobile app even when that app shared characteristics with their preferred modes of learning. This perceptual barrier suggests the need to include debriefing activities where the student can review his or her app activities with the instructor or a peer. This debriefing strategy could also support deeper learning by encouraging reflection and other regulatory processes or exposing learners to different perspectives [23–25], thus moderating the learning that takes place through app use.

## 4.2 Study 2

For this study, 84 of the 85 students completed all four assigned cases. This improvement in case completion over Study 1 is likely due to the use of a video tutorial to streamline the training process. Students were also consulting the whiteboard to make sure they were getting through all of the assigned scenarios, which may indicate a greater interest in the subject or that students were more performance oriented.

**App Usage.** Observational data indicates students were highly engaged when using the app. This was seen through their slow and careful initial app navigation, time spent on scenarios, repeated attempts at the same scenarios, and trying additional scenarios. This was also supported by their visible emotional responses to the scenarios and the feedback the app gave them about their performance. Some students seemed surprised when their answers were incorrect and commented that this was a benefit of the app: the “app made me want to explore other similar apps to better prepare me for clinical. I thought it was helpful especially seeing how some answers I would think would be right to tell the patient/patient’s family are not.”

Their physical interactions with the app varied, with some sitting while participating and others choosing to stand. They also increased the speed of their interactions as time went on with no one being seen using the fourth interaction pattern from Study 1 (skimming items). Rather, a new interaction pattern emerged where students would return to previous stages in a scenario to back trace their decision points and make sense of how they could have done better. Additionally, many students were seen zooming in on image resources indicating a desire to see more clearly or to obtain more information, which suggests they were more engaged or that their engagement was not harmed as a result of the software bugs that may have disrupted student engagement during Study 1.

**Learner Experience.** Students generally felt they benefitted from interacting with the app: 84 of them thought “the game provided information [they could] use in the real world” and 83 agreed that “seeing the different ways in which someone could respond to a situation was helpful”. Student responses to these Likert-scale items help explain why they ( $n = 82$ ) felt that “the scenarios helped [them] to learn about mental status assessment”, which was a primary goal behind PsychOut’s development.

When comparing the app to the lecture, 41 students agreed the lecture “presented new information”, while 38 felt both formats provided new information. However, students commented that “the lecture was more obvious” even though the app was “very informative”. PsychOut ( $n = 23$ ) performed slightly better than the lecture ( $n = 10$ ) with

respect to helping them “understand some of the challenges of mental health assessment”. Students felt this key learning objective was best met by the app or when the app was used in conjunction with the lecture ( $n = 50$ ). They commented, “the lecture provided factual information while the scenarios gave real life tools.” They also wrote that “the lecture gave [them] good ideas on what to look for in an assessment while the [app] was more of how to approach” conducting the mental status exam. These views demonstrate the complementary nature of the two learning methods with students “wish[ing] the lecture would have taken place before the game.”

**Perceived Usability.** Only 1 of the 85 felt the scenarios were unrealistic, while 80 felt the scenarios were realistic. Students also felt the game was usable, as shown through 76 of them disagreeing with the statement that “the game controls were confusing” and 74 of them agreeing that “it was easy to make [their] way through the scenarios.”

Students encountered a few bugs that were related to memory leaks. These bugs occurred less frequently than those encountered during Study 1. Thus, system bugs did not impede app use to the same extent as in the previous study: Students still required some support, but they became less frustrated and recovered more quickly.

**Student Learning.** Student scores and the number of test-items they skipped can be seen in Table 1, with students from each condition appearing to perform similarly on the pre-test: no differences were detected in their scores ( $t(83) = 0.182, p = .856, d = .039$ ) or the number of test items they skipped ( $t(83) = -0.754, p = .453, d = 0.161$ ).

This lack of measurable difference allows the comparison of their performance on later tests. These comparisons indicate that students benefitted differentially based on whether they used PsychOut or attended the lecture first: those in the lecture first condition skipped fewer items on test 2 ( $t(83) = -2.210, p = .030, d = 0.470$ ). However, there was no measurable difference in their test scores ( $t(83) = 1.376, p = .173, d = 0.296$ ), which means the lecture may have boosted their confidence more than those who used the app without differentially benefitting their learning. This increased confidence in the absence of increased knowledge runs the risk of causing harm, as those with greater confidence are less likely to question their diagnoses even when they are incorrect. Consequently, there is a benefit to using the app because it does not artificially inflate learner confidence in a domain where false confidence can be harmful.

The test results from after students experienced both the game and the app also show no difference in their scores ( $t(83) = 1.585, p = .117, d = 0.376$ ) or the number of items they skipped ( $t(83) = -1.414, p = .161, d = 0.189$ ). This lack of detectable difference indicates neither condition was directly linked to student scores more than the other was. Seeing as student knowledge and confidence increased from the pre-test to the post-test (see the 95% confidence intervals in Table 1), these findings also suggest that combining both approaches might be best for supporting student learning.

**Table 1.** Study 2, student test scores (max. 10) and the number of items they skipped (max. 10) by test and condition.

Condition	Pre			Mid			Post		
	M	SD	95% CI	M	SD	95% CI	M	SD	95% CI
<i>Scores</i>									
App first	7.3	1.86	[6.7, 7.9]	8.30	1.79	[7.7, 8.9]	9.20	1.68	[8.7, 9.8]
Lecture first	7.2	1.81	[6.6, 7.8]	8.69	1.65	[8.1, 9.3]	9.69	1.65	[9.1, 10.2]
<i>No. Items Skipped</i>									
App first	0.9	1.39	[0.5, 1.3]	0.4	0.76	[0.2, 0.6]	0.04	0.21	[0, 0.09]
Lecture first	0.7	0.93	[0.3, 1.1]	0.1	0.30	[0.2, 0.6]	0	0.21	[0, 0.05]

### 4.3 Triangulation Across Studies: Implications for Mobile Integration in Classroom Settings

**App Integration Procedures.** Students generally felt that using PsychOut<sup>1</sup> to complement their learning was beneficial, and Study 2 provides evidence that app usage improves student learning experiences without harming their learning. However, the depth of interaction between students and the content was shallower than desired. This may have been the result of their working independently.

To increase students' depth of interaction, other approaches to integrating the application into the course should be considered. Technology integration plans that encourage socio-collaborative approaches are most likely to encourage an increase in the type of interaction that would benefit students' ability to respond appropriately to complex mental-health assessment situations [26, 27]. These types of interactive activities are also deeply desired by learners when using mobile solutions [28], and they can be used to help overcome other barriers to the effective integration of mobile technologies in educational settings [29].

Specific approaches could include individual work followed by group reflection and comparison (i.e., think-pair-share) to encourage an analysis of their performance and enhance student metacognition. This could help if students received guidance specifying that they should compare the feedback the app gave to each of them based on their actions within the simulation. This type of feedback can be beneficial when mechanisms, such as learning dashboards [30], are put in place to support app use. The pairing of students is one such mechanism that also takes advantage of social comparison to encourage students to focus on and improve their learning [31]. An alternative to this think-pair-share approach would be to have students navigate the scenarios in groups where they could discuss and debate how to proceed. This would expose students to different perspectives, which can support learning in complex settings where there are many issues to consider [11, 23].

Regardless of which integration procedures are explored next, a greater variety of approaches to using apps in classrooms needs to be explored if we are to understand how this technology can effectively support professional training.

<sup>1</sup> App development was funded by the School of Nursing at the University of Pittsburgh.



**Quantity and Timing of App Usage.** Many students, regardless of which study they participated in, felt the app session lasted too long. One way to better support learning and allow those who enjoy learning through technology-enabled scenarios would be to incorporate one or two scenarios into a single classroom session so that students experience more than one content-delivery method at a time. This has the added benefit that student coverage of the materials would be spaced over time, with cases being covered at different points within the term. This type of spacing can support learning by enabling information to be stored in students' long-term memory [32]. It may also help them to change their behaviors within the app because they will be given more time to reflect on their activities, which could help them adjust their approach to interacting with the app and, by extension, their patients.

This integration could be timed to ensure that the scenarios that will be integrated are relevant to students' current needs, which should help improve the engagement of those who preferred lectures over app usage [4]. Synchronizing app usage with other activities in this way will also reinforce the lessons that are being delivered through other media and learned through student interaction in clinical environments, which should support student retention of key information. It could also help to better situate student learning within upcoming clinical experiences, thus helping students to join their professional communities [13].

**Dealing with Problems.** Across both studies, we saw students who were able to recover quickly from software bugs or crashes. However, many students were not as resilient. Others were able to recognize that they needed help and requested it, while yet others gave up, as shown through their ceasing to interact with the app or tablet. These behaviors indicate that additional processes and technologies are needed to monitor student app and activity status. This monitoring would allow instructors to recognize when a student is having problems and intervene as appropriate [29] to support student recovery following an app failure or to bring students back on track should they become distracted after something goes wrong.

**Limitations.** This study relied on human observation of learner interactions with the system. As a result, some learner behaviors may have been missed. The use of multiple observers helps account for this potential limitation and provides a reasonable example of how information about learner activities can be collected when apps do not integrate detailed logging. This pragmatic approach enabled the study of app usage in real classrooms which can later be augmented with app usage logs.

## 5 Conclusion

These early studies explored a novel mobile app to support the situated learning of nursing students in a classroom setting. Study data show app usage contributed to student learning and improved student confidence, but the approach to integrating this learning technology could be improved. Potential improvements include adding group reflections and debriefing activities and shorter periods of use to meet learner attentional needs and better support classroom orchestration. We recently piloted a study to more closely

examine different approaches to integrating apps into nurse training programs. This study should shed additional light on the challenges and benefits of incorporating mobile apps into professional learning environments.


## References

1. Sperling, J.D., Clark, S., Kang, Y.: Teaching medical students a clinical approach to altered mental status: simulation enhances traditional curriculum. *Med. Educ. Online* **18**, 19775 (2013)
2. Tiwari, A., Lai, P., So, M., Yuen, K.: A comparison of the effects of problem-based learning and lecturing on the development of students' critical thinking. *Med. Educ.* **40**, 547–554 (2006)
3. Kanerva, A., Lammintakanen, J., Kivinen, T.: Nursing staff's perceptions of patient safety in psychiatric inpatient care: nursing staff's perceptions of patient safety in psychiatric inpatient care. *Perspect. Psychiatr. Care* **52**, 25–31 (2016)
4. Choi, J.-I., Hannafin, M.: Situated cognition and learning environments: roles, structures, and implications for design. *Education Tech. Research Dev.* **43**, 53–69 (1995)
5. Almeida, T., Wood, G., Comber, R., Balaam, M.: Interactivity: looking at the vagina through labella. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 3635–3638. ACM, New York, NY, USA (2016)
6. Madson, L.: Role-play to teach the mental status exam. *MedEdPORTAL* **7**, 8363 (2011). [https://doi.org/10.15766/mep\\_2374-8265.8363](https://doi.org/10.15766/mep_2374-8265.8363)
7. MacDonald-Wicks, L., Levett-Jones, T.: Effective teaching of communication to health professional undergraduate and postgraduate students: a systematic review. *JBIS Database Syst. Rev. Implementation Rep.* **10**, 1–12 (2012)
8. Wehbe-Alamah, H., et al.: Development of an extensible game architecture for teaching transcultural nursing. *Online J. Cult. Competence Nurs. Healthc.* **5**, 64–74 (2015)
9. Herrington, J., Oliver, R.: Critical characteristics of situated learning: implications for the instructional design of multimedia. In: *ASCILITE*, Melbourne, Australia pp. 253–262, (1995)
10. McKenzie, A., McCalla, G.I.: Serious games for professional ethics: an architecture to support personalization. In: *WKSP on Intelligent Ed. Games at Artificial Intelligence in Education (AIED)*, Brighton, England, pp. 69–78 (2009)
11. Sharipova, M.: Supporting students in the analysis of case studies for ill-defined domains. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) *ITS 2012. LNCS*, vol. 7315, pp. 609–611. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-30950-2\\_86](https://doi.org/10.1007/978-3-642-30950-2_86)
12. Demmans Epp, C.: Mobile adaptive communication support for vocabulary acquisition. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) *AIED 2013. LNCS (LNAI)*, vol. 7926, pp. 876–879. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_135](https://doi.org/10.1007/978-3-642-39112-5_135)
13. Lave, J., Wenger, E.: *Situated Learning Legitimate Peripheral Participation*. Cambridge Univ. Press, Cambridge [u.a.] (2011)
14. Lave, J.: Situating learning in communities of practice. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (eds.) *Perspectives on Socially Shared Cognition*, pp. 63–82. American Psychological Association, Washington, DC (1991)
15. Mentis, H.M., Chellali, A., Schwaizberg, S.: Learning to see the body: supporting instructional practices in laparoscopic surgical procedures. In: *Proceedings of CHI the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2113–2122. ACM, New York, NY, USA (2014)

16. Jalali, S., et al.: LiveBook: competence assessment with virtual-patient simulations. In: IEEE International Symposium on Computer-Based Medical Systems (CBMS), pp. 47–52 (2017)
17. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: *Human-Computer Interaction*. Pearson/Prentice-Hall, Harlow (2004)
18. Anderson, T., Shattuck, J.: Design-based research a decade of progress in education research? *Educ. Res.* **41**, 16–25 (2012)
19. Fowler, F.J.: *Survey Research Methods*. Sage Publications, Thousand Oaks (2009)
20. Spector, P.: *Summated Rating Scale Construction: Introduction*. Sage Publications, Newbury Park (1992)
21. Charmaz, K.: *Constructing Grounded Theory*. Sage Publications, Thousand Oaks (2010)
22. Cattell, R.B.: The clinical use of difference scores: some psychometric problems. *Multivar. Exp. Clin. Res.* **6**(2), 87–98 (1983)
23. Vygotsky, L.S.: *Mind in society: the development of higher psychological processes*. Harvard University Press, Cambridge, MA, USA (1978)
24. Järvelä, S., et al.: Socially shared regulation of learning in CSCL: understanding and prompting individual- and group-level shared regulatory activities. *Intern. J. Comput.-Support. Collab. Learn.* **11**, 263–280 (2016)
25. Zimmerman, B.J.: Models of self-regulated learning and academic achievement. In: Zimmerman, B.J., Schunk, D.H. (eds.) *Self-Regulated Learning and Academic Achievement*. Springer Series in Cognitive Development, pp. 1–25. Springer, New York (1989). [https://doi.org/10.1007/978-1-4612-3618-4\\_1](https://doi.org/10.1007/978-1-4612-3618-4_1)
26. Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., Bier, N.L.: Learning is not a spectator sport: doing is better than watching for learning from a MOOC. In: *Learning @ Scale*, pp. 111–120. ACM Press (2015)
27. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cogn. Sci.* **36**, 757–798 (2012)
28. Demmans Epp, C.: Migrants and mobile technology use: gaps in the support provided by current tools. *J. Interact. Media in Educ. Spec. Collect. Migr. Educ. Technol.* **2017**, 1–13 (2017)
29. Demmans Epp, C., Phirangee, K., Despres-Bedward, A., Wang, L.: Resourceful instructors and students: overcoming barriers to integrating mobile tools. In: Power, R., Ally, M., Cristol, D., Palalas, A. (eds.) *IAMLearning: Mobilizing and Supporting Educator Practice*, p. E-book. IAmLearn (2017)
30. Demmans Epp, C., Bull, S.: Uncertainty representation in visualizations of learning analytics for learners: current approaches and opportunities. *IEEE TLT* **8**, 242–260 (2015)
31. Hsiao, I.-H., Bakalov, F., Brusilovsky, P., König-Ries, B.: Open social student modeling: visualizing student models with parallel introspective views. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) *UMAP 2011*. LNCS, vol. 6787, pp. 171–182. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-22362-4\\_15](https://doi.org/10.1007/978-3-642-22362-4_15)
32. Settles, B., Meeder, B.: A trainable spaced repetition model for language learning. In: *Association for Computational Linguistics (ACL)*, pp. 1848–1858 (2016)



# Exploring Gamification to Prevent Gaming the System and Help Refusal in Tutoring Systems

Otávio Azevedo<sup>(✉)</sup>, Felipe de Moraes<sup>(✉)</sup>, and Patricia A. Jaques<sup>(✉)</sup> 

Programa de Pós-Graduação em Computação Aplicada (PPGCA),  
Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, Brazil  
{obazevedo,felipmoraes}@edu.unisinos.br,  
pjaques@unisinos.br

**Abstract.** Intelligent Tutoring Systems (ITSs) have shown to be almost as effective as one-to-one tutoring. Nonetheless, the students' improper use of the ITS help system and its intelligent assistance, i. e. gaming the system or help refusal, can impair learning. This paper presents the use of gamification elements, more specifically, points and difficulty levels, as an approach to prevent the behaviors of gaming the system (help abuse and trial-and-error) and help refusal. This system was integrated into a step-based algebraic ITS and it was evaluated in an experiment, during six weeks, involving 60 students from three classes of the 7<sup>th</sup> year of an elementary school. Each class of students was assigned to one of the three groups: fully gamified, partially gamified and non-gamified, being that they differ by the level of gamification implemented. The students in the two gamified groups had a lower rate of trial-and-error behavior than the non-gamified group. However, we haven't found statistically significant difference between the fully and partially gamified groups for the trial and error. Also, no differences were observed between the gamified groups and the non-gamified one for the help refusal and help abuse behaviors. The results of this research confirm previous finding that gamification can be used as a non-restrictive approach for the trial-and-error behavior, a form of gaming. On the other hand, we were not able to show that gamification can prevent help refusal and abuse.

**Keywords:** Gamification · Intelligent Tutoring System · Help abuse  
Help refusal · Gaming the system

## 1 Introduction

Gamification has its origin in digital games and can be defined as the use of game elements in systems or contexts that are not games, aiming at encouraging the participation of users [12]. This technique can be applied in many different areas [7], including education. In the context of education, gamification has been mainly used for increasing learners' motivation and engagement.

An Intelligent Tutoring System (ITS) is a computer learning program designed to provide individualized assistance for students during problem solving. ITSs have shown to be almost as effective as one-to-one tutoring [17]. Typically they offer individualized assistance for students in three main ways. First, they provide minimal feedback, which informs whether a solved step or the final answer is correct. If the step/answer is wrong, the tutor provides an error feedback that returns a help message. Students can also request a hint manually; in this case, the system returns a help message, guiding the students to solve the next step of the exercise or to achieve the final answer. Both feedback and hints are provided by the ITS help system.

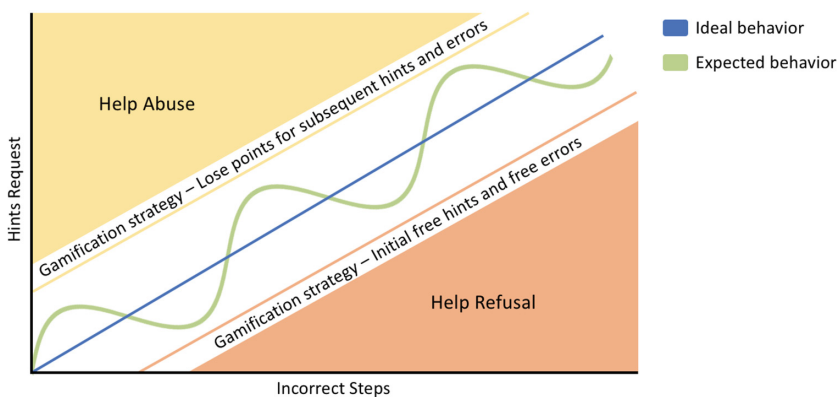
In spite of its efficiency, the rich help system of ITS and their intelligent assistance can lead students to unwanted behaviors, such as gaming the system. Students who game try to mislead the tutor to advance faster. This behavior generally happens in two different ways [1]: (i) when the student takes advantage of the progressing hint system of an ITS to have the answer (help abuse); (ii) when the student tries several possible answers in a task without making cognitive efforts to solve it (trial-and-error behavior). On the other hand, students may refuse to ask for help from the system (help refusal), usually because they mistakenly believe that asking for help shows lack of intelligence [16]. Both behaviors impair learning. Baker [1] observed that students that game the system scored lower in the assessment than who did not. Students who do not ask for help when they have some difficulties may also perform poorly, as they will not be able to make adequate progress in the proposed activities [16]. Gamification can change user behaviors, as many research studies have already shown [2,8,9], including the more specific ones such as gaming the system [14].

In this context, the general goal of this work is to analyze whether the elements of gamification can minimize the students' resistance to request help or their tendency to game the system. To achieve this goal, a computational model of gamification has been developed and integrated into a step-based intelligent tutoring system, PAT2Math [11] (<http://pat2math.unisinos.br>). The ITS assists students to solve first-degree equations by providing three types of assistance for each step: minimal feedback, error-feedback, and hints. Three different versions of the ITS have been implemented. The first version consisted of the tutor with no elements of gamification (non-gamified). The second (partially gamified) and the third (fully gamified) versions are gamified and use points and levels as gamification elements. The main difference between the gamified versions is that the second version only shows a score per equation and it discounts points for any error or hint request that the student has made, whereas the third version, besides a score per equation, has a total score and also provides some free errors and hints points. In this way, while the partially gamified version only handles gaming the system (by always discounting points for any error or hint request), the fully version prevents students from help refusal, by providing some initial free errors and hints points, besides discounting points for additional errors or hint requests to also avert gaming behavior. Furthermore, we also wanted to

check whether the more extensive use of the gamification elements (for example, total score besides points per equation) has a stronger impact in gaming and help refusal behaviors.

Gamification has been investigated as an engagement strategy in ITSs [5, 13, 15], having achieved promising results in the short-term. However, despite its broad use, there is insufficient evidence to support the long-term benefits of gamification in educational contexts [4]. In the specific context of behaviors related to the help system, the work of [14] used gamification to minimize gaming the system. The results of a 17 months duration experiment showed a strong impact of the gamification on the decrease of gaming the system, observed by the student's permanence time in the solution of mathematical problems. The students spent more time solving problems when using the gamified system (67% more than in the tutor without gamification elements).

Nevertheless, an open question is how to use gamification to prevent gaming the system (trial-and-error and help abuse) without making students avoid requesting help when needed, i.e., help refusal. Another important open question is to verify whether a more exhaustive use of gamification elements have a stronger impact on students' gaming and help refusal behaviors, i.e., more gamification leads to less gaming the system and less help refusal. This paper describes an experiment to verify whether it is possible to use gamification as a strategy to prevent gaming the system and, at the same time, as a strategy to encourage students requesting for help when they need to, avoiding the help refusal behavior.



**Fig. 1.** Gamification strategies to balance between help refusal and help abuse (Color figure online)

Figure 1 illustrates the use of gamification as a strategy to create a balance between help abuse and help refusal. The ideal behavior (blue line) shows a perfect balance between the number of incorrect steps and a student's request for hints; it maximizes students learning. However, the student is expected to have

behavior that is not perfect, but considered as expected. The expected behavior (green line) illustrates the behavior of a student that does not commit a large number of errors without asking for help, falling into the help refusal behavior, nor that s/he requests an excessive number of hints, falling into the help abuse behavior. However, the number of hints requested will not always be optimal as in the ideal behavior. Thus, two gamification strategies were implemented in the fully gamified version of the system to prevent students from the help refusal or gaming behaviors. The first strategy is to provide free hints and free errors for the student to motivate her/himself requesting help when needed. The second strategy is to penalize subsequent hints and errors through the punctuation system so that the student does not request more help than it is necessary and commit mistakes in a trial and error manner.

## 2 The Different Versions of the ITS

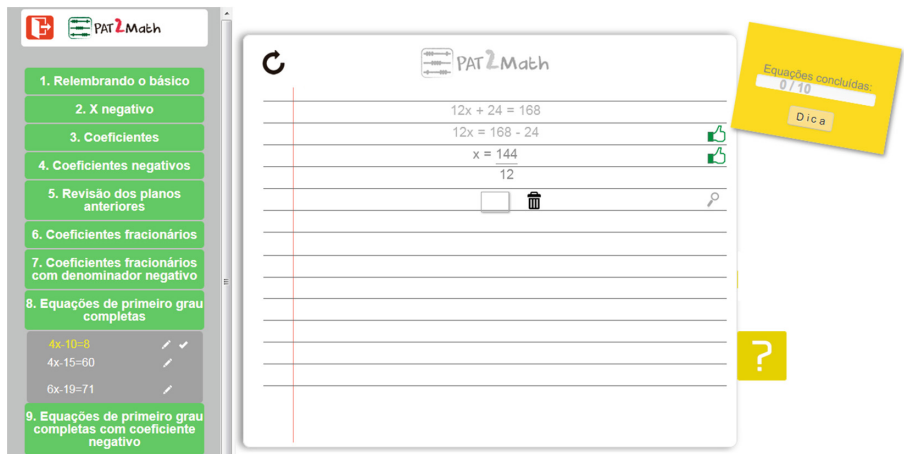
Three versions of PAT2Math, an algebraic step-based ITS, have been implemented for the experiment: one non-gamified version, containing the core and the basic functions of the ITS, and two gamified versions (called partially and fully gamified), having the same features of the non-gamified version plus some game elements. The game elements are the same for both gamified versions. However, the difference is in the amount of gamification applied to each of them and in the strategies employed.

The following sections give an overview for each version of the system, ending with a summary of the differences between them.

### 2.1 Non-gamified Version

The non-gamified version is used as the base for the other versions. It has no gamification elements, neither strategies to avoid help abuse or refusal behaviors. To use the ITS, based on an initial equation, the students use the keyboard of the computer to enter the steps to solve the equation. For every step given, the ITS returns a feedback to the student. This feedback could be just an ok (minimal feedback) if the step is correct, or an error feedback otherwise. Besides, the student can request for hints, if s/he would need some help to proceed. The ITS has a level-based hints system that follows a Point, Teach and Bottom-out approach [10], with the former offering more generic hints and which require greater reasoning, intermediate ones showing more refined hints, and the last level indicating the answer for the current step. The Fig. 2 shows the non-gamified version of the system with two correct steps.

In the algebraic ITS, the equations for the student to solve are distributed in equations plans, and each plan has the equations of the same format. The plans are organized in increasing level of difficulty, with the first plan being the simplest (with equations in the format  $x + b = c$ , where  $b$  and  $c$  are constants) and the latter being the most complex, with equations involving fractional numbers and distributive property to be solved. Students must solve all equations in the



**Fig. 2.** Interface of the non-gamified version of the ITS

current plan to unlock the next. There are two types of equations plans: content or revision. Content plans contain similar equations, i.e., equations that involve the same algebraic operations to be solved and are rated at the same difficulty level. These plans have 5 to 10 equations, depending on their complexity. Differently, review plans have equations that cover several previous plans aiming to reinforce student learning. Review plans have between 10 to 20 equations, depending on the amount of reviewed contents and their complexities. Only for content plans, the first equation of every plan is a worked example.

Since there are no game elements in the non-gamified version, such as points, students can see the worked example whenever they want without cost. In addition, hints and errors are also free. Thus, students can request hints and make mistakes as many times as they want until the system gives them the answer.

## 2.2 Partially Gamified Version

The partially gamified version was implemented based on the non-gamified version. Thus, the tutor works in the same way. The only difference is in the insertion of gamification. In this version, regardless of the level of the current equation plan, the student loses 3 points for each hint request and 5 points for each error, which aims to prevent help abuse and trial-and-error solving. The partially gamified version does not handle help refusal. In this version, students can visualize the worked examples as many times as they want without losing points.

Equations plans were gamified in both gamified versions: they were distributed in five levels of difficulty - Basic, Intermediate, Advanced, Expert and Season Finale - and have titles that resemble the names of electronic game phases. Each of the levels, except for the last one that contains equations from all previous levels, has three to four content plans and one review plan. The buttons colors of the levels and plans, the points display and the concluded equations



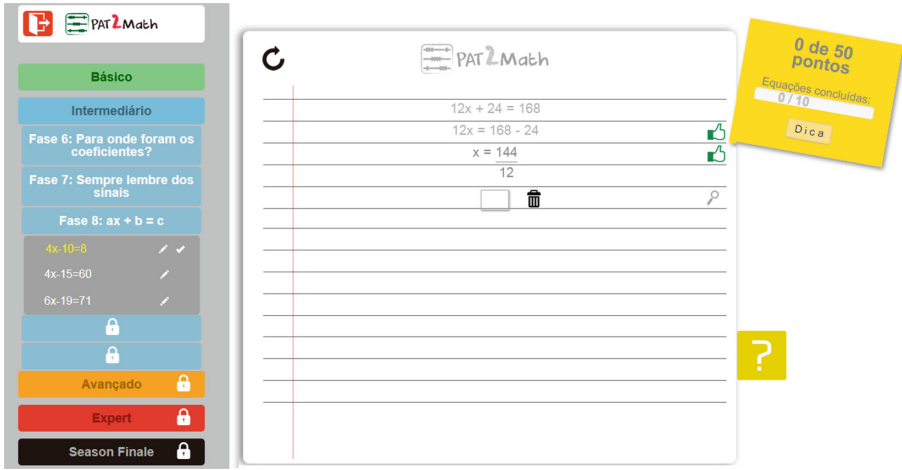


Fig. 3. Interface of the partially gamified version of the ITS

bar are the same for the fully gamified version. Figure 3 shows the interface of the partially gamified version, with different levels and subsequent levels locked.

### 2.3 Fully Gamified Version

Again, the fully gamified version was implemented based on the partially gamified version; therefore it handles help abuse and trial-and-error behaviors. However, in this version, some additional strategies were implemented to also prevent help refusal. Students have a certain amount of hints that can be requested for free, i.e., without losing any points (free hints); in the same way, students can perform a certain number of errors without losing points (free errors). The number of free hints and errors was defined according to the level of difficulty and the number of equations of each plan: the more difficult an equation plan is and/or more equations it has, the more points it will count (because students will need more steps to solve it and each step counts 5 points) and the more free hints and errors it will have. When the student uses all free hints, the next requested hints will discount students' points according to their level of detail. Offering free hints and discounting points for hints should provide a balance between help refusal and help abuse behaviors. In the fully gamified version students also have access to one free preview of the worked example for each equations plan, for the first time a plan is selected. But, they have the option to skip viewing the worked example.

Figure 4 shows the fully gamified ITS interface. Scores by level (Fig. 4a), a level is formed by several equations plans with same difficulty level, and total scores (Fig. 4b) were added in the equations plan menu. The buttons colors of the levels (Fig. 4c) and plans (Fig. 4d) are associated with their respective complexity: the stronger the color, the greater the level of difficulty. The interface also

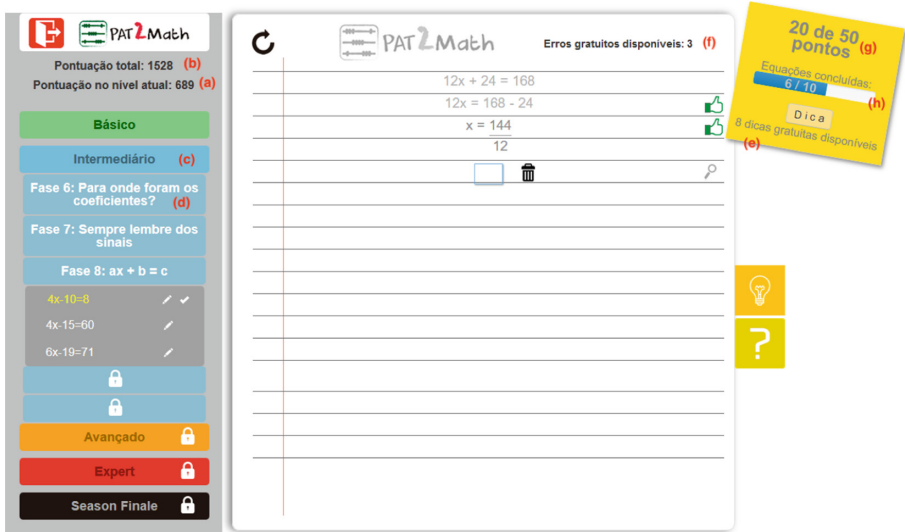


Fig. 4. Interface of the fully gamified version of the ITS

features an indicator of free hints (Fig. 4e) and free errors (Fig. 4f). These indicators disappear from the interface when they reach zero, i.e., when the student has no more free hints or free errors. We opted for removing this information because students may choose not to use them to avoid losing points, thus leading to help refusal. It is also possible to visualize the points for the current equation (Fig. 4g) and also the number of equations solved in the current plan (Fig. 4h). The last two scores (Fig. 4g and h) are only available in the fully gamified version.

## 2.4 Differences Between the Three Versions of the Tutor

This section summarizes the main difference between the three versions. The non-gamified version has no gamification elements; hence it does not control gaming the system or help refusal behaviors. Both partially and fully gamified versions use points and difficulty levels as gamification elements. However, while the partially gamified version only aims to prevent help abuse and trial-and-error solving by discounting points for any error or hint request; the fully gamified version also offers some free errors and hints to handle help refusal. The fully gamified version, besides showing the points per equation, also shows the accumulative total number of points. Table 1 shows the gamification characteristics for each version of the developed ITS for the experiment.

### 3 Evaluation

An evaluation was conducted to verify the impact of the gamification elements (points and difficulty levels) in reducing gaming the system (help abuse and trial-and-error solving) and help refusal behaviors.

**Table 1.** Gamification characteristics for the three versions of the ITS

<b>Current features in all the three versions</b>			
New system of worked examples, where the student decides at the beginning of each lesson plan if s/he wants to check the worked example (the visualization of the worked example was mandatory previously).			
	<b>Fully Gamified</b>	<b>Partially Gamified</b>	<b>Non-Gamified</b>
<b>Hints</b>	It allows free hints according to the complexity level of the lesson plan. Exceeding hints cost 3 points.	Each requested hint costs 3 points, without free hints.	There are no points and other rewards or game mechanisms, so the hints, errors and worked examples have no costs.
<b>Error Feedbacks</b>	It allows errors without loss of points (free errors) according to the difficulty of the lesson plan. Exceeding errors cost 5 points.	Each error costs 5 points, without free errors.	
<b>Worked Examples</b>	First preview in each lesson plan is free, the others cost 8 points.	All views are free.	
<b>Points</b>	Total, by difficulty level and by equation.	Only by equation.	
<b>Difficulty Levels</b>	Well-defined difficulty levels.		Difficulty levels not defined
<b>Lesson Plan Titles</b>	Lesson plan titles as if they were phases of a game.		Lesson plan titles indicate only their content.

The experiment was carried out in a private school in the south of Brazil. Three seventh-grade classes took part in the experiment (total of 60 students, 30 boys and 30 girls, 12 to 14 years old) during five weeks, where each class had a weekly session that lasted 50 min. Each student used one of the three versions of the ITS in the computer lab of the school, which counted on one computer per student, for four weeks. In the fifth week, students were invited to fill out a questionnaire to verify how much students enjoy to use the ITS.

Due to the apparent differences between the three versions of the tutor’s graphical interface, the students were not randomly assigned to the ITS different versions (non-gamified, partially, fully). If they perceive they were using different versions of the program, the evaluation could be negatively impacted (i.e., Hawthorne effect). In this way, each class of students was assigned to a different version of the tutor, which characterizes this work as a quasi-experiment [3]. A class with 18 students (10 boys and 8 girls) used the non-gamified version; another class with 20 students (9 boys and 11 girls) used the partially gamified version of the tutor; and the third class with 22 students (11 boys and 11 girls) had access to the fully gamified version. The three classes had the same math teacher. Thus, the content, teaching methodology, and student assessment were the same for all classes.

For the students’ log data to be used in this evaluation, a consent term, validated by the ethics committee of our university, was given for the parents to sign. In addition, the experiment occurred during a period in which the school proposed to use the ITS as a complementary activity, this being an initiative of the school managers. Thus, after the period specified by this experiment, the students continued to use the system as a complementary activity to the classroom. In the case of parents choose not to let the student participate in the experiment, the data of that student was not considered in our research. However, in the case of our experiment, all parents signed the consent term.

## 4 Results Analysis and Discussion

The data extraction of the experiment was done through system logs and a questionnaire administered at the end of the experiment. From the log of students interactions in the ITS, it was possible to identify the students most likely to game the system and refuse help, as well as to estimate the intensities of these behaviors. The questionnaire was developed by us and aimed to verify how much students liked to use the tutor version and also their opinion about gamification and the tutor’s characteristics.

### 4.1 Gaming the System and Help Refusal

To identify the level of the trial and error, and help refusal and abuse behaviors, we created three formulas with the help and validation of two math teachers, who have 10 and 20 years of teaching experience. These formulas are given in Table 2.

**Table 2.** Formulas for identifying help refusal and gaming behaviors

Gaming the System		Help Refusal
Trial and Error	Help Abuse	
$3 \times \left( 1 + \frac{\frac{\#errors\ in\ sequence}{intervals\ between\ errors\ mean\ (in\ seconds)}}{10} \right)$	$\frac{\#hints + \#errors}{\#errors \times 6}$	$\frac{\#errors\ in\ sequence\ without\ requesting\ hints}{\#errors}$

The trial and error formula assumes that students can miss up to twice in an equation step without being categorized into trial and error, since the first error may have been a typing error or a simple lack of attention. From the third error in sequence, there is a greater chance to characterize gaming the system; for this reason, the numerator of the formula is divided by 3. It is also necessary to take into consideration the time between the sequential errors, since the closer the errors are in sequence, the greater the probability of trial

and error (considering the hypothesis that the student did not analyze correctly the wrong step before retrying). Thus, teachers suggested that students should wait at least 10s between mistakes to have enough time to figure out why they miss. Finally, we added the number 1 in the multiplication so that the trial and error coefficients do not get too high if the student has the average time between errors of less than 10.

The hints abuse formula returns the percentage of hints used concerning all errors made by students. When a student misses a step, s/he also gets a hint, which is why the number of errors is also considered in the numerator of the formula. The denominator follows the following principle: for each error, the student could have asked four hints, considering that the hints of each algebraic operation have four levels. In the case of error feedback, they usually have two levels, which in addition to the other four levels of the hints results in six possibilities. Thus, the more hints the student uses, the greater the chance of help abuse. It is important to note that it is acceptable for students to use up to half of the available hints without it being considered help abuse. Finally, the help refusal formula is based on the following principle: the greater the number of sequential errors without requesting for hints, the greater the chance that the student will exhibit this behavior; if the student had asked for help s/he probably would not have missed so many times in a step.

All the results from the three formulas were tested against Shapiro-Wilk normality test, presenting significantly non-normal distributions. Thus, assuming non-parametric tests, the help refusal and abuse were tested using Kruskal-Wallis for the formulas of trial and error, help abuse and help refusal of students using the ITS version as the factor (Table 3). Kruskal-Wallis revealed a significant difference among the different versions for the trial and error behavior, a type of manifestation for gaming the system ( $H(2) = 10.26, p = .0059$ ).

**Table 3.** Kruskal-Wallis and posthoc tests for trial and error, help refusal and help abuse formulas values

Formula	FG		PG		NG		Kruskal-Wallis		Multiple Comparison KW		
	M	SD	M	SD	M	SD	H(2)	p	FG×NG	FG×PG	PG×NG
Trial and error	.83	2.208	.66	1.1703	1.41	3.5751	10.26	.0059	True	False	True
Help Abuse	.19	.025	.20	.0261	.21	.0221	4.5893	.1008	-	-	-
Help Refusal	1.92	2.6273	1.69	1.7034	2.06	3.1174	.0014	.9993	-	-	-

FG = Fully Gamified, PG = Partially Gamified, NG = Non-Gamified, M = Mean, SD = Standard Deviation, H(2) = statistics for the Kruskal-Wallis test with 2 degrees of freedom, p = p-value, diff = difference is significant

Posthoc comparisons were conducted using multiple comparison test after Kruskal-Wallis [6]. Results of pairwise comparison showed that students in the non-gamified version performed more trial and error than students in the partially (difference=66.76) and fully (difference=64.03) gamified versions, for the critical differences equal to 57.48 and 53.76, respectively. However, no differences were found between the fully and the partially gamified versions (difference=2.73), for the critical difference equals to 53.04. In all cases, assuming

$\alpha = .05$  correct for the number of tests. These results indicate that there is no difference between the gamified versions, partially and fully gamified. However, there is a difference between the non-gamified and the gamified versions of the ITS when trying to reduce the trial and error behavior. One possible explanation for this result is that to enrich the gamification of a system does not lead to less trial and error behavior. Another possible explanation is that the additional elements were not enough to stand out the partial gamification group, due to the great similarity between the two versions. For help abuse and refusal, no difference was found between the three versions. Therefore, we were not able to show that gamification can prevent help refusal and abuse.

### 4.2 Questionnaire

We elaborated a personalized questionnaire with 36 questions that aim to verify how much students liked to use the ITS version and their opinions about hints and gamification, besides questions related to help refusal and help abuse behaviors. The questions followed a Likert scale, with scores ranging from 1 to 5. Due to space limitation, we chose to show only the 9 most relevant questions in Table 4.

Unfortunately, it was not possible to use a questionnaire already recognized in the literature, since we are not aware of validated and free instruments in the Portuguese language to identify help behaviors and engagement in learning software.

**Table 4.** Kruskal-Wallis and posthoc tests for the questionnaire items

Question	FG			PG			NG			Kruskal-Wallis		Multiple Comparison KW			
	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	H(2)	p	FG×NG	FG×PG	PG×NG	
1. Do you like to solve equations in the tutoring system?	4.77	5	.5284	4.50	5	.607	4.76	5	.5623	4.2578	.119	-	-	-	
2. Would you like to use the tutoring system to do your homework?	4.54	5	.671	4.50	4.5	.513	4.71	5	.4697	1.4747	.4784	-	-	-	
3. Do you prefer to solve equations in the tutoring system instead of paper and pencil?	4.43	5	1.0757	3.67	4	1.3904	4.47	5	.8745	5.8578	.0535	False	False	False	
4. Were the hints useful?	3.68	4	.9455	4.05	4	.6048	3.94	4	.9663	1.4711	.4792	-	-	-	
5. The hints helped me to best understand the equations.	3.90	4	1.1792	4.10	4	.7881	4.18	4	1.0146	.5999	.7408	-	-	-	
6. I requested hints to get faster at the final answer.	1.86	2	1.0372	1.55	1	.6863	2.18	2	1.2367	2.53	.2822	-	-	-	
7. When I had difficulties, I would rather try to solve the equation by myself instead of requesting a hint.	3.95	4	1.1329	3.85	4	1.04	4.00	4	.7906	.2751	.8715	-	-	-	
											Wilcoxon rank sum test				
											W	p			
8. The points system turns the tutoring system funnier.	4.18	4	1.0065	3.75	4	.9105	-	-	-	154.5	-	-	.0827	-	
9. One of my main goals was to achieve a good score.	4.00	5	1.3452	3.35	3.5	1.268	-	-	-	149.5	-	-	.0666	-	

FG = Fully Gamified, PG = Partially Gamified, NG = Non-Gamified, M = Mean, SD = Standard Deviation, H(2) = statistics for the Kruskal-Wallis test with 2 degrees of freedom, p = p-value, diff = difference is significant

As illustrated in Table 4, the three groups liked to use the system, and, in most questions, the means were very close. In order to verify if there was any significant difference between groups, the Kruskal-Wallis test was applied to every question data. Kruskal-Wallis test was chosen instead of ANOVA because the data were tested for normality using Shapiro-Wilk, and the result for all the questions was above .05 [6]. No statistic significance was found for any of the questions. Question 3 reached a marginally significant result. However, through the multiple comparison Kruskal-Wallis applied in this question data, it was possible to see that even close, there was no significant difference between the partial (difference=9.68) and fully (difference=.75) gamified versions against the non-gamified version, with critical difference equals to 13.41 for both. Also, no difference between the gamified groups was found (difference=10.43), for the critical difference equals to 12.69. In all cases, assuming  $\alpha = .05$  correct for the number of tests.

Questions 6 and 7 are directly related to the general goal of this work, and the non-gamified group presented the highest mean, which is in agreement with the results showed in the previous section. Related to items 8 and 9, these two questions were related to just the gamified versions of the system. The data was collected just from the students who were in the gamified groups, fully or partially. Again, Shapiro-Wilk was used to test for normality. The results of the test showed that the data does not follow a normal distribution. Thus, Wilcoxon rank-sum test was applied instead of using an unpaired t-test [6]. The results of the test showed a marginally significant evidence for both questions. So, there is evidence that the students of the fully gamified group liked the scoring system better (Q8),  $W = 154.5$ ,  $p = .0827$ ,  $r = -.2677$ , and they were more competitive than the partially gamified group (Q9),  $W = 149.5$ ,  $p = .0666$ ,  $r = -.283$ . Both results present a small to medium effect size.

## 5 Conclusion

Intelligent tutoring systems are getting almost as effective as one-to-one tutoring due to the evolution of technology and artificial intelligence [17]. They have a help system that can guide students to solve problems of the most varied contents and disciplines, and they can also help students to correct mistakes in real time. However, if this intelligent assistance is not used correctly, it can generate undesirable effects, and even impair students' learning [1]. The most unwanted well-known effects are gaming the system (help abuse and trial and error solving) and help refusal.

This paper presented a solution proposal to the previously mentioned behaviors, which may be caused by the inappropriate use of the help system and minimal feedback offered by the ITS. This misuse is usually caused by students' lack of motivation in the studied subject, or by the act of studying itself [1]. Students may abuse the minimal feedback and hints to get to the final answer faster and without reflecting on the problem, or they may not use them because they believe asking for help is bad or wrong. Therefore, the developed system aims to minimize the occurrence of these behaviors using gamification.

The elements of gamification inserted into the ITS acted in the following aspects: (*i*) providing free hints and errors (with no loss of points) based on the complexity of the lesson plan and the number of equations; (*ii*) distribution of lesson plans in levels of difficulty and phases with gamified names; and, (*iii*) the optimization of the scoring system already present in the ITS, including the total score and level of difficulty. The partially gamified version aims to handle help abuse and trial and error behavior by making the student lose points for any hint request or error. In the fully gamified version, the tutor prevents help refusal by offering some initial free hints and free errors but continuing to prevent help abuse and trial and error solving by discounting points for subsequent hints and errors.

In order to verify this goal, a quantitative experimental evaluation was carried out involving three classes of the 7<sup>th</sup> grade of a Brazilian private school. Each class was associated with a group, which could be fully gamified, partially gamified and non-gamified, thus being considered a quasi-experiment [3]. Although we have not found any difference among the groups in relation to the help refusal and help abuse behaviors, the students of the gamified groups, fully and partially, presented a lower index in the trial and error behavior when compared to the non-gamified group. However, the gamified groups did not present a relevant difference between them, which may mean that more gamification does not necessarily lead to a more significant change in the trial and error behavior. On the other hand, the questionnaire showed that the students of the fully gamified group liked more the scoring system and were more engaged to score better than students in the partially-gamified group. Possibly, the richer scoring system of the fully gamified version was more attractive for students.

The results found in our study corroborate with the findings of [14], which shows that gamification can be used to minimize gaming the system. However, [14] compared the time students spent reading hints and the task statement, while in our work we compare the intensity of the behaviors themselves. Besides, in our study, although we were able to find an impact of the gamification on the trial and error behavior, we haven't found any difference between the two gamified versions. It possibly means that more gamification does not necessarily lead to less trial and error. In addition, we were not able to show an impact of the gamification in the help refusal or help abuse behaviors.

Nevertheless, it is still possible to elaborate alternative explanations for the results we have found: (*i*) the two gamified versions are similar, (*ii*) limitations of the design of the quasi-experiment (one of the classes could have students who are more comfortable with the content), and (*iii*) short duration of the experiment. Another limitation of this study is related to the formulas to detect help refusal and gaming. These formulas were elaborated with the help of experienced teachers who have previously used the tutor with their students, but we have not formally validated them. Although we believe that they can offer an approximate estimate of help refusal and gaming in a first study, future works should develop validated mechanisms for detection of these behaviors. As future work, we also want to study the impact of different elements of gamification on these same behaviors, as well as on students' engagement and learning.



**Acknowledgements.** This work is supported by the following research funding agencies of Brazil: CAPES, CNPq and FAPERGS.

## References

1. Baker, R.S.: Designing intelligent tutors that adapt to when students game the system. Ph. D. thesis, Carnegie Mellon University, Pittsburgh (2005)
2. Barna, B., Fodor, S.: An empirical study on the use of gamification on IT courses at higher education. In: Auer, M.E., Guralnick, D., Simonics, I. (eds.) ICL 2017. AISC, vol. 715, pp. 684–692. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-73210-7\\_80](https://doi.org/10.1007/978-3-319-73210-7_80)
3. Campbell, D.T., Stanley, J.C.: Experimental and Quasi-Experimental Designs for Research. Ravenio Books, Cortina (2015)
4. Dichev, C., Dicheva, D.: Gamifying education: what is known, what is believed and what remains uncertain: a critical review. IJAIED **14**(1), 9 (2017)
5. Faghihi, U., Brautigam, A., Jorgenson, K., Martin, D., Brown, A., Measures, E., Maldonado-Bouchard, S.: How gamification applies for educational purpose specially with college algebra. Procedia Comput. Sci. **41**, 182–187 (2014)
6. Field, A., Miles, J., Field, Z.: Discovering Statistics using R. Sage, London (2012)
7. González, C., Mora, A., Toledo, P.: Gamification in intelligent tutoring systems. In: International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 221–225. ACM Press, New York (2014)
8. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work? – a literature review of empirical studies on gamification. In: HICSS, pp. 3025–3034. IEEE (2014)
9. Huang, W.H.Y., Soman, D.: Gamification of Education. Behavioural Economics in Action. Rotman School of Management, University of Toronto (2013)
10. Hume, G., Michael, J., Rovick, A., Evens, M.: Hinting as a tactic in one-on-one tutoring. J. Learn. Sci. **5**(1), 23–47 (1996)
11. Jaques, P.A., Seffrin, H., Rubi, G., Morais, F., Ghilardi, C., Bittencourt, I.I., Isotani, S.: Rule-based expert systems to support step-by-step guidance in algebraic problem solving: the case of the tutor PAT2MATH. Expert Syst Appl. **40**(14), 5456–5465 (2013)
12. Kapp, K.M.: The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education. Wiley, New York (2012)
13. Long, Y., Alevan, V.: Gamification of joint student/system control over problem selection in a linear equation tutor. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 378–387. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07221-0\\_47](https://doi.org/10.1007/978-3-319-07221-0_47)
14. Pedro, L., Isotani, S.: Explorando o Impacto da Gamificação na Redução do Gaming the System em um Ambiente Virtual de Aprendizagem. In: Anais dos Workshops do V CBIE (Concurso de Teses e Dissertações), pp. 81–90 (2016)
15. da Rocha Seixas, L., Gomes, A.S., de Melo Filho, I.J.: Effectiveness of gamification in the engagement of students. Comput. Hum. Behav. **58**, 48–63 (2016)
16. Vanlehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. **16**(3), 227–265 (2006)
17. Vanlehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)



# Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese

Valter Neto<sup>1</sup>, Vitor Rolim<sup>1</sup>, Rafael Ferreira<sup>1,2(✉)</sup>, Vitomir Kovanović<sup>3</sup>,  
Dragan Gašević<sup>2,4</sup>, Rafael Dueire Lins<sup>1</sup>, and Rodrigo Lins<sup>1</sup>

<sup>1</sup> Departamento de Computação, Universidade Federal Rural de Pernambuco,  
Recife, Brazil

{valter.neto,vitor.rolim,rafael.mello,rafael.lins,  
rodrigo.linsrodrigues}@ufrpe.br

<sup>2</sup> University of Edinburgh, Old College, South Bridge, Edinburgh EH8 9YL, UK  
{rafael.ferreira,dragan.gasevic}@ac.ed.uk

<sup>3</sup> University of South Australia, 160 Currie St, Adelaide, SA 5000, Australia  
Vitomir.Kovanovic@unisa.edu.au

<sup>4</sup> Monash University, 19 Ancora Imparo Way, Clayton, VIC 3800, Australia  
dragan.gasevic@monash.edu

**Abstract.** This paper presents a method for automated content analysis of students' messages in asynchronous discussions written in Portuguese. In particular, the paper looks at the problem of coding discussion transcripts for the levels of cognitive presence, a key construct in a widely used Community of Inquiry model of online learning. Although there are techniques to coding for cognitive presence in the English language, the literature is still poor in methods for others languages, such as Portuguese. The proposed method uses a set of 87 different features to create a random forest classifier to automatically extract the cognitive phases. The model developed reached Cohen's  $\kappa$  of .72, which represents a "substantial" agreement, and it is above the Cohen's  $\kappa$  threshold of .70, commonly used in the literature for determining a reliable quantitative content analysis. This paper also provides some theoretical insights into the nature of cognitive presence by looking at the classification features that were most relevant for distinguishing between the different phases of cognitive presence.

**Keywords:** Community of Inquiry (CoI) model · Content analytics  
Online discussions · Text classification

## 1 Introduction

The adoption of Learning Management Systems (LMSs) has increased significantly in the last few years [30]. Such systems provide resources that can enable social interactions between students, as well as between students and their teachers. Among the resources available in LMSs, asynchronous discussion forums are

widely used for encouraging student course participation, answering questions, and sharing resources [17]. Online discussions play an important role in the educational experience of students, especially in fully online learning courses, given the absence of face to face interactions.

The Community of Inquiry (CoI) model [14] emphasizes the social nature of modern online learning and it is one of the most researched and validated pedagogical model in the domain of distance education. It defines three constructs (known as presences) that shape students online learning, with the central construct being the *cognitive presence*, which captures the development of the critical and in-depth thinking skills [14] of the students. The *Quantitative Content Analysis (QCA)* method [37, 42] is widely adopted to assess the three CoI presences, making valid and reliable inferences from the analysis of textual data [5]. The CoI model defines three QCA coding schemes, one for each presence which can be used to analyze the discussion messages of the students online at the three presence levels. Although widely adopted in the social sciences within CoI community, content analysis has been primarily used for retrospection and research after the courses are over, without much impact on the actual student learning and outcomes [41]. In this regard, automated methods for text analysis commonly used within learning analytics [13] have a potential for making an assessment of CoI presences easier and less labor intensive, with the ultimate goal of using CoI model to drive instructional interventions and affect student learning outcomes [21].

There have been promising approaches for automating the assessment of cognitive presence [8, 22, 23, 31, 44], but the focus of those studies have been exclusively on English language courses, limiting their use to English-speaking countries only. Likewise, the availability of text analytics tools to languages other than English is even more limited, causing a significant deleterious effect on the accuracy of the systems developed for those languages. The different student demographics and course context within non-English courses can have a substantial effect on the predictive power of the developed analytics. The growing need for high-quality education in developing countries, implies in the need to examine how such findings can be replicated within courses in languages other than English and how analytics findings can be used for supporting students in non-English-speaking countries.

This paper describes the results of the study which examined the use of automated text analytics methods for assessing the cognitive presence from online discussion transcripts written in Portuguese. The study was based on the previous work within English-language courses [22, 24, 44] and adopted a similar classification approach, albeit with some modifications due to the differences between English and Portuguese text analytics tools. The classification method of Kovanović et al. [24] was successfully adopted showing some evidence of the potential of employing existing text analytics to non-English courses. Moreover, despite of the fact that Portuguese analysis tools and libraries are slightly less developed, the classification accuracy of 83% and Cohen's  $\kappa$  of .72 obtained in the experiments performed were better than the ones reported by the previous studies [22, 24, 44] showing the role of the context on the final analytics findings. The results and their implications are further discussed in this paper.

## 2 Background

### 2.1 The Community of Inquiry (CoI) Model

The Community of Inquiry (CoI) model is a widely adopted framework that describes the different facets of students' online learning [15]. Three dimensions or presences provide an overview of online learning experience: (i) *Cognitive presence* captures the development of desirable learning outcomes such as critical thinking, problem-solving, and knowledge (co-)construction [14, 16]; (ii) *Social presence* focuses on social interactions within a group of students (i.e., cohesion, affectivity, and open communication) [36]; and (iii) Teaching presence encompasses the instructors' role before (i.e., course design) and during (i.e., facilitation and direct instruction) a course [2]. This study focuses on the cognitive presence, which captures the development of critical and deep-thinking skills [14]. The cognitive presence is operationalized through a four-phase model of practical inquiry by Lipman [29]:

- 1 *Triggering event*: A problem or dilemma is identified and conceptualized. In an educational context, discussions are usually triggered by instructors; however, they can also be initiated by any participant in the discussion.
- 2 *Exploration*: The students explore the potential solutions to a given problem, typically by information seeking and brainstorming different ideas.
- 3 *Integration*: The students synthesize new ideas and knowledge by employing social (co-)construction.
- 4 *Resolution*: Finally, students solve the original dilemma or problem triggered at the beginning of the learning cycle. Here, students evaluate the newly-created knowledge through hypothesis testing, vicarious application, or consensus building.

Despite the fact that the CoI model is well established as a very effective model for assessment of social interactions in distance learning, the coding process requires a considerable amount of manual work which leads to a problem related to the scalability of its adoption [12]. The development of the CoI survey instrument [4] was one effort to reduce the need for manual content analysis of the discussion messages. However, the CoI survey instrument relies on self-reported data which makes it not applicable for real-time monitoring and guidance of student learning. Thus, automatic methods for coding are essential to enable a broader adoption of the CoI model.

### 2.2 Automating Cognitive Presence Analysis

Within the published literature, there have been several studies that looked at the automation of cognitive presence content analysis. Early proposals based their approach primarily on word and phrase counts [8, 31], such as the ones provided by the General Inquirer category model [40] adopted by Mcklin [31] or fully custom dictionaries adopted by Corich et al. [8]. Using such an approach Mcklin [31], the performance figures achieved 0.69 in Holsti's Coefficient

of reliability [18] and in 0.31 Cohen’s  $\kappa$ . Similarly, reference Corich et al. [8] reported in 0.71 Holsti’s coefficient of reliability, albeit using a sentence-level coding and assessment rather than the more widely used message level.

Some more recent studies examined the use of other different features and classifiers. Kovanović et al. [22] examined the use of a combination of bag-of-words (n-gram) approach and Part-of-Speech (POS) N-gram features for classifying cognitive presence using the Support Vector Machines (SVMs) classifier. While the authors reported 0.41 Cohen’s  $\kappa$ , they also pointed out at the issue of high class imbalance (lower level exploration messages are much more common than other three types of messages), as well as overfitting the data with very high number of features (more than 20,000) on a comparatively small dataset (1,747 messages). In order to address those challenges, Kovanović et al. [24] proposed the use of features based on Coh-Metrix [32], LIWC [43], LSA similarity, named entities, and discussion context [44]. Thereby, the authors reduced the feature space from more than 20,000 features to just 205 features. In their study, Kovanović et al. [24] developed a random forest classifier [6], which also allowed for the analysis of the influence of the different features on the final classification results. For example, their findings indicated that longer and more complex messages were generally more closely related to higher levels of cognitive presence, whereas question marks and first-person singular pronouns were indicative of the lower levels of cognitive presence. This work reached the best classification values (0.63 Cohen’s  $\kappa$ ) so far reported in the literature [24].

Since the focus of this study is on examining the use of text analytics for assessing the cognitive presence online discussion messages in Portuguese, studies that addressed the CoI model within Portuguese online courses were also examined. Although, there are some studies that looked at the CoI model within Portuguese courses [3,38], there is no publication that looked at the automation of cognitive presence assessment neither in Portuguese, nor for any language other than English, to the best of the knowledge of the authors of this paper.

**Table 1.** Course topics by weeks.

Week	Theme	Messages (%)
1	Uses of microscopes	511 (34.06%)
2	Cell theory	400 (26.66%)
3	Genetics	314 (20.93%)
4	DNA and cloning	275 (18.35%)
Total		1,500 (100.00%)

**Table 2.** Distribution of cognitive presence.

ID	Phase	Messages (%)
0	<i>Other</i>	196 (13.07%)
1	Triggering event	235 (15.67%)
2	Exploration	871 (58.07%)
3	Integration	154 (10.27%)
4	Resolution	44 (2.92%)
Total		1,500 (100.00%)

### 3 Method

#### 3.1 Dataset

The dataset used in the research reported here, comes from a biology undergraduate-level course offered through a fully online instructional condition at a Brazilian public university. The dataset has 1,500 discussion messages produced by 215 students over four weeks of the course (Table 1). On average, each student produced seven messages containing 89 words on average. The purpose of the online discussions was on a theme proposed by the instructor, with participation accounting for 20% of the final course mark. However, the discussions were mostly of the type question-answer rather than online debates. The whole dataset was coded by the two coders for the four levels of cognitive presence enabling for a supervised learning approach. The inter-rater agreement was excellent (percent agreement = 91.4% and Cohen's  $\kappa = 0.86$ ). A third coder resolved the disagreements (128 in total).

Table 2 shows the distribution of the four phases of the cognitive presence, along with the category “other” which was used for messages that did not exhibit the indicators of any cognitive presence phase. The most frequent were exploration messages, accounting for more than 58% of the data, while the least frequent were resolution messages, accounting only for 2.93% of the data. The substantial difference between the frequencies of cognitive presence phases was expected [15] and also reported in the previous studies of the CoI model [22, 24].

There are several explanations for this pattern [1]. In this particular case, the forum showed characteristics of a question-answer discussion. Thus, it does seem reasonable that students will spend more time asking questions (triggering event) and especially exploring different answers (exploration). Moreover, as discussions were designed to occur between the first and the fourth week of the course, students did not typically move onto the resolution phase that early in the course.

#### 3.2 Feature Extraction

This work follows the same approach presented by Kovanović et al. [24], in which traditional text classification features (e.g., N-gram, POS, dependency triplets) were not adopted in order to: (i) decrease the number of features, reducing the chances for over-fitting the training data; (ii) the traditional features are very “dataset dependent”, as data itself defines the classification space; (iii) N-grams and other simple text mining features are not based on any existing theory of human cognition related to the CoI model; such features can lead to models which hard to understand their theoretical meaning.

Kovanović et al. [24] evaluated 205 features mainly based on LIWC [43] and Coh-Metrix [32]. As the resources and tools for Portuguese text analytics are limited, only 87 features were explored, but all of the best ones found in [24] were included.

**LIWC Features.** The LIWC (Linguistic Inquiry and Word Count) tool [43] extracts a large number of word counts which are indicative of different psychological processes (e.g., affective, cognitive, social, perceptual). As there is no implementation of LIWC for Portuguese, the features extracted were the ones that: (i) reached the best results for the state-of-art cognitive presence classifier in English [24], and (ii) can be analyzed using NLP techniques (i.e., given the dictionary-based approach of LIWC, some words can only be empirically determined as representative of the psychological processes). A total of 24 features adapted from LIWC were extracted.

**Coh-Metrix Features.** Coh-Metrix is a computational linguistics tool that provides different measures of text coherence ((i.e., co-reference and structural cohesion) linguistic complexity, text readability, and lexical category use [32]. Coh-Metrix has been adopted in the collaborative learning domain, for example, to predict the student performance [9] and the development of social ties [20] based on the language used in the discourse. The Portuguese version of Coh-Metrix [39] has 48 different measures (while the English version has 108). It is important to mention that the features that are missing in the Portuguese version have not achieved good results in the cognitive presence classification for English.

**Discussion Context Features.** In order to incorporate more context to the feature space of the current study, the features proposed by Waters et al. [44] and used by [24] were included: (i) *Number of replies*: An integer variable indicating the number of responses a given message received; (ii) *Message Depth*: An integer variable showing a position of a message within a discussion tree; (iii) *Cosine similarity to previous/next message*: The idea of these features is to obtain how much the current message builds on the previously presented information; (iv) *Start/end indicators*: It uses an indicator (0/1) showing whether a message is first/last in the discussion.

The features above are relevant to the problem under study due to the process nature of the CoI model [15], in which the students' cognitive presence is viewed as being developed over time through discourse and reflection. Moreover, due to the social-constructivist view of learning in the CoI model, the different phases of the cognitive presence tend to change over time. Thus, one expected that triggering and exploration messages would be more frequent in the early stages of the discussions, while integration and resolution messages would be more common in the later stages.

**Word Embedding Similarity.** Kovanović et al. [24] made a parallel about the cognitive phases and the information presented in the various stages of the learning process. In summary, the triggering phase introduces a topic, while the exploration phase introduces new ideas and answers. The integration phase keeps talking about the same ideas (by constructing the meaning from the ideas

previously introduced), and resolution concludes the discussion presenting the explicit guidelines for applying knowledge constructed [33].

Due to the reasons listed above, it is beneficial to have a feature that can identify if the context of each message changes over time in a discussion. The main difference related to the original work by Kovanović et al. [24] is that the current study adopted word embeddings to represent the word similarity instead LSA. In brief, word embeddings are neural networks algorithms to translate words into numerical vectors based on their occurrences in a text [26]. Thus, the problem of identifying the relationship between words becomes a simple measure of the cosine similarity between their vectors. In the current study, the word embeddings algorithms and trained dataset available in the spaCy tool<sup>1</sup> were applied.

**Number of Named Entities.** Previous work in the literature suggested that the number named entities (e.g., named objects such as people, organizations, and geographical locations) would be different for the different phases of cognitive [14]. Exploration messages, which are characterized by the exploration of new concepts and opinions, are expected to bear more named entities than integration and resolution messages. The spaCy library<sup>2</sup> was used to extract the number of named entities.

### 3.3 Data Preprocessing

The first step of the data analysis performed here divided the data into training and test datasets (75% and 25% of the whole corpus, respectively), as often done in machine learning [11]. This step was performed to avoid overestimating the model performance which can occur if the model accuracy estimated on the same data as the model parameters [11] were learned. It is important to mention that stratified samples concerning coding categories (i.e., Triggering, Exploration, Integration, Resolution, and Other) were created to preserve their distribution in both train and test subsets. The split dataset included 1,125 and 375 instances for the training and test datasets, respectively (Table 3).

After the corpus partitioning, the problem of class imbalance was addressed as shown in Table 3. The imbalance can lead to negative effects on the results of the classification analyses [35]. In this step, the approach suggested by Kovanović et al. [24] was followed, using the SMOTE algorithm [7], which creates additional synthetic data points as a linear combination of the existing data points. The SMOTE processes the data points in an  $n$ -dimensional feature space (for instance  $X = f_1, f_2, f_3, \dots, f_n$ ) of a specific class selected for resampling as follows: (i) Find  $K$  (in our case five) nearest neighbors of  $X$  belonging to the minority class chosen; (ii) Randomly select one of the identified neighbors (called  $Y$ ), (iii) Generate a new synthetic data point  $Z$  as:  $X + r * Y$  where  $r$  is a random number between 0 and 1.

<sup>1</sup> <https://spacy.io>.

<sup>2</sup> <https://spacy.io>.



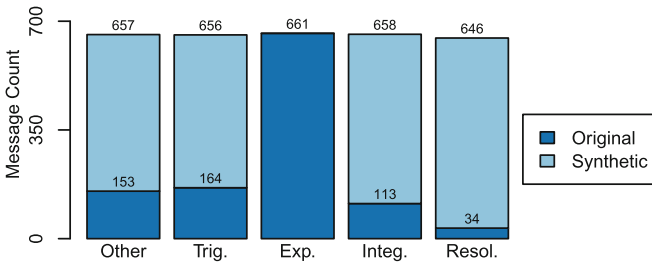
Figure 1 presents the final result of the SMOTE algorithm application in the training set. The size of the codes Other, Triggering Event, and Integration, were increased 4 to 5-fold, while the Resolution category was increased 19-fold (from 34 to 646).

**Table 3.** Distribution of coding categories in test and train data sets.

Phase	Dataset				
	Train		Test		Total
Other	153	(13.6%)	43	(11.47%)	196 (13.07%)
Triggering event	164	(14.58%)	71	(18.93%)	235 (15.67%)
Exploration	661	(58.76%)	210	(56%)	871 (58.07%)
Integration	113	(10.04%)	41	(10.93%)	154 (10.27%)
Resolution	34	(3.02%)	10	(2.67%)	44 (2.92%)
Total	1125	(100%)	375	(100%)	1500 (100%)

### 3.4 Model Selection and Evaluation

There are several machine learning techniques to build supervised models. Fernández-Delgado et al. [10] performed a sizeable comparative analysis of 179 general-purpose classification algorithms over 121 different datasets identified that random forests and Gaussian kernel SVMs were the top performing algorithm. This work adopted the random forests because it is a white-box algorithm in addition to its excellent performance. This means that it is possible to evaluate the extent to which each feature contributes to the classifier [6].



**Fig. 1.** SMOTE preprocessing for class balancing.

The main idea of the random forest classifier is to combine a large number of decision trees that depend on a random independently sampled vector with the same distribution for all trees. With such a mechanism, the algorithm maintains a low variance without increasing the bias [6]. It is important to mention that each tree is constructed on a different bootstrap sample of the training data,

and evaluated on the data points that were not included in the initial sample. The outcome is decided using a simple majority voting scheme.

As previously stated, the random forest algorithm allows the evaluation of the importance of the classification features. In this context, the most used measure is Mean Decrease Gini (MDG) index, which accounts for the separability of a given feature regarding the categories [6].

Finally, the two parameters used in the random forest classifiers [6] were set up: (i) `ntree`: the number of trees generated by the algorithm; and (ii) `mtry`: the number of random features selected by each tree. Here, different values for each parameter were evaluated over the training data using 10-fold cross-validation. In both cases, the values that maximize the final performance were selected.

### 3.5 Implementations

The classifier was mainly coded in Python and in R programming languages. The key software packages and libraries used were:

- spaCy<sup>3</sup>, for natural language processing,
- Coh-Matrix, the Portuguese version by Scarton et al. [39],
- scikit-learn [34], for stratified sampling of test and train data,
- randomForest R package [28], for classifier development, and
- caret R package [25], for model training, selection, and validation.

## 4 Results

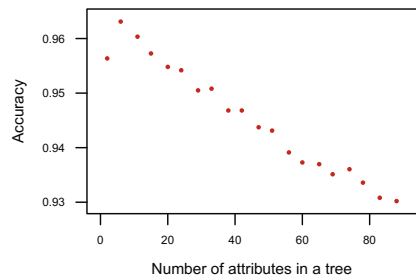
### 4.1 Model Training and Evaluation

Figure 2 shows the results of the tuning procedure performed in the random forest model. In the best case, the proposed classifier achieved a performance of .96 (SD = .01) classification accuracy and Cohen's  $\kappa$  of 0.95 (SD = .01). This result was reached with six features per decision tree on the training dataset (`mtry` = 6).

**Table 4.** Parameter tuning summary.

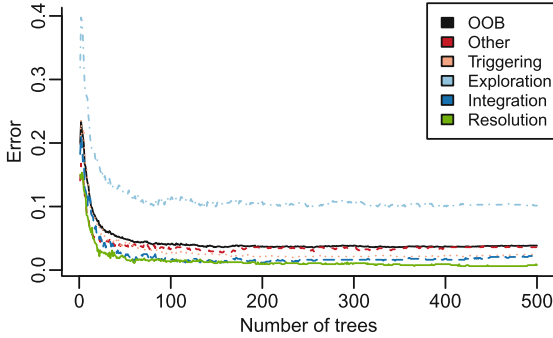
	<code>mtry</code>	Accuracy	Kappa
Min	87	0.93 (0.01)	0.91 (0.02)
Max	6	0.96 (0.01)	0.95 (0.01)
Difference		0.03	0.04

**Table 5.** Parameter tuning results.



<sup>3</sup> <https://spacy.io>.

The improvement between the best- and worst-performing model was 0.03 and 0.04 for classification accuracy and  $\kappa$  respectively, which shows the importance of the parameter optimization in the final performance.



**Fig. 2.** Best random forest configuration performance.

Table 5 shows the performance of the random forest model using the optimal  $m_{try} = 6$  on the training set. There are three essential results to be analyzed in this figure: (i) the selected number of trees (500) is enough to guarantee a good classifier performance, as it stabilized with a little less than 100 decision trees; (ii) the average out-of-bag (OOB) error rate reached result

**Table 6.** Test data confusion matrix without the SMOTE application.

Actual	Predicted					
	Other	Triggering event	Exploration	Integration	Resolution	Error rate
Other	39	0	2	2	0	0.09
Triggering event	5	62	4	0	0	0.12
Exploration	3	2	197	8	0	0.06
Integration	1	0	24	16	0	0.60
Resolution	0	0	10	0	0	1.00

**Table 7.** Test data confusion matrix with the SMOTE application.

Actual	Predicted					
	Other	Triggering event	Exploration	Integration	Resolution	Error rate
Other	39	0	2	2	0	0.10
Triggering event	5	62	4	0	0	0.13
Exploration	3	2	197	8	0	0.07
Integration	1	0	24	16	1	0.61
Resolution	0	0	9	1	0	1.00

under .1, suggesting that less than 10% of the data points were misclassified; (iii) the highest error rate was observed for Exploration; this result was expected as this category was not resampled.

Tables 6 and 7 present the confusion matrix for the test data, the 25% that was left as the holdout (Table 3), before and after the application of the SMOTE algorithm. Both tables show the same result, where the error rate for the Exploration is the lowest, followed closely by the error rate for the Triggering event and Other. The tables also show that Integration and Resolution were mostly misclassified. This probably happened because these two phases had the smallest number of instances in the test dataset (Table 3), making hard for the classifier to effectively learn how to recognize messages in the two-phase.

Finally, it is important to notice that the proposed random forest model obtained .83 classification accuracy (95% CI[0.79, 0.86]) and Cohen's  $\kappa$  of 0.72 on the test set, which is considered a "substantial" agreement above the level of pure chance [27].

## 4.2 Analysis of the Feature Importance

This study also analyzed the contributions of the different features to the final performance of the classifier. Figure 3 shows the MDG scores for all classification features. It is possible to recognize that 50% of the features reached MDG score below than median (25.26) and 65% obtained an MDG score lower than the average (29.55). On the other hand, some features achieved very high MDG scores reaching 154.65 for the best feature.

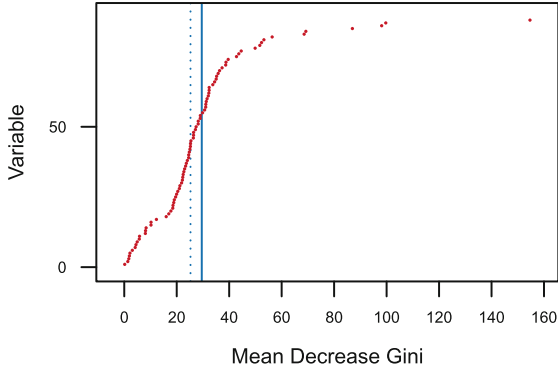
Table 8 presents a detailed analysis of top twenty most relevant features. Although 87 features were evaluated, 54 had above average MDG scores; thus, due to space limitations, only the top twenty were analyzed here. It is important to note that LIWC was not used, as there is no Portuguese version of it; Thus, some features were re-implemented. The *liwc* prefix was used to refer to the features that were based on the original implementation of LIWC.

One can see that the most relevant variable was *liwc.QMark* (the number of question marks in a message), which is directly related to the Triggering phase. The average sentence length, average word per sentence, number of words and number of words bigger than six letters, number of tokens showed a similar trend, with higher values associated to Exploration and Resolution, followed closely by Integration.

Several conclusions can be drawn from the Coh-Matrix features analyzed. First, the givenness (i.e., how much information in a text is previously given) had the highest association with the higher levels of cognitive presence. The highest values for the variables of lexical diversity of the student vocabulary (VOCD and content words) were found to "other" messages. Finally, the variables related to content words and type to token ratio reached the highest values for Other and Triggering.

Regarding the features based on LIWC, they were mainly based on quantitative values (number of articles, prepositions, quantifiers, and pronouns) achieving the highest values in the exploration and resolution phases.

Finally, the variable related to the position of the message within a discussion obtained the highest values for other and triggering. This result is not usual, and the design of the discussion (debate and question-answers with a large number of the instructor’s interventions) can justify it. Most of the triggering messages were posted by the instructor trying to encourage the engagement of the students. Integration and Resolution also reached high values due to the fact that these phases usually happens after triggering and exploration messages.



**Fig. 3.** Feature importance by Mean Decrease Gini (MDG) measure. Dotted blue line shows median MDG score (25.26), while solid blue line shows average MDG score (29.55).

## 5 Discussion

The evaluation of the automatic classification of cognitive presence over the testing dataset showed that the features based on LIWC and Coh-Metrix are effective to classify forums message in Portuguese. Cohen’s  $\kappa$  of 0.72 represents a “substantial” inter-rater agreement [27], and it is above the 0.70 Cohen’s which is the CoI research community commonly used as the threshold limit required before coding results are considered valid. The optimization of the *mtry* parameter (i.e., the number of attributes used in each tree of the forest) improved the final result for 0.04 Cohen’s  $\kappa$  and 0.003 classification accuracy (Table 4). Although the authors of this paper did not find any other related work which performed a similar analysis to compare, it is important to mention that the approach presented here reached accuracy results better than the classifiers of cognitive presence developed for English [22, 24, 44].

This study conducted a detailed analysis of the features used. First, the model was trained on only 87 features and did not use a bag-of-words vector as an attribute. Thus, the chances of over-fitting the training data decrease substantially. To draw any future conclusions about the generalizability of the classifier, it will be important to apply it to different subject domains and pedagogical

**Table 8.** Twenty most important features for distinguishing between cognitive presence phases and their values in different cognitive presence phases.

#	Variable	Description	MDG	Other	Cognitive presence phase			
					Triggering	Exploration	Integration	Resolution
1	liwc.QMark	Number of question marks	154.66	0.07 (0.36)	1.17 (0.80)	0.17 (0.81)	0.08 (0.37)	0.20 (0.82)
2	cm.AveSen	Average sentence length	99.69	5.98 (3.48)	8.15 (4.17)	25.3 (15.4)	23.5 (14.0)	25.7 (10.8)
3	message.depth	Position within discussion	98.12	2.65 (1.22)	2.58 (1.17)	1.52 (1.01)	2.56 (1.17)	2.18 (1.63)
4	liwc.6Word	Number of words bigger than six letters	86.96	3.04 (2.65)	5.17 (6.63)	40.4 (36.4)	17.6 (17.4)	39.9 (28.8)
5	cm.WPerSen	Avg. word per sentence	69.23	8.31 (9.53)	10.5 (6.29)	27.3 (17.9)	25.8 (16.6)	27.2 (11.0)
6	liwc.Art	Number of articles	68.58	0.86 (1.31)	1.81 (2.19)	13.6 (12.8)	5.64 (5.85)	12.7 (10.2)
7	cm.Tokens	Number of tokens	56.39	11.5 (10.98)	19.1 (20.9)	127 (111)	60.0 (51.6)	131 (95.8)
8	cm.Giveness	Average givenness of each sentence	53.24	0.47 (0.18)	0.54 (0.19)	0.73 (0.13)	0.73 (0.13)	0.69 (0.11)
9	liwc.PreP	Number of prepositions	52.33	1.16 (1.78)	1.92 (2.74)	17.1 (15.8)	7.32 (6.42)	17.9 (12.4)
10	cm.NumWord	Number of words	51.67	12.0 (12.89)	19.1 (20.5)	126 (111)	59.8 (52.2)	131 (95.3)
11	liwc.Conj	Number of conjunctions	49.89	0.35 (0.75)	0.97 (1.36)	6.80 (6.73)	3.56 (3.85)	7.43 (5.57)
12	cm.MContWord	Min. among content words frequency	44.58	520 (123)	213 (69.2)	69.3 (72.3)	39.1 (10.3)	38.4 (9.15)
13	liwc.Verb	Number of verbs	43.66	1.38 (1.69)	2.79 (3.44)	18.3 (16.4)	9.27 (7.89)	19.3 (14.6)
14	cm.LDVOCD	Lexical diversity, VOCD	42.79	0.69 (0.17)	0.59 (0.12)	0.42 (0.07)	0.46 (0.08)	0.43 (0.07)
15	cm.LDContW	Lexical diversity, content words	39.62	0.65 (0.15)	0.57 (0.11)	0.40 (0.06)	0.43 (0.05)	0.43 (0.07)
16	cm.TTR	Type to token ratio	38.79	0.95 (0.13)	0.95 (0.07)	0.74 (0.11)	0.84 (0.10)	0.76 (0.12)
17	liwc.Quant	Number of quantifiers	38.67	0.56 (0.90)	0.43 (0.81)	3.13 (3.13)	1.85 (1.95)	3.95 (3.57)
18	cm.ContWord	Content words frequency	37.29	730 (186)	629 (109)	587 (57.5)	599 (65.0)	590 (46.4)
19	liwc.3Pron	No. of pronouns in third person singular	36.33	0.04 (0.19)	0.13 (0.41)	1.55 (2.13)	0.67 (1.02)	1.68 (1.91)
20	cm.PronNP	Mean pronouns per noun phrase	35.81	0.02 (0.07)	0.04 (0.11)	0.02 (0.05)	0.05 (0.07)	0.01 (0.02)

contexts. Second, the results indicated that a small subset of features had highly predictive indicators of the different phases of cognitive presence (Fig. 3).

It is important to highlight that the most relevant classification indicators (Table 8) were aligned with the theory of cognitive presence [19]. Higher levels of cognitive presence were related to messages that are: (i) longer, with more words and sentences; (ii) complex, with complex words (words bigger than 6 letters) and longer sentences; (iii) have lower lexical diversity, as shown here by two measures of lexical diversity; (iv) have higher givenness of the information; (v) use more third-person singular pronouns; (vi) use fewer question marks. The conclusions drawn above are consistent with the findings of previous studies, for instance, 45% of the top 20 features found in the current study match those found by Kovanović et al. [24]. Future research is needed to better understand the reasons behind the differences in contributions of the features across different studies.

Finally, one can see that the Other category produces indicators with values close to the triggering phase. The Other category had messages with general requests, solicitation, or course exception rather to contribute towards knowledge construction about topics discussed. Such a category had large diversity in relation to other messages (as seen in lexical diversity and TTR features) and tended to be more informal (with fewer words, and sentences). Besides that, Other messages occurred more towards the end of a discussion, which is expected as many students would use their final post for thanking each other for their contributions.

## 6 Final Remarks

This paper has two main contributions. First, a new classifier to code students' transcripts on the level of cognitive presence for messages written in Portuguese was proposed. The developed approach obtained 83% accuracy and Cohen's  $\kappa$  of 0.72 which is considered substantial agreement above the level of pure chance [27]. This result shows the potential to provide an automated system for coding cognitive presence in Portuguese.

Second, a detailed relevance analysis of the proposed features was presented, which were mainly based on Coh-Metrix and LIWC. In such a context, the experiments performed showed that long and complex messages, along with bigger givenness and more use of third-person singular were related to higher levels of cognitive presence. Higher lexical diversity and a greater number of question marks were associated with lower levels of cognitive presence. Such conclusions corroborate the results of the related work [24].

The main limitations of the approach presented here are related to the dataset. First, the collected data was from a single study domain (i.e., biology) with discussions designed with a particular pedagogical purpose (i.e., question-answer discussion) from the same course at a Portuguese speaking university. Thus, the study may not be entirely representative of the different interactions that can lead to different cognitive presence messages. Second, the dataset size

and unbalanced categories, although consistent with the findings in the literature, may affect the performance of the classifier.

Along the lines for further work, the authors plan to test the generalization of the classifier in another education context (i.e., blended vs. fully online vs. MOOC; and undergraduate vs. graduate) and the effectiveness of the proposed features to other languages (e.g., Spanish).

## References

1. Akyol, Z., Arbaugh, J.B., Cleveland-Innes, M., Garrison, D.R., Ice, P., Richardson, J.C., Swan, K.: A response to the review of the community of inquiry framework. *Int. J. E-Learn. Distance Educ.* **23**(2), 123–136 (2009)
2. Anderson, T., Rourke, L., Garrison, D.R., Archer, W.: Assessing teaching presence in a computer conferencing context. *J. Asynchronous Learn. Netw.* **5**, 1–17 (2001)
3. de Araújo, E.M., de Oliveira Neto, J.D.: Avaliação do pensamento crítico e da presença cognitiva em fórum de discussão online utilizando a análise estatística textual. In: *Proceedings of International Conference on Engineering and Computer Education*, vol. 8, pp. 113–117 (2013)
4. Arbaugh, J., Cleveland-Innes, M., Diaz, S.R., Garrison, D.R., Ice, P., Richardson, J.C., Swan, K.P.: Developing a community of inquiry instrument: testing a measure of the community of inquiry framework using a multi-institutional sample. *Internet High. Educ.* **11**(3–4), 133–136 (2008). <https://doi.org/10.1016/j.iheduc.2008.06.003>
5. Bauer, M.W.: Content analysis. An introduction to its methodology-by Klaus Krippendorff from words to numbers. Narrative, data and social science-by roberto franzosi. *Br. J. Sociol.* **58**(2), 329–331 (2007)
6. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
8. Corich, S., Hunt, K., Hunt, L.: Computerised content analysis for measuring critical thinking within discussion forums. *J. E-learn. Knowl. Soc.* **2**(1), 1–8 (2006)
9. Dowell, N.M., Skrypnik, O., Joksimovic, S., Graesser, A.C., Dawson, S., Gašević, D., Hennis, T.A., de Vries, P., Kovanovic, V.: Modeling learners' social centrality and performance through language and discourse. *International Educational Data Mining Society* (2015)
10. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.* **15**(1), 3133–3181 (2014)
11. Friedman, J., Hastie, T., Tibshirani, R.: *The Elements of Statistical Learning*. Springer Series in Statistics, vol. 1. Springer, New York (2001). <https://doi.org/10.1007/978-0-387-21606-5>
12. Gašević, D., Adesope, O., Joksimović, S., Kovanović, V.: Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *Internet High. Educ.* **24**, 53–65 (2015). <https://doi.org/10.1016/j.iheduc.2014.09.006>
13. Gašević, D., Kovanović, V., Joksimović, S.: Piecing the learning analytics puzzle: a consolidated model of a field of research and practice. *Learn. Res. Pract.* **3**(1), 63–78 (2017). <https://doi.org/10.1080/23735082.2017.1286142>



14. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. *Am. J. Distance Educ.* **15**(1), 7–23 (2001). <https://doi.org/10.1080/08923640109527071>
15. Garrison, D.R., Anderson, T., Archer, W.: The first decade of the community of inquiry framework: a retrospective. *Internet High. Educ.* **13**(1–2), 5–9 (2010)
16. Heo, H., Lim, K.Y., Kim, Y.: Exploratory study on the patterns of online interaction and knowledge co-construction in project-based learning. *Comput. Educ.* **55**(3), 1383–1392 (2010). <https://doi.org/10.1016/j.compedu.2010.06.012>
17. Hew, K.F., Cheung, W.S.: Attracting student participation in asynchronous online discussions: a case study of peer facilitation. *Comput. Educ.* **51**(3), 1111–1124 (2008)
18. Holsti, O.R.: *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley Pub. Co., Reading (1969)
19. Joksimovic, S., Gasevic, D., Kovanovic, V., Adesope, O., Hatala, M.: Psychological characteristics in cognitive presence of communities of inquiry: a linguistic analysis of online discussions. *Internet High. Educ.* **22**, 1–10 (2014)
20. Joksimović, S., Kovanović, V., Jovanović, J., Zouaq, A., Gašević, D., Hatala, M.: What do cMOOC participants talk about in social media?: a topic analysis of discourse in a cMOOC. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pp. 156–165. ACM (2015)
21. Kovanović, V., Gašević, D., Hatala, M.: Learning analytics for communities of inquiry. *J. Learn. Anal.* **1**(3), 195–198 (2014)
22. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M.: Automated cognitive presence detection in online discussion transcripts. In: *Proceedings of the Workshops at the LAK 2014 Conference Co-Located with 4th International Conference on Learning Analytics and Knowledge (LAK 2014)*, Indianapolis, IN (2014). <http://ceur-ws.org/Vol-1137/>
23. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M., Siemens, G.: Content analytics: the definition, scope, and an overview of published research. In: Lang, C., Siemens, G., Wise, A., Gašević, D. (eds.) *Handbook of Learning Analytics and Educational Data Mining*, pp. 77–92. SoLAR, Edmonton (2017). <https://doi.org/10.18608/hla17.007>
24. Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., Siemens, G.: Towards automated content analysis of discussion transcripts: a cognitive presence case. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK 2016)*, pp. 15–24. ACM, New York (2016)
25. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al.: *Caret: classification and regression training*. R package version 4 (2017)
26. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: *International Conference on Machine Learning*, pp. 957–966 (2015)
27. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977)
28. Liaw, A., Wiener, M., et al.: Classification and regression by random forest. *R News* **2**(3), 18–22 (2002)
29. Lipman, M.: *Thinking in Education*. Cambridge University Press, New York (1991)
30. McGill, T.J., Klobas, J.E.: A task technology fit view of learning management system impact. *Comput. Educ.* **52**(2), 496–508 (2009)
31. Mcklin, T.E.: *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. thesis, Atlanta, GA, USA (2004). aAI3190967

32. 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)
33. Park, C.L.: *Replicating the use of a cognitive presence measurement tool* (2009)
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: *Scikit-learn machine learning in Python*. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
35. Rosé, C., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: *Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning*. *Int. J. Comput. Support. Collab. Learn.* **3**(3), 237–271 (2008)
36. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: *Assessing social presence in asynchronous text-based computer conferencing*. *J. Distance Educ.* **14**(2), 50–71 (1999). <http://www.ijede.ca/index.php/jde/article/view/153>
37. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: *Methodological issues in the content analysis of computer conference transcripts*. *Int. J. Artif. Intell. Educ. (IJAIED)* **12**, 8–22 (2001)
38. Rozenfeld, C.C.D.F.: *Fóruns online na formação crítico-reflexiva de professores de línguas estrangeiras: uma representação do pensamento crítico em fases na/pela linguagem*. *Alfa Rev. Linguíst. (São José do Rio Preto)* **1**, 35–62 (2014)
39. Scarton, C., Gasperin, C., Aluisio, S.: *Revisiting the readability assessment of texts in Portuguese*. In: Kuri-Morales, A., Simari, G.R. (eds.) *IBERAMIA 2010. LNCS (LNAI)*, vol. 6433, pp. 306–315. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-16952-6\\_31](https://doi.org/10.1007/978-3-642-16952-6_31)
40. Stone, P.J., Dunphy, D.C., Smith, M.S.: *The general inquirer: a computer approach to content analysis* (1966)
41. Strijbos, J.W.: *Assessment of (computer-supported) collaborative learning*. *IEEE Trans. Learn. Technol.* **4**(1), 59–73 (2011)
42. Strijbos, J.W., Martens, R.L., Prins, F.J., Jochems, W.M.: *Content analysis: what are they talking about?* *Comput. Educ.* **46**(1), 29–48 (2006)
43. Tausczik, Y.R., Pennebaker, J.W.: *The psychological meaning of words: LIWC and computerized text analysis methods*. *J. Lang. Soc. Psychol.* **29**(1), 24–54 (2010)
44. Waters, Z., Kovanović, V., Kitto, K., Gašević, D.: *Structure matters: adoption of structured classification approach in the context of cognitive presence classification*. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) *AIRS 2015. LNCS*, vol. 9460, pp. 227–238. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-28940-3\\_18](https://doi.org/10.1007/978-3-319-28940-3_18)



# Fine-Grained Cognitive Assessment Based on Free-Form Input for Math Story Problems

Bastiaan Heeren<sup>1</sup>(✉), Johan Jeuring<sup>1,2</sup>, Sergey Sosnovsky<sup>2</sup>, Paul Drijvers<sup>3</sup>,  
Peter Boon<sup>3</sup>, Sietske Tacoma<sup>3</sup>, Jesse Koops<sup>4</sup>, Armin Weinberger<sup>5</sup>,  
Brigitte Grugeon-Allys<sup>6</sup>, Françoise Chenevotot-Quentin<sup>6</sup>, Jorn van Wijk<sup>1</sup>,  
and Ferdinand van Walree<sup>1</sup>

<sup>1</sup> Faculty of Management, Science and Technology, Open University  
of the Netherlands, P.O. Box 2960, 6401 DL Heerlen, The Netherlands  
[bastiaan.heeren@ou.nl](mailto:bastiaan.heeren@ou.nl)

<sup>2</sup> Department of Information and Computing Sciences, Universiteit Utrecht,  
Utrecht, The Netherlands

<sup>3</sup> Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

<sup>4</sup> Cito, Arnhem, The Netherlands

<sup>5</sup> Department of Educational Technology, Saarland University,  
Saarbrücken, Germany

<sup>6</sup> Laboratoire de Didactique André Revuz, Université Paris Est Créteil, Paris, France

**Abstract.** We describe an approach to using ICT for assessing mathematics achievement of pupils using learning environments for mathematics. In particular, we look at fine-grained cognitive assessment of free-form answers to math story problems, which requires determining the steps a pupil takes towards a solution, together with the high-level solution approach used by the pupil. We recognise steps and solution approaches in free-form answers and use this information to update a user model of mathematical competencies. We use the user model to find out for which student competencies we need more evidence of mastery, and determine which next problem to offer to a pupil. We describe the results of our fine-grained cognitive assessment on a large dataset for one problem, and report the results of two pilot studies in different European countries.

**Keywords:** Math story problems · Step-based assessment  
Free-form input · Solution strategies · User modelling

## 1 Introduction

Competence in mathematics has been identified at EU level as one of the key competencies for personal fulfilment, active citizenship, social inclusion, and employability in the knowledge society of the 21st century.<sup>1</sup> In 2009, concerns

<sup>1</sup> Mathematics Education in Europe: Common Challenges and National Policies, 2011.

about low student performance led to the adoption of an EU-wide benchmark in basic skills, which states that ‘by 2020 the share of 15-year-olds with insufficient abilities in reading, mathematics and science should be less than 15%’.<sup>2</sup> An extensive review of research evidence on ‘what works for children with mathematical difficulties’ has concluded that ‘interventions should ideally be targeted towards an individual child’s particular difficulties’ [4].

In a time where many European countries face shortages of teachers,<sup>3</sup> it is not to be expected that time of teachers available for assessment and determining competencies will increase. ICT can be of help here. The actions of a pupil when working in a digital environment can be collected in, and interpreted by, a so-called user model. The model describes the current level of mathematics achievement of a pupil. It is essential that this model not only analyses final answers, but also intermediate steps [13]. Intermediate steps contain essential information about how a pupil arrived at an answer (the method that was used), and can be used to identify misconceptions and pinpoint errors. We thus want to analyse intermediate steps to get precise diagnostic information.

This paper describes an approach to fine-grained cognitive assessment, including information about steps, of free-form input to math story problems on the domain of ‘Relationships’ targeting 15-year-olds. Figure 1 shows such a problem in Numworx, which we use as our digital assessment environment. We recognise steps and solution approaches in free-form input and use this information to update a user model of mathematical competencies. We use the user model to find out for which competencies we need more evidence of mastery, or proof of absence of mastery, and determine which next problem to offer to a pupil. We describe the results of our fine-grained cognitive assessment on a large dataset for one problem, and report the results of two pilot studies in different European countries. The contributions of this paper are:

- a novel approach to fine-grained cognitive assessment of free-form solutions to math story problems;
- a novel user modelling approach that uses the results from the fine-grained assessment;
- the results of applying our fine-grained assessment to a large data set for a single task, and for several smaller datasets containing solutions for multiple tasks.

This paper is organized as follows. In Sect. 2 we review several cognitive assessment methods, and motivate our focus on free-form input. Section 3 describes a high-level architecture for our approach. Section 4 gives the competencies we assess, and illustrates these with one particular task: the ‘Magical trick’. Section 5 illustrates the various components of our architecture with instances for this task, and Sect. 6 discusses the results of applying our components to a large dataset for the magical trick task, and the results of two small-scale pilots with multiple tasks. Section 7 concludes the paper and discusses future work.

<sup>2</sup> Strategic Framework for European Cooperation in Education and Training, ET 2020.

<sup>3</sup> Key Data on Education in Europe 2012.

## 2 Supporting Fine-Grained Cognitive Assessment

Conventional assessment tests developed in the traditions of psychometrics aim at supporting high-stakes decisions such as selection, placement, or licensing. In these circumstances, the focus of the test design is made on characteristics such as validity and reliability. The results of such assessment tests are usually unidimensional: a single value on a single scale [1]. While reliable ranking of test takers is important, a single score provides little information about the source of potential learning problems that have prevented a test taker from scoring high on the test. In this paper, the focus is made less on the absolute reliability of the test, and more on obtaining the detailed picture of a student's strengths and weaknesses. As we will explain later, such a fine-grained cognitive assessment needs to be organized on the basis of free-form answers students give to algebraic story problems, which adds another layer of complexity. While the inference of user knowledge based on the evidence produced by the result of solution analysis is rather straightforward, it is the analysis of the free-form student input that poses the biggest challenge for the cognitive assessment mechanism.

The screenshot shows a digital assessment interface for 'numw@rx'. The page title is 'Setting up algebraic expressions >'. Below the title, there is a 'LESSON' section with the same title. The main content area is titled 'Task 05 Magical trick?'. It contains a text block describing a math problem: 'A student says to her peer: "Choose a number, add 8, multiply the result by 3, subtract 4, add the initial number, divide by 4, add 2, and subtract the initial number. You will end up with 7."'. To the right of the text is an image of a black top hat with a red band and a black cane. Below the text is a question: 'Is this true for any starting number? Explain your answer.'. To the right of the text is a 'Your work' section with a text input area containing the following algebraic steps:  $(x+8) \cdot 3 - 4 + x / 4 + 2 - x = 7$ ,  $(3(x+8) - 4 + x) / 4 + 2 - x = 7$ ,  $(3x + 24 - 4 + x) / 4 + 2 = 7$ ,  $(4x + 20) / 4 + 2 - x = 7$ ,  $x + 5 + 2 - x = 7$ , and  $7 = 7$ . Below the input area is a 'Submit' button. At the bottom of the page, there is a progress bar with 10 steps, where step 05 is highlighted.

**Fig. 1.** The ‘Magical trick’ math story problem in a digital assessment environment

A direct way to facilitate fine-grained cognitive assessment at the solution analysis phase is to use plain assessment exercises that solicit easily verifiable input from students, such as multiple-choice questions. In this case, the potential uncertainty at the analysis phase is minimal: a student answer is either correct or

not, and it is clear which concept is responsible for the outcome. One example of building an adaptive cognitive assessment based on such exercises is SIETTE [3].

However, it is often the case that the underlying assessment method requires a more advanced type of exercises that engages students in complex tasks and requires application of multiple concepts over several solution steps. One way to address this problem is to keep the solution analysis phase simple by allowing students to do all the intermediate computations outside the system, and requiring only the final answer. The uncertainty is then transferred to the part of the system intelligence that defines the concepts responsible for student mistakes. Often, the ‘blame’ is simply shared among the related concepts. For example, a single self-assessment exercise may involve more than a dozen of concepts [11].

A more reliably way to detect the concept(s) responsible for a mistake made in multi-step exercises is to structure the interaction between the student and the system. In this case, a student is restricted by the interface in what she can enter, and the source of a mistake is easier to identify. Narciss et al. [9] provide one example of such an approach based on a dedicated interface element, where a student chooses the type of the operation before performing it. The Andes system structures the entire student solution so that at every step the operation, its operands, and the purpose of the operation is known to the system [2].

Finally, a system can ask additional follow-up questions that address intermediate steps of the student solution, thus identifying the source of a possible mistake. ASSISTment is one of the systems that employ this approach [5].

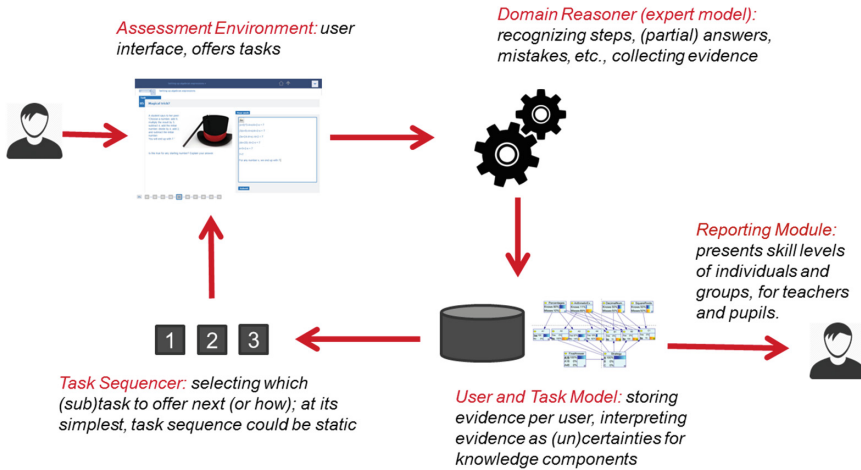
Unfortunately, neither of the listed methods is fully suitable. ‘Sharing the blame’ between potentially responsible concepts does not necessarily guarantee modelling accuracy. Structuring and restricting interaction inescapably provides scaffolding to a student, which is fine for a tutoring system, but is less desirable in an assessment scenario. Finally, follow-up questions unnecessarily extend the assessment session, and can also potentially reveal knowledge to test takers. Our research attempts to go beyond the state of the art by employing a range of AI techniques for analysing free-form student input to math story problems in an assessment setting.

### 3 A High-Level Architecture

Our goal is to support fine-grained cognitive assessment based on free-form input for math story problems. For this purpose, we analyse free-form student input to find out which steps a student takes when solving a task, and we use the information we find to update the user model, and to determine which next task to offer to the student. This section introduces the main components of the pipeline we use for assessment.

Tutoring systems consist of an outer loop and an inner loop [12]. The inner loop supports a pupil solving a task by taking steps to reach a solution. In the setting of assessing competencies in mathematics, the inner loop is responsible for analysing the steps that a pupil takes, and recognising the high-level solution approach used by the pupil. Giving hints and feedback, thereby providing

opportunities for learning, is not part of an assessment system. The outer loop selects the next task to offer to a pupil. This selection is based on information that is stored in a user model. In the case of assessment, the goal is to quickly find out competencies and misconceptions from a pupil by collecting evidence of mastery, or absence of mastery. Figure 2 identifies the main components needed for assessment, and the flow of information between these components. Next, we describe each of the components.



**Fig. 2.** Information flow in the outer loop

The *assessment environment* is the user interface responsible for offering the tasks to pupils. Such an environment manages user accounts for teachers and pupils, and may offer a teacher area with access to student and class work. For tasks in mathematics, some special tools are needed, for example a formula editor, a calculator, and a graphing tool.

The *domain reasoner* component (also known as the expert knowledge module [10]) contains expert knowledge in the task domain (i.e., the domain of relationships) and uses this knowledge for reasoning about the steps that together form the answer. Input is analysed at different levels of granularity, in particular at the fine-grained step level at which the use of variables, calculational mistakes, sloppy notation, and precedence errors can be detected (among other types of steps), and the high-level solution approach, which may be algebraic, numerical, only partially correct, etc. The parts that are recognised are translated to evidence for the competencies in which we are interested.

The *user and task models* are responsible for the inference, storage and update of student knowledge based on the evidence collected by the domain reasoner. To help manage the potential uncertainty of the diagnosis produced by the domain reasoners, these models are represented as Bayesian Networks (BNs). The user model structures concepts and competencies into a hierarchy,

and the task models relate concepts and competencies involved in a particular assessment task to its solutions steps and the characteristics of the final answer.

The *task sequencer* determines which task should be used next in the assessment. For this, information from the user model is used to calculate which task can best be used to remove uncertainties (in the user model) about competencies. At its simplest, a task sequencer offers tasks in a predetermined order.

The *reporting module* presents the skill levels of individuals in an easy-to-understand way (such as a skill-meter), targeting both teachers and pupils. The module can aggregate information and display information about groups.

A learning environment is a highly complex educational software applications [8]. The technology for calculating diagnostics and building a detailed user model is offered as an open set of services that can be used by multiple learning systems. A service-oriented approach promotes large-scale reuse and counteracts the complexity found in educational software applications [7]. The service-oriented approach, and the integration of these services into existing systems, is one of the innovative aspects of our approach.

## 4 Tasks and Competencies

Our assessment is based on the answers of a pupil on a set of tasks. The tasks concern the domain of Relationships and target 12–15 years old pupils. In particular, the tasks involve setting up algebraic expressions, equations, and inequalities, as well as simplifying and solving them. The tasks require multi-step solutions, and usually there are multiple ways to solve these tasks. The assessment presented here consists of ten tasks in the domain of Relationships. In this section we introduce the tasks, the competencies they address, and the ways they can be solved by means of an example.

We want to assess the learning goals R1–R3 in the domain of Relationships listed in Table 1. The diagnostic assessment analyses each pupil’s answer on the three dimensions described above, but not only by means of correct/incorrect. The diagnostic system provides a set of codes that characterize the answer according to an a priori analysis.

Next, for each of the tasks, the possible solution steps and alternative routes are described, as well as the mistakes students might make. Consider, for example, the ‘Magical trick’ task, which is also shown in Fig. 1:

A student says to her peer: ‘Choose a number, add 8, multiply the result by 3, subtract 4, add the initial number, divide by 4, add 2, and subtract the initial number. You will end up with 7.’

Is this true for any starting number? Explain your answer.

The magical trick task is a rich task from a diagnostic point of view. Its goal is to identify whether or not a student is able to generalize and prove a property (R1 and R3) with algebraic strategies. It also provides information about the types of connections (R2) between two representations (in this case a numerical and an algebraic representation), and about the arguments used by a student.



**Table 1.** Codes for characterizing answers in the domain of Relationships [6]

- R1: Construct algebraic objects, e.g., set up expressions, formulas, equations
- Correct global (R11), or step-by-step (R12) algebraization
  - Incorrect algebraization (R13)
  - Numerical solution (R14), either global (R141) or step-by-step (R142)
- R2: Recognise and relate different representations of a mathematical relationship
- Correct relation between two representations with congruent rules (R21)
  - Incorrect relation between two representations without reformulation (when non-congruent) (R23)
  - Incorrect relation between two representations with schematization (R24)
- R3: Calculate and simplify, e.g., expand/factor algebraic expressions, solve equations
- Correct calculation with argumentation and correct semantic and syntactic rules (R31), or without argumentation (R32)
  - Incorrect with false rules (R33): errors of parentheses (R332), errors of signs (R333)
  - Incorrect with operator priority or concatenation (R34)

There are (at least) two strategies to solve this task: an arithmetic strategy using a particular number (Table 2), and an algebraic strategy that involves a variable (Table 3). For both strategies, we distinguish between a global approach (first set up the complete expression, then perform the simplifications) and a step-by-step approach (set up the expression for the next step and simplify). The tables also present some mistakes that illustrate incorrect techniques.

## 5 Components

This section discusses two components for assessing solutions in more detail.

### 5.1 Domain Reasoner

The domain reasoner receives free-form input, determines which steps are taken, and tries to recognise the solution strategy. Based on this analysis, competencies and misconceptions are determined. The analysis proceeds in three phases:

1. Extract mathematical expressions from the textual input, ignoring most of the natural language;
2. Parse the extracted expressions and equations into (structured) mathematical objects;
3. Recognise the solution strategy by parsing the sequence of mathematical objects (i.e., approach strategy recognition as a parsing problem).

Numworx has a formula editor for entering mathematical content. Pupils are allowed to use this editor, which is particularly useful for entering square roots, powers, and other operations that do not have an obvious textual equivalent. Each phase will be described in more detail below.

*Math Extraction (Phase 1)*. A pupil enters input as free text and may use a formula editor that contains mathematical symbols. Learning environments have precise information of what a pupil wrote. We use the MathML standard for transferring the semantics of these graphical mathematical representations from the learning environment to the services.

**Table 2.** Arithmetic strategy (for number 5) and possible mistakes

Solution	Reasoning	Coding
$((5+8)*3-4+5)/4+2-5 = 7$	Correct arithmetic strategy with global expression, but with a missing generalization for any starting number	R141, R21, R31
$5+8 = 13; 13*3 = 39;$ $39-4 = 35; 35+5 = 40;$ $40/4 = 10; 10+2 = 12;$ $12-5 = 7$	Correct arithmetic strategy with step-by-step approach, but with a missing generalization for any starting number	R142, R22, R31
$5+8*3-4+5/4+2-5 = 7$	Erroneous arithmetic strategy with global expression without parentheses	R14, R23, R33
$5+8 = 13*3 = 39-4 = 35+5 = 40/4 = 10+2 = 12-5 = 7$	Erroneous arithmetic strategy with step-by-step calculations	R14, R24

Since the input from a pupil may contain malformed mathematical expressions, it is not always possible to represent the pupil's input in a standard mathematical representation such as MathML Content. Thus, we allow a learning environment to send input following a subset of the MathML Presentation language. Lexical analysis converts MathML code back to plain text. In this way, a pupil can use specialized symbols such as the root symbol in expressions, while we handle everything as plain text during the recognition phase.

For each task, parts of the math extraction phase can be specialized. For example, some tasks suggest the use of multi-letter variables such as `dist` and `cost`: these variables can be whitelisted. In other tasks there is no distinction between uppercase and lowercase characters, hence we allow both. Furthermore, depending on the language specified in the request, some pre-processing is performed. For example, German words such as `mal`, `plus`, `quadrat`, and `hoch` are converted to their mathematical representations (`*`, `+`, `^2`, and `^`, respectively). Currently, we support English, German, French, and Dutch.

Certain symbols have multiple interpretations. For instance, compare `x+3` (`x` is a variable) with `3x5=15`, in which `x` is probably used to denote multiplication. Similar ambiguities arise for certain punctuation symbols. We are careful not to blow up the search space by considering all interpretations, but instead use heuristics to resolve most of the ambiguous interpretations. For example, any use of `x` surrounded by numbers is interpreted as multiplication. This approach seems to work quite well, but does not prevent misinterpreting expressions such as `1/2 x a x b`. Tasks for which variable `x` has no obvious meaning can be configured to always interpret it as multiplication.

*Parsing Expressions (Phase 2).* Parsing expressions is rather straightforward, although we do have to take care of equations that are incorrectly chained, such as  $5+8 = 13*3 = 39$  (see the possible mistakes in Table 2), which we split into two equations with an annotation for the incorrect chaining. Parsed equations are particularly helpful, because they can be checked for equality to spot mistakes (e.g.,  $5+8 = 40$ ).

**Table 3.** Algebraic strategy and possible mistakes

Solution	Reasoning	Coding
$(x+8)*3-4+x)/4+2-x$ $= (3x+24-4+x)/4+2-x$ $= (4x+20)/4+2-x$ $= x+5+2-x$ $= 7$	Algebraic proof with global expression	R11, R21, R31
$(x+8)*3 = 3x+24$ $3x+24-4 = 3x+20$ $3x+20+x = 4x+20$ $(4x+20)/4 = x+5$ $x+5+2 = x+7$ $x+7-x = 7$	Algebraic proof with step-by-step approach	R12, R31
$(x+8)*3-4+x/4+2-x$ or $(x+8*3-4+x)/4+2-x$ or $x+8*3-4+x/4+2-x$	Order or priority of operations is missed	R13, R23
$x+8*3-4+x/4+2-x$ $= 2x+20/4+2-x$ $= 2x+5+2-x$ $= x+7$	Simplification mistakes, such as ignoring parentheses or priority rules	R13, R23, R33
$(x+8)*3 = 3x+24 = 27x$ $27x-4+x = 24x$ $24x+2-x = 23x+2 = 25x$	Simplification mistakes, such as dilemma process-product: $a+b \rightarrow ab$	R13, R21, R341

*Strategy Recognition (Phase 3).* In the third phase we try to recognise solution strategies. The general approach is to consider recognition as a parsing problem, and to express the solution strategies as context-free grammars. During the parsing, we keep track of variables that are introduced, numbers that are chosen, and definitions that can be propagated (e.g.  $d = a^2+14$ , followed by  $4 * d$ ).

The recogniser must be flexible enough to recover from different types of mistakes and imperfections by providing some error correction, for example steps that are taken implicitly, (basic) calculation mistakes, algebraic misconceptions, missing parentheses or incorrect options when modelling, and so on. See Tables 2 and 3 for more mistakes that are recognised during this phase. Note that there is a trade-off between flexibility and computation time: we use real data, collected from earlier user studies and pilots, to calibrate the recogniser.

### 5.2 Task Model and User Model

Not every student answer provides high-quality input for the domain reasoner to analyse. Hence, the evidence produced by the analysis step can be scattered. It is often clear whether or not a student has given a correct final answer and which strategy a student has applied. Sometimes, individual steps are clearly indicated in the solution and can be easily diagnosed. However, as a rule, we have assumed (and the following evaluation confirms this) that the diagnosis evidence is not guaranteed for any step of the solution, nor for the overall answer. Yet, the implemented solution has to produce a detailed cognitive assessment under such uncertainty. Due to these consideration, we have decided to employ BNs as a well-known mechanism to support inference under uncertainty.

BNs are widely used for modelling students' knowledge based on students' responses to multi-step learning exercises [1, 2]. We follow this tradition by representing both the model of student knowledge and the task models as probabilistic networks. However, the fact that users in the assessment environment produce unrestricted, unstructured, unscaffolded input, adds another layer of uncertainty into our inference pipeline and reflects on the design of the task modelling BNs.

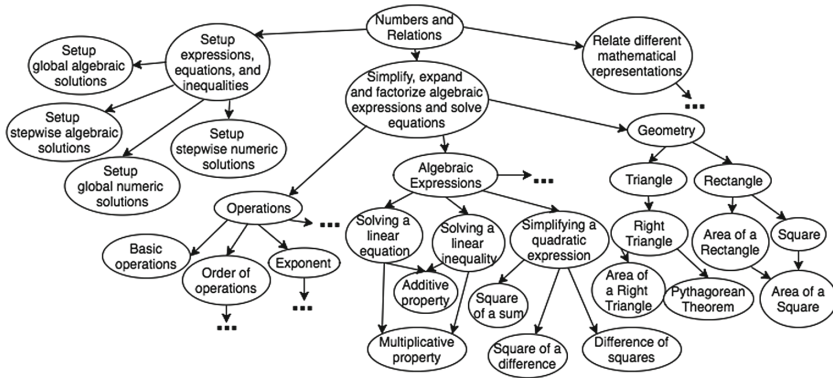


Fig. 3. Part of the user model

Figure 3 visualizes the upper-level structure of a first version of the user model. At the top of the hierarchy, we have three learning goals (R1–R3, see Table reftable:codes). They are further categorized into smaller competencies and concepts that participate in individual models of tasks. For example, Fig. 4 represents a tentative BN for the ‘Magical trick’ exercise. The top nodes represent competencies corresponding to setting up a solution (R1): these nodes connect the task model with the user model. The top nodes in Fig. 4 are connected to more general characteristics of the solution (correctness and properties of the chosen strategy). These characteristics are more likely to be diagnosed. At the same time, every combination of these characteristics is probabilistically related to a certain sequence of solutions steps (there are four possible strategies to

solve the ‘Magical trick’ exercise, corresponding to four such sequences of steps). Within each sequence, the probability of a next step to be applied correctly depends on the previous step and the corresponding concept nodes that model the probability that a student has mastered these concepts. When a student starts working on a task, the prior probabilities for all concept and competency nodes are copied from this student’s user model. Once a student submits a solution, some of the nodes are set to 1 or 0 in the task model, depending on the results of the analysis phase. This triggers the probability update of the concept and competency nodes given the new evidence. Finally, the updated probabilities are carried over to the user model; they will inform the task model of the next assessment task the student will attempt.

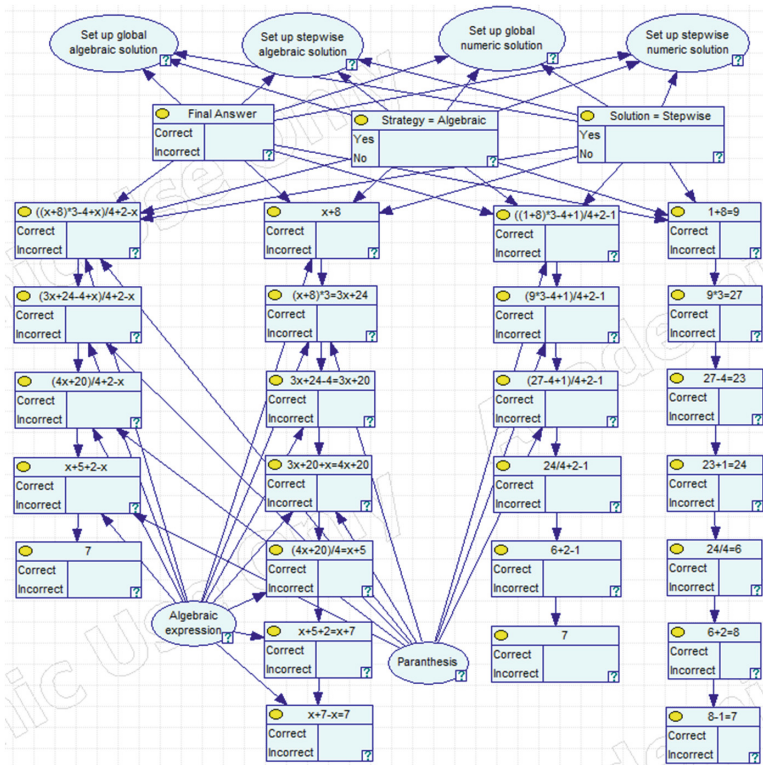


Fig. 4. Magical trick task model

## 6 Evaluation

We have evaluated our assessment technology in two ways. First, we tested the domain reasoner on a large collection of free-form answers for the Magical trick task, which we describe in Sect. 6.1. Second, we conducted two small-scale pilot studies in 2018, and report about these pilots in Sect. 6.2.

## 6.1 Magical Trick Dataset

Between 2011 and 2015, several experiments were conducted on the LaboMep platform, which was developed by Sésamath, a French maths' teachers association.<sup>4</sup> The goal of these experiments was to study the integration of automated diagnosis tools in the usual teaching practices, and to study the evolution of students' cognitive profiles in algebra during several years [6]. This has resulted in a large collection of student responses (grade 8–10). From the dataset, we analysed 2956 free-form answers for the Magical trick task with the domain reasoner. Running the analysis took 155 s, which is fast enough for online assessment. From the 2956 answers, we extracted 18,302 mathematical expressions: 99.2% of these expressions could be parsed. Most of the parse errors are caused by unbalanced parentheses, for example  $[(x+8)*3-4+x])/4+2-x = 7$ .

The results of recognising the solution strategy are as follows: 115 answers (3.89%) only contain natural language, and no mathematical content; 677 answers (22.90%) follow the algebraic strategy; 1527 answers (51.66%) follow the arithmetic strategy; 637 answers (21.55%) could not be recognised.

We also found combinations of solution strategies, e.g. the arithmetic strategy with different numbers, or the algebraic strategy followed by the arithmetic strategy. Some analysis results were checked ad hoc, but since we do not have a golden standard to compare against, we cannot rule out false positives or true negatives. Nevertheless, the results indicate that a substantial part of the free-form input could be analysed automatically, including the assessment of high-level learning goals based on the solution strategy that was followed.

## 6.2 Analysis of Pilots

In the context of the Erasmus+ Advise-Me project, we organized two small-scale pilots to test our assessment method and the free-form input for math story problems. The first pilot ( $N = 19$ ) was organized on March 15, 2018, in Germany, and the second pilot ( $N = 22$ ) was organized on April 9, 2018, in the Netherlands. Pupils were asked to solve ten math story problems, and to answer eight statements about the tasks, the software, and their attitude towards mathematics in a questionnaire. The experiments were carried out in 90 min: 10 min for logging-in, briefing, and the first questionnaire, 70 min for doing the tasks, and 10 min for the final survey. Some tasks have multiple parts (a–c). All interactions with the assessment environment were logged for further analysis.

From the questionnaires, we learned that pupils found the tasks clear and that they think they did well in the test. Doing well typically improves the experience. They found it relatively easy to use the assessment software and the text editing field. They do not often do math tasks on the computer.

Table 4 summarizes how often the algebraic or arithmetic solution strategy was recognised for the pilot studies. The 'graphical' strategy corresponds to

<sup>4</sup> <http://www.labomep.net>.

**Table 4.** Recognising solution strategies in pilots (Al = Algebraic, Ar = Arithmetic, Gr = Graphical, all as proportions); we also report the proportion of empty answers (Em) and unrecognised answers (Un).

Task		German pilot					Dutch pilot					
		N	Al	Ar	Gr	Em	Un	N	Al	Ar	Em	Un
1	Making a square	19	.53			.05	.42	22	.55			.45
2	Matryoshka	17		.76		.06	.18	12		.17	.33	.50
3	Car rental	18	.39		.39	.06	.17	22	.82			.18
4	Pattern	18	.11	.67			.22	19	.11	.53		.37
5	Magical trick	18	.06	.11		.56	.28	20			.70	.30
6a	Rectangle area	18	.94			.06		22	.95			.05
6b	Rectangle area	18	.67			.06	.28	22	.77			.23
6c	Rectangle area	18				.67	.33	22	.36		.55	.09
7b	Theatre rate	18	.33		.11	.50	.06	21	.76		.05	.19
9a	Area of triangle	15	.33			.47	.20	22	.77		.05	.18
9b	Area of triangle	15	.20			.60	.20	22	.73		.05	.23
9c	Area of triangle	15	.13			.73	.13	22	.77		.23	
10.	V-pattern	15	.53			.27	.20	22	.73		.05	.23
	<i>Overall</i>	222	.33	.12	.04	.30	.21	270	.59	.04	.14	.22

approximating the solution directly from a graph. Task 7a (Theatre rate) and task 8 (Area and expression) are omitted: the former requires an answer in natural language, and for the latter, pupils have to click areas instead of writing mathematical expressions. Overall, the solution strategy could be recognised for nearly 80% of the answers. For the remaining 20%, recognised steps and the final answer can still provide valuable information. A closer inspection of the unrecognised answers resulted in the following categorization of difficulties in recognising answers: unclear or ambiguous notation, use of natural language, and unanticipated errors.

## 7 Conclusion and Future Work

We have developed a framework for fine-grained cognitive assessment of free-form solutions to math story problems for pupils of around 15 years old. The framework uses a domain reasoner to analyse the input from pupils. The domain reasoner extracts the mathematics from the free-form input, parses the mathematical expressions, and then tries to recognise a solution strategy in the solution. The diagnosis from the domain reasoner is taken as input by a Bayesian task model, which in its turn is used to populate a user model.

We have evaluated our framework in various ways. We have tested that one of our domain reasoners for a particular task can analyse more than 80% of pupil

solutions in a dataset of almost 3000 solutions. In a number of small pilots we determined the main causes for our domain reasoner to fail to recognise pupil solutions. Some of the main causes are pupils using only natural language, or mixing up notation. In a qualitative evaluation, pupils were mildly positive about solving this kind of tasks in an online assessment system.

In the future, we want to analyse the quality of the user models resulting from our analyses. We want to determine ways to deal with situations in which we do not recognise a solution from a pupil. Here we envisage several approaches: we might ask a pupil simpler questions, or we might combine our analysis with natural language processing software for recognising mathematical language. Furthermore, we want to perform more extensive evaluations with our framework and compare the fine-grained cognitive assessments with a manual analysis produced by experts.

**Acknowledgements.** The Advise-Me project has received funding from the European Union's ERASMUS+ Programme, Strategic Partnerships for school education for the development of innovation, under grant agreement number 2016-1-NL01-KA201-023022. For more information, visit <http://advise-me.ou.nl>.

## References

1. Almond, R.G., Mislevy, R.J., Steinberg, L.S., Yan, D., Williamson, D.M.: Bayesian Networks in Educational Assessment. Springer, New York (2015). <https://doi.org/10.1007/978-1-4939-2125-6>
2. Conati, C., Gertner, A., VanLehn, K.: Using Bayesian networks to manage uncertainty in student modeling. *User Model. User Adapt. Interact.* **12**(4), 371–417 (2002)
3. Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J.L., Ríos, A.: SIETTE: a web-based tool for adaptive testing. *Int. J. Artif. Intell. Educ.* **14**(1), 29–61 (2004)
4. Dowker, A.: *Children with Difficulties in Mathematics: What Works?*. DfES Publications, London (2004)
5. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. *User Model. User Adapt. Interact.* **19**(3), 243–266 (2009)
6. Grugeon-Allys, B., Chenevotot-Quentin, F., Pilet, J., Prévot, D.: Online automated assessment and student learning: the *PEPITE* project in elementary algebra. In: Ball, L., Drijvers, P., Ladel, S., Siller, H.-S., Tabach, M., Vale, C. (eds.) *Uses of Technology in Primary and Secondary Mathematics Education. ICME-13*, pp. 245–266. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-76575-4\\_13](https://doi.org/10.1007/978-3-319-76575-4_13)
7. Heeren, B., Jeuring, J.: Feedback services for stepwise exercises. *Sci. Comput. Program.* **88**, 110–129 (2014)
8. Murray, T.: An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: Murray, T., Blessing, S.B., Ainsworth, S. (eds.) *Authoring Tools for Advanced Technology Learning Environments*, pp. 491–544. Springer, Dordrecht (2003). [https://doi.org/10.1007/978-94-017-0819-7\\_17](https://doi.org/10.1007/978-94-017-0819-7_17)
9. Narciss, S., et al.: Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Comput. Educ.* **71**, 56–76 (2014)



10. Nwana, H.: Intelligent tutoring systems: an overview. *AI Rev.* **4**(4), 251–277 (1990)
11. Sosnovsky, S., Brusilovsky, P., Lee, D.H., Zadorozhny, V., Zhou, X.: Re-assessing the value of adaptive navigation support in e-Learning context. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) *AH 2008. LNCS*, vol. 5149, pp. 193–203. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-70987-9\\_22](https://doi.org/10.1007/978-3-540-70987-9_22)
12. VanLehn, K.: The behavior of tutoring systems. *J. AIED* **16**(3), 227–265 (2006)
13. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**(4), 197–221 (2011)



# Extending the SIPS-Model: A Research Framework for Online Collaborative Learning

Karel Kreijns<sup>1(✉)</sup> and Paul A. Kirschner<sup>1,2</sup>

<sup>1</sup> Open Universiteit Nederland, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands  
{karel.kreijns,paul.kirschner}@ou.nl

<sup>2</sup> University of Oulu, Pentti Kaiteran Katu 1, 90570 Oulu, Finland

**Abstract.** The SIPS-model, introduced to emphasize social aspects of online collaborative learning (OCL) expresses the degree to which online environments for collaborative learning support social aspects through social affordances by the sociability attribute. However, OCL-environments are primarily meant to support collaborative learning. Hence, SIPS was extended by adding an educability attribute to express the degree to which these environments have educational affordances for collaborative learning (CL). In this paper, we propose a second extension, adding hedonicity to express the extent to which OCL-environments give pleasure and enjoyment during the interacting with them. By adding hedonicity, we stress that learning should not only be effective and efficient but also enjoyable. That aspect, though missing in SIPS, is an important element in learning. To reduce complexity of the SIPS-model caused by the two extensions, SIPS is split into three distinct sub-models: the PIP-, SIP-, and HES-model. By characterizing OCL-environments by the attributes hedonicity, educability, and sociability, we can more accurately evaluate the impact of OCL-environments on social presence, participation, social interaction, and social space which are needed for socio-cognitive (where group learning/knowledge construction takes place) and socio-emotional processes (where group forming/dynamics takes place) in groups. The TEL-community should take up the non-trivial task of designing OLC-environments that possess hedonicity, educability, and sociability through their respective affordances.

**Keywords:** Online collaborative learning · Hedonicity · Educability · Sociability  
Social presence · Social space · Affordances · Extended SIPS-model · CSCL

## 1 Introduction

Collaborative learning is “the instructional use of small groups so that students work together to maximize their own and each other’s learning” [20; p. 87]. A variety of pedagogical techniques was developed to implement collaborative learning (CL) such as structured academic controversy [19], and jigsaw [3]. In contrast to these so called direct approaches, Johnson and Johnson [18] suggested a conceptual approach, which entails that every successful collaborative pedagogical technique should fulfill five conditions: (1) positive interdependence, (2) group and individual accountability, (3) promotive interaction, (4) group processing, and (5) social skills. CL was first applied

in face-to-face classrooms but as technology developed and internet became the dominant way to connect computers, computer supported classroom collaborative learning (CCL) and online collaborative learning (OCL)—collectively known as computer supported collaborative learning (CSCL)—became possible. Computer-supported CCL is basically synchronous collaboration whereas OCL supports a-synchronous CL. While a-synchronous collaboration has certain benefits such as relaxation of time and place constraints enabling collaboration between distance education students, it has also drawbacks [30]. First, social interaction for socio-cognitive processes risks not occurring unless specific pedagogical techniques are developed that takes the asynchronous mode of OCL into account. Second, while group dynamics processes naturally take place in face-to-face settings, they are hampered in online settings unless explicit attention is paid to them by recognizing that social interaction is not only necessary for socio-cognitive processes but also for the socio-emotional processes underlying group forming and group dynamics. It is hampered because the social interaction has to take place via communication media which are mostly text-based, which cannot easily communicate the expressiveness and richness—in terms of verbal and non-verbal cues—of face-to-face social interaction. These cues are needed for impression formation which is at the basis for developing the interpersonal relationships so important in group dynamics [54].

Group forming and group dynamics and all the variables that may affect these processes are all social aspects of OCL. Kreijns, Kirschner, and Vermeulen [29] proposed the SIPS-model (SIPS: Sociability, social Interaction, social Presence, social Space; see also [57]) to emphasize the social aspects of OCL. In the SIPS model, the degree to which online environments for CL support social aspects through social affordances is expressed by their sociability attribute. But as the purpose of OCL-environments is to support CL, Kirschner, Kreijns, Phielix, and Franssen [25] extended the SIPS model, adding an educability attribute expressing the degree to which these environments have educational affordances to support collaborative learning. In this paper, we propose a second extension, namely the hedonicity attribute which expresses the extent to which OCL-environments give pleasure and enjoyment during the interaction with them. By adding hedonicity, we stress that learning should not only be effective and efficient but also enjoyable. That last aspect was missing in SIPS but considered an important element in learning [23]. Not considering hedonicity in OCL would mean an incomplete picture of all the variables that may affect social interaction and, thus, CL, group forming and group dynamics.

To reduce complexity of the SIPS-model caused by the two extensions, the model is split into three distinct sub-models: the PIP-model (PIP: Participation, social Interaction, Performance), the SIP-model (SIP: Social Information Processing) based on Walther's SIP-theory [54, 55] and the HES-model (HES: Hedonicity, Educability, Sociability). In the next sections each of the sub-models (PIP, SIP, and HES) will be described.

## 2 The Extended SIPS-Model

### 2.1 The PIP Model: Participation, Social Interaction, Performance

The PIP-model (*Participation, social Interaction, Performance*), introduced by Kreijns, Kirschner, and Jochems [30, 31], is meant to show the dual function of social interaction, namely for the meta-cognitive and socio-cognitive processes and for the social and social-emotional processes, and how these processes affect learning and social performances. Meta-cognitive and socio-cognitive processes are those processes in which the group learning and knowledge co-construction takes place and are seen as being important for regulating CL in groups.

Figure 1 displays the PIP-model along with a number of variables that affect participation and social interaction, and some outcome variables. The next sub-sections will discuss pedagogical techniques, academic and social skills, the dispositions OCL-group members may have, and finally social space and social presence.

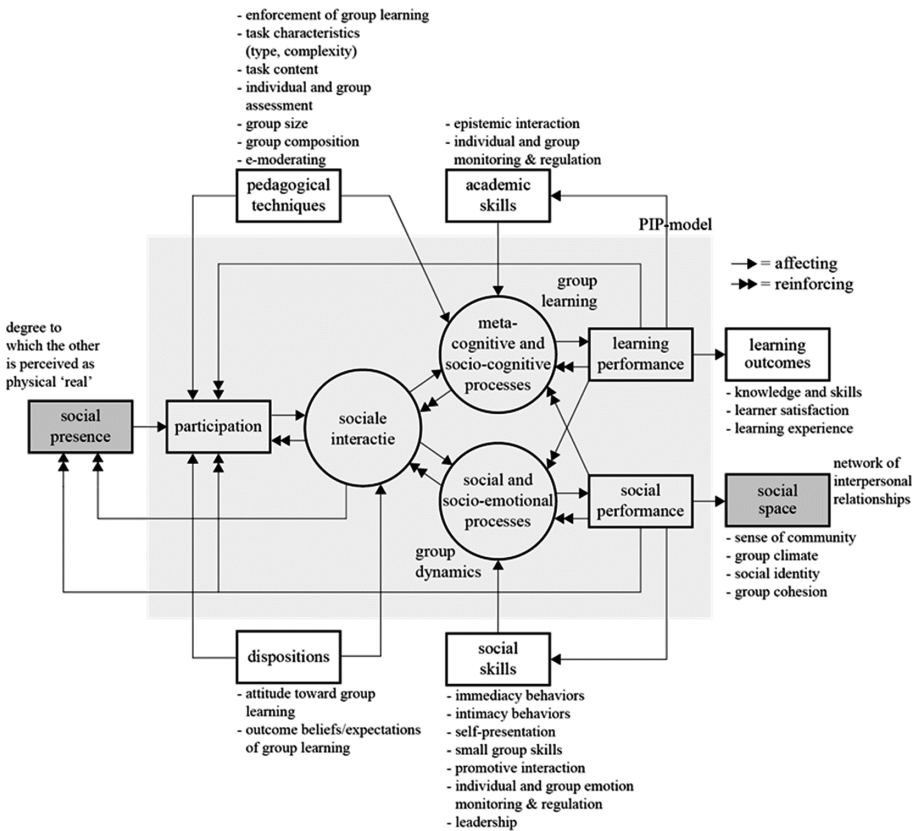


Fig. 1. The PIP-model applied to collaborative/group learning.

**Pedagogical Techniques.** Researchers have developed pedagogies specifically suited to CSCL. One stream exploited the graphical possibilities of computer displays by introducing shared graphical workspaces. Knowledge Forum® is a knowledge building environment in which shared discourse is supported by the textual and graphical representations of ideas that can be reorganized or reconstructed [46]. A second stream investigated the effectivity of scripting on the degree to which productive social and cognitive interactions emerged between members of a CL-group by showing prompts/cues on the computer screens to which they have to respond [10, 58]. Through scripting, CL-members are more engaged in problem solving, fostering mutual understanding, and giving elaborated explanations than when there is no script guidance. With scripting the probability of learners sharing knowledge construction is increased; without scripting learners risk diverging from the topic [58]. Recently a third stream of CSCL-researchers are augmenting cognitive load theory [52] so that it can be applied in groups as well. They stated that when group task complexity exceeds the complexity level that an individual can process alone, the task should be divided among more individuals working together but under the condition that transactional costs—because of communication and coordination—is kept acceptable [24]. However, these pedagogical techniques are primarily for synchronous computer-supported CCL and may not all be well suited for a-synchronous OCL.

**Academic Skills.** Academic skills refer to the “ability to identify and use different ways of knowing, to understand their different forms of expression and evaluation and to take the perspectives of others who are operating within a different epistemic framework” [39; p. 109]. Ohlsson [42] proposed seven epistemic activities associated with academic skills: (1) describing, (2) explaining, (3) predicting, (4) arguing, (5) critiquing/evaluating, (6) explicating, and (7) defining. Some researchers point to the ability to perform these epistemic activities as argumentation competence that can be supported by argument scaffolds, a specific kind of scripting [59]. By performing epistemic activities, CL-group members acquire domain-specific knowledge.

**Social Skills.** In addition to academic skills, social skills are also necessary and complement academic skills. Johnson and Johnson [18; p. 369] included small group skills in their five conditions because “participants must (a) get to know and trust each other, (b) communicate accurately and unambiguously, (c) accept and support each other, and (d) resolve conflicts constructively [...]. Interpersonal and small-group skills form the basic nexus among individuals, and if individuals are to work together productively and cope with the stresses and strains of doing so, they must have a modicum of these skills.” Except for these skills, social skills also encompass many other skills including leadership and self-presentation in an online environment.

**Dispositions.** Dispositions like attitude and beliefs towards CL must be taken into account because they affect participation and social interaction in both the educational and social dimensions. The OECD Programme for International Student Assessment (PISA) 2015 [41] found females to be more positive than males about CCL when assessed on its relational potential (i.e., working with peers) whereas the opposite was true when CCL was assessed on its potential for efficient teamwork (e.g., make better decisions). A study by Kreijns [27; Chapter 10] showed that the majority of distance education students

involved in OCL had negative attitudes towards CL. Distance education students are often adults with families and full-time work and therefore, the freedom to study whenever they wish, in their own pace, and from any location made them decide to enroll in distance courses. CL jeopardizes freedom of pace and forces them to coordinate their activities with each other. Indeed, Rourke and Anderson [44; p. 270] pointed out that there is a “group of students [that] may select distance education because it has traditionally allowed students to work towards their goals independently without having to interact with others.”

**Social Space.** Effective CL can only take place when a group is productive and well-functioning with a positive group climate, mutual trust, a sense of belonging and of community making the group a psychologically safe place to engage in critical discourse and share knowledge [18, 49, 53]. These features are manifestations of a sound social space; the network of social relationships amongst group members [29]. As Jacques [17; p. 72] stated “lack of attention to the socio-emotional dimension means that many of the task aims cannot be achieved. Without a climate of trust and cooperation, students will not feel taking the risk of making mistakes and learning from them.” Kreijns, Kirschner, and Jochems developed a social space measure [34].

**Social Presence.** Whether social interaction is used for socio-cognitive or for socio-emotional processes, it is affected by the communication media’s limited capacity to communicate verbal and non-verbal cues. To build a theory around these media effects and how they affect participation and social interaction, OCL-researchers (e.g., [13, 14, 61]) adopted the concept of social presence from communication researchers, defined by Short, Williams, and Christie [48; p. 65] as the “degree of salience of the other person in the interaction [first part] and the consequent salience of the interpersonal relationship [second part].” Kreijns, Weidlich, and Rajagopal [28] redefined the first part as the “degree to which the other person is perceived as physically ‘real’ in the communication” and identified this as ‘social presence’, for which they developed a social presence measure to assess this realness. However, not all social presence researchers agree with this definition as illustrated by Lowenthal and Snelson [36]. The second part of the definition was identified as ‘social space,’ which is mentioned above. Social presence research claims that social presence influences participation, social interaction, learner satisfaction, and learner outcomes [13, 14, 61].

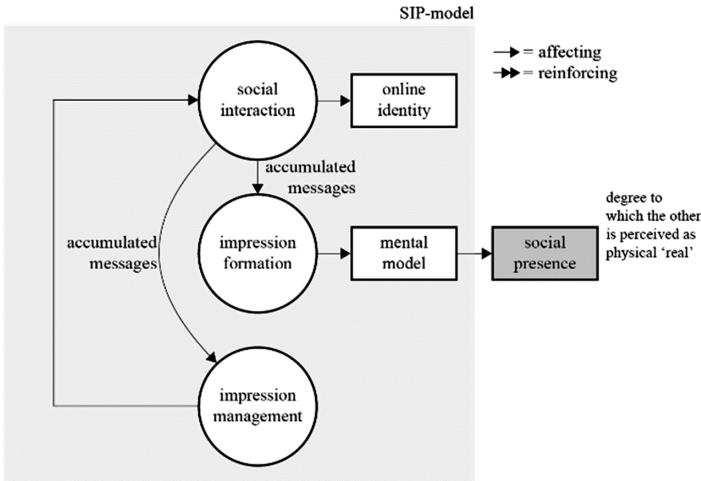
## 2.2 The SIP Model: Social Information Processing

**Impression Formation.** Walther’s [54] Social Information Processing (SIP) theory states that despite the fact that online communication lacks the full richness of face-to-face communication in terms of the extent to which communication media can transfer the physical signals conveying verbal and non-verbal cues, communicating partners still can develop interpersonal relationships. SIP-theory was a response to existing theories (e.g., media richness theory [7], cues-filtered-out theory [50], and social presence theory [48]) denying that interpersonal relationships can develop in lean media. According to these theories, if verbal and non-verbal cues cannot be transferred, behaviors that rely on these cues and which play an important role in developing interpersonal relationships

[48] such as intimacy [2] and immediacy [60] will be hampered. SIP-theory states that communicating partners develop interpersonal relationships over time even in lean media with possibly the same relational dimensions and qualities as face-to-face relationships. Given enough time, messages accumulate and through this accumulation and the compensation of non-transferable physical signals to express intimacy and immediacy behaviors (e.g., emoticons or particular spatial arrangement of words in the messages), communication partners form individuating impressions of each other resulting in corresponding mental models.

**Impression Management.** Impression formation and mental models are the bases on which the interpersonal relationships develop [54] and communication (i.e., social interaction) transforms them from impersonal into interpersonal and, in some cases, even into hyperpersonal [55]. To elaborate the latter, Walther's SIP-theory also includes a process of impression management; that is, the process in which communication partners determine how they will present themselves online and how to sustain this. Usually, communication partners create more favorable impressions of themselves to others by deciding what to share about themselves and what not. They are informed by the same accumulated messages—which now function as a feedback channel—whether they succeeded in this endeavor or if they have to make some adjustments. On the other hand, communicating partners also tend to evaluate and judge the accumulated messages more positively than they are, thereby idealizing the other communication partners, which is reflected in the mental models formed. The selective self-presentation and the idealized mental models cause the hyperpersonal effect. Walther [56] also showed that this hyperpersonal effect diminishes once communicating partners meet each other in a face-to-face setting.

The SIP-theory of impression formation and impression management, that explain how mental models of the communicating partners are formed and how communicating partners create online identities will ultimately have an effect on social presence as realness. The SIP-model in Fig. 2 graphically depicts the SIP-theory.



**Fig. 2.** Walther’s [54] SIP-model. Accumulated messages for impression management are filtered on feedback information about one’s own online identity; accumulated messages for impression formation are filtered on information about the other.

**2.3 The HES Model: Hedonicity, Educability, and Sociability**

The last model is the HES-model (*Hedonicity, Educability, and Sociability*), which represents an affordance perspective on online environments used for CL. The attributes hedonicity, educability, and sociability characterize OCL-environments. As such, these attributes contribute to the usefulness of the OCL-environment.

**Hedonicity.** Hedonicity expresses the extent to which OCL-environments give pleasure and enjoyment during the interacting with them. To do so, these OCL-environments should possess hedonic affordances. Gamification widgets are obvious choices for bringing hedonic affordances to the OCL-environment. Gamification is the application of game-design elements and game principles in non-game contexts [15]. But other, not gamification-based features in the OCL-environment, may also possess hedonic affordances and should be considered as well. In that respect, human-computer interaction (HCI) research on *funology* studies how we should understand and design for fun as a user experience [4]. Findings from HCI-research may inform the design of OCL-environments that exhibit hedonic affordances. Our search for literature on hedonic affordances in CSCL, however, made clear that the CSCL-research community is not yet exploring hedonic affordances that are built in OCL-environments and how they affect participation and learning and social performances with the exception of Suh and Wagner [51]. In that respect, the CSCL-research community lags behind the e-commerce community that has collected empirical evidence on the role of hedonicity and purchase intention of users visiting web-stores (see, for instance: [6]).

While the OCL-environment by itself may possess hedonic affordances, collaborative tasks may also have these affordances. For example, a difficult problem-solving task

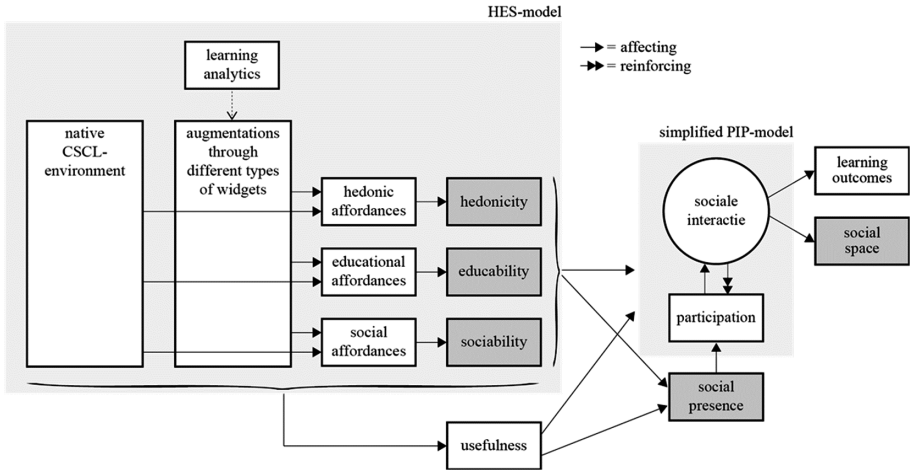


may cause enjoyment among group members when it is finally solved after hard work. Interestingly, in some cases negative hedonic value (e.g., frustration) in the short term combined with positive hedonic value in the long term may ultimately result in higher learning gains than when there was solely positive hedonic value throughout the task performance [21]. This suggests that striving for positive hedonic value all the time may not always be the best strategy. Hedonic value through gamification can also be designed into the collaborating tasks. Research has found gamification-based hedonic affordances in (collaborative) tasks to be important in increasing students' motivation to persevere [35]. The relationship between hedonic affordances and motivation for learning originates from the observed enjoyment and persistence when young people play computer games to reach next levels until the game is over. However, gamification may not always result in positive learning gains [8, 15]. When meaningful gamification is brought in the collaborative tasks, and the OCL-environment supports this type of gamification, it actually adds to the educability of the OCL-environment and, as a kind of spin-off, also its hedonicity.

**Educability.** Educability expresses the degree to which an online environment has educational affordances to support CL. If the online environment is oriented towards CL, these affordances are requisite.

**Sociability.** Sociability is the degree to which the OCL-environment supports social aspects; that is, the emergence of a sound social space with its associated qualities (e.g., positive group climate, sense of community, mutual trust) [32, 33]). Social affordances—elements in the OCL-environment that have potential for evoking specific actions—affect sociability of the OCL-environment; here, social interaction that serves social and socio-emotional processes. One kind of social affordance is aimed at reducing transactional distance. According to Moore [38], the distance in distance education is more than just geographical. It implies a psychological and a communication distance both between fellow students and with instructors. He designated this kind of distance as transactional distance which can be reduced through virtual proximity (or teleproximity) [32]. Research on the effects of physical proximity has shown that proximity facilitates impromptu encounters and informal or casual conversations. Festinger, Schachter, and Back [9] found that proximity leads to social relationships and even close friendships between people. One way to create virtual proximity in an OCL-group is to provide real-time group awareness information about all the other group members through group awareness widgets embedded in the virtual environments whether these are for learning, collaboration, information exchange, and so on. Group awareness is the condition in which one is informed about a number of issues including the availability of other persons, their whereabouts, their activities, and with whom a conversation can be started [31].

**Learning Analytics.** As can be seen from Fig. 3, the HES-model explicitly incorporates learning analytics to feed awareness information into the different types of widgets. Learning analytics are “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” [47]. These widgets visually display the awareness information in the OCL-environment.



**Fig. 3.** The HES-model applied for collaborative/group learning. In so far, the native CSCL-environment is lacking functionalities, augmentations are added through different types of widgets. These widgets—in the context of this paper—provide group awareness.

**Usefulness.** Figure 3 also shows another variable, namely usefulness which refers to both the utility and usability [40]. Utility refers to the functionalities available in a system, here the OCL-environment. The attributes hedonicity, educability, and sociability represent underlying functionalities that are required in the OCL-environments as advocated in this paper but in varying degrees present in current available environments for OCL. Usability is the ease-of-use of a system so that users can interact and perform their tasks in an intuitive way [40]. According to Preece [45; p. 27], a system with good usability “supports rapid learning, high skill retention, low error rates and high productivity [and] is consistent, controllable, and predictable, making it pleasant and effective to use.” It is also clear that usability also influences the degree of social presence and the social interaction; in a clumsily designed OCL-environment with bad usability, members are busier fighting the system than with learning.

**Support for the HES-Model.** The HES-model and its affordance perspective on OCL-environments seems to fit the uses and gratification theory (UGT; Katz, Bumler and Gurevitch [22; see also 37]). UGT purports that the extent to which media are selected and used depends on the degree to which four general motivational needs are gratified, namely: (1) integration and social interaction: the need to socialize by meeting new people and sustaining existing contacts via a sense of belonging and connectedness; (2) information: the need to self-educate, acquire new knowledge and understanding; (3) entertainment: the need for relaxation and enjoyment; and (4) personal identity: the need to reaffirm one’s individual identity by getting involved in activities of others who have similar interests or other things in common.

Brandtzæg and Heim [5], studying why people use social networking sites, confirmed these four motivational needs. If at least one of the four motivational needs is not fulfilled, the medium is at risk of non-use. In other words—and from the

perspective of HCI—if media misses functionalities for achieving some purposes (i.e., gratification of one or more of the motivational needs), its utility is neglectable and, as a result, it is designated as being useless; the medium will not be used [40]. Once again, hedonicity, educability, and sociability, if present, will avoid such a risk as they simultaneously address the four motivation needs: hedonicity addresses the entertainment need, educability the information need, and sociability the need for integration and social interaction (i.e., socialization). The three attributes together address the need for establishing personal identity. A recent study [1] using UGT on the linkage between social media and job performance saw three categories of media use, namely, the hedonic, cognitive, and social use to be responsible for job performance via social capital, thereby supporting the validity of the HES-model as these categories of uses correspond very well with the three attributes of it.

### 3 Putting it all Together

The extended SIPS-model integrates the three sub-models (i.e., PIP, SIP, and HES); see Fig. 4 with simplified versions of the sub-models. Furthermore, the extended SIPS-model is drawn to resemble earlier versions of it (see, Kreijns, Kirschner, and Jochems [30]; Kreijns, Kirschner, and Vermeulen [29]; and Kirschner, Kreijns, Phielix, and Franssen [25]).

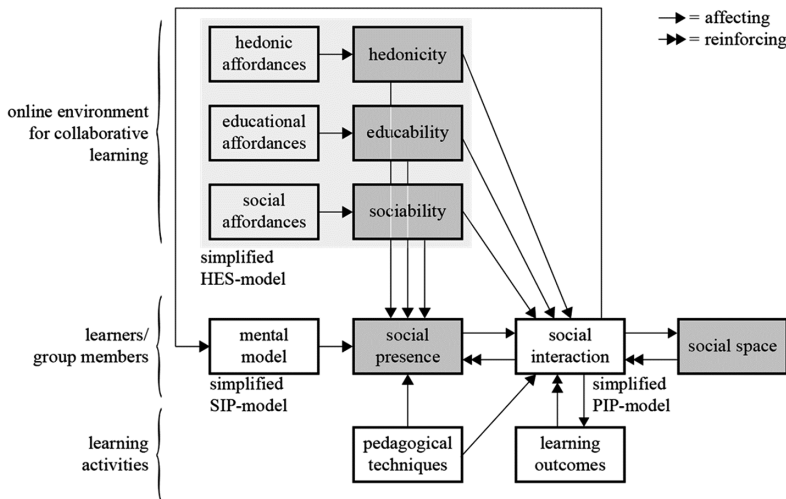


Fig. 4. The extended SIPS-model

#### 3.1 Discussion and Conclusion

This conceptual paper [12] extends the SIPS-model by introducing hedonicity in addition to educability and sociability. Three distinct sub-models were introduced, namely the PIP-, SIP-, and HES-models. The PIP-model— centering around social interaction

for socio-cognitive (where group learning/knowledge construction takes place) and socio-emotional processes (where group forming/dynamics takes place)—is of particular interest to the CSCL-community. It shows (not surprisingly) that pedagogical techniques directly affect participation and social interaction. Most research, therefore, concentrate on finding effective and efficient pedagogies such as those based on scripting. This research is mostly done in the context of computer supported CCL but rarely in the context of OCL. The PIP-model also shows group dynamics to be essential for OCL. Unfortunately, research on the effects of mediated communication in OCL on group dynamics is seldom an item on the CSCL research agenda. The PIP-model further shows that apart from academic skills, social skills are also important.

Impression formation and impression management as shown in the SIP-model may not be of interest in the context of computer-supported CCL, but is essential in the context of OCL as it affects the degree of social presence, either perceived (through impression formation; [13]) or projected (through impression management; [11]). OCL-group members, therefore, have to acquire the social skills for appropriate impression management. Especially, when social networking sites are used, impression management is becoming even more an important issue [26].

The HES-model is concerned with the OCL-environment. It is, therefore, of particular interest to the TEL-community. If OCL-environments are not well-designed (e.g., they lack functionalities such as a shared text-editor) or have badly implemented user interfaces, it will directly affect the OCL-members dispositions in that they will dislike the OCL-environment and not use it. Furthermore, the TEL-community should answer questions about how to design OCL-environments that possess hedonicity, educability, and sociability through their respective affordances. This is not a trivial matter. One way to realize these affordances is by means of group awareness widgets [31, 32] and gamification widgets.

We hope that the extended SIPS-model and its sub-models (PIP, SIP, HES) are helpful as a research framework for OCL—and potentially also for computer-supported CCL—because they capture all the important issues of CSCL-research and show important relationships between the many variables involved. But as was already made clear in Kreijns, Kirschner, and Vermeulen [29], many of the relationships are still hypothetical and future research should investigate them.

## References

1. Ali-Hassan, H., Nevo, D., Wade, M.: Linking dimensions of social media use to job performance: The role of social capital. *J. Strateg. Inf. Syst.* **24**(2), 65–89 (2015)
2. Argyle, M., Dean, J.: Eye contact, distance and affiliation. *Sociometry* **28**, 289–304 (1965)
3. Aronson, E., Patnoe, S.: *The jigsaw classroom*. Addison Wesley Longman, New York (1997)
4. Blythe, M., Monk, A. (eds.): *Funology 2: From usability to enjoyment*. Springer, Cham (2018)
5. Brandtzæg, P.B., Heim, J.: Why people use social networking sites. In: Ozok, A.A., Zaphiris, P. (eds.) *OCSC 2009*. LNCS, vol. 5621, pp. 143–152. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-02774-1\\_16](https://doi.org/10.1007/978-3-642-02774-1_16)
6. Chen, W.-K., Chang, D.-S., Chen, C.-C.: The role of utilitarian and hedonic values on users' continued usage and purchase intention in a social commerce environment. *J. Econ. Manag.* **13**(2), 193–220 (2017)

7. Daft, R.L., Lengel, R.H.: Information richness: a new approach to managerial behavior and organizational behavior, vol. 6, pp. 191–233. JAI Press, Greenwich (1984)
8. Dichiva, D., Dichev, C., Agre, G., Angelova, G.: Gamification in education: A systematic mapping study. *Educ. Technol. Soc.* **18**(3), 1–14 (2015)
9. Festinger, L., Schachter, S.S., Back, K.W.: Social pressures in informal groups: A study of human factors in housing. Stanford University Press, Stanford (1950)
10. Fischer, F., Kollar, I., Stegmann, K., Wecker, C.: Toward a script theory of guidance in computer-supported collaborative learning. *Educ. Psychol.* **48**(1), 56–66 (2013)
11. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking in a text-based environment: computer conferencing in higher education. *Internet High. Educ.* **2**(2), 87–105 (2000)
12. Gilson, L.L., Goldberg, C.B.: Editors' comment: so, what is a conceptual paper? *Group Organ. Manag.* **40**(2), 127–130 (2015)
13. Gunawardena, C.N.: Social presence theory and implications for interaction and collaborative learning in computer conferences. *Int. J. Educ. Telecommun.* **1**(2&3), 147–166 (1995)
14. Hostetter, C., Bush, M.: Community matters: social presence and learning outcomes. *J. Sch. Teach. Learn.* **13**(1), 77–86 (2013)
15. Huang, W.H.-Y., Soman, D.: A practitioner's guide to gamification of education. Research Report Series: Behavioural Economics in Action, University of Toronto, Rotman School of Management (2013)
16. Huotari, K., Hamari, J.: Defining gamification: a service marketing perspective. In: Proceedings of the 16th International Academic MindTrek Conference, pp. 17–22. ACM, New York
17. Jacques, D.: Learning in groups, 2nd edn. Kogan Page, London (1992)
18. Johnson, D.W., Johnson, R.T.: An educational psychology success story: social interdependence theory and cooperative learning. *Educ. Res.* **38**(5), 365–379 (2009)
19. Johnson, D.W., Johnson, R.T.: Critical thinking through structured controversy. *Educ. Leadersh.* **45**(8), 58–64 (1988)
20. Johnson, D.W., Johnson, R.T., Smith, K.A.: Cooperative learning: improving university instruction by basing practice on validated theory. *J. Excell. Coll. Teach.* **25**(3–4), 85–118 (2014)
21. Kapur, M.: Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educ. Psychol.* **51**(2), 289–299 (2016)
22. Katz, E., Blumler, J.G., Gurevitch, M.: Uses and gratifications research. *Public Opin. Q.* **37**(4), 509–523 (1973)
23. Kirschner, P.A., Gerjets, P.: Instructional design for effective and enjoyable computer-supported learning. *Comput. Hum. Behav.* **22**(1), 1–9 (2006)
24. Kirschner, P.A., Sweller, J., Kirschner, F., Zambrano, J.R.: From cognitive load theory to collaborative cognitive load theory. *Int. J. Comput.-Support. Collab. Learn.* (2018) (First online)
25. Kirschner, P.A., Kreijns, K., Phielix, C., Fransen, J.: Awareness of cognitive and social behaviour in a CSCL environment. *J. Comput. Assist. Learn.* **31**(1), 59–77 (2015)
26. Krämer, N.C., Winter, S.: Impression management 2.0: the relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *J. Media Psychol.* **20**(3), 106–116 (2008)
27. Kreijns, K.: Sociable CSCL environments: social affordances, sociability, and social presence. Unpublished PhD dissertation. Open Universiteit Nederland, Heerlen (2004)
28. Kreijns, K., Weidlich, J., Rajagopal, K.: The psychometric properties of a preliminary social presence measure using Rasch analysis. In: V. Pammer-Schindler et al. (Eds.): EC-TEL 2018, LNCS 11082, pp. X–XY. Springer, AG (2018)

29. Kreijns, K., Kirschner, P.A., Vermeulen, M.: Social aspects of CSCL environments: a research framework. *Educ. Psychol.* **48**(4), 229–242 (2013)
30. Kreijns, K., Kirschner, P.A., Jochems, W.: Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Comput. Hum. Behav.* **19**(3), 335–353 (2003)
31. Kreijns, K., Kirschner, P. A., Jochems, W.: Supporting social interaction for group dynamics through social affordances in CSCL: group awareness widgets. Paper presented at the 10th European Conference for Research on Learning and Instruction (EARLI). Padova, Italy (2003)
32. Kreijns, K., Kirschner, P.A., Jochems, W.: The sociability of computer-supported collaborative learning environments. *J. Educ. Technol. Soc.* **5**(1), 8–22 (2002)
33. Kreijns, K., Kirschner, P.A., Jochems, W., van Buuren, H.: Measuring perceived sociability of computer-supported collaborative learning environments. *Comput. Educ.* **49**(2), 176–192 (2007)
34. Kreijns, K., Kirschner, P.A., Jochems, W., van Buuren, H.: Measuring perceived quality of social space in distributed learning groups. *Comput. Hum. Behav.* **20**(5), 607–632 (2004)
35. Landers, R.N., Armstrong, M.B.: Enhancing instructional outcomes with gamification: an empirical test of the technology-enhanced training effectiveness model. *Comput. Hum. Behav.* **71**, 499–507 (2017)
36. Lowenthal, P.R., Snelson, C.: In search of a better understanding of social presence: an investigation into how researchers define social presence. *Distance Educ.* **38**(2), 1–19 (2017)
37. McQuail, D.: *Mass communication theory: An introduction*. Sage, London (1994)
38. Moore, M.G.: Theory of transactional distance. In: Keegan, D. (ed.) *Theoretical principles of distance education*, pp. 22–38. Routledge, Abingdon (1993)
39. Morrison, B., Collins, A.: Epistemic fluency and constructivist learning environments. In: Wilson, B. (ed.) *constructivist learning environments*, pp. 107–119. Educational Technology Press, Englewood Cliffs (1996)
40. Nielsen, J.: *Usability engineering*. Morgan Kaufmann Publishers, San Francisco (1994)
41. OECD: *PISA 2015 Results (Volume V): Collaborative Problem Solving*. PISA, OECD Publishing, Paris (2017)
42. Ohlsson, S.: Learning to do and learning to understand: a lesson and a challenge for cognitive modeling. In: Reimann, P., Spada, H. (eds.) *Learning in humans and machines*, pp. 37–62. Pergamon, Oxford (1996)
43. Richardson, J.C., Maeda, Y., Lv, J., Caskurlu, S.: Social presence in relation to students' satisfaction and learning in the online environment: a meta-analysis. *Comput. Hum. Behav.* **71**, 402–417 (2017)
44. Rourke, L., Anderson, T.: Exploring social communication in asynchronous, text-based computer conferences. *J. Interact. Learn. Res.* **13**(3), 259–275 (2002)
45. Preece, J.: *Online communities: designing usability, supporting sociability*. Wiley, New York (2000)
46. Scardamalia, M., Bereiter, C.: Knowledge building: theory, pedagogy, and technology. In: Sawyer, K. (ed.) *Cambridge Handbook of the learning sciences*, pp. 97–118. Cambridge University Press, New York (2006)
47. Siemens, G.: *Learning and Academic Analytics* (2011). <http://www.learninganalytics.net/?p=131>. Accessed 17 Jun 2018
48. Short, J., Williams, E., Christie, B.: *The social psychology of telecommunications*. Wiley, London (1976)
49. Smith, M., Kollock, P. (eds.): *Communities in cyberspace*. Routledge, London (1998)

50. Sproull, L., Kiesler, S.: *Connections: New ways of working in the networked organization*. MIT Press, Cambridge (1991)
51. Suh, A., Wagner, C.: How gamification of an enterprise collaboration system increases knowledge contribution: an affordance approach. *J. Knowl. Manag.* **21**(2), 416–431 (2017)
52. Sweller, J., Ayres, P., Kalyuga, S. (eds.): *Cognitive load theory*. Springer, New York (2011)
53. Von Krogh, G., Nonaka, I., Ichijo, K.: *Enabling knowledge creation*. Oxford University Press, New York (2000)
54. Walther, J.B.: Impression development in computer-mediated interaction. *W. J. Commun.* **57**, 381–398 (1993)
55. Walther, J.B.: Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Commun. Res.* **23**(1), 3–43 (1996)
56. Walther, J.B.: Visual cues and computer-mediated communication: don't look before you leap. Paper presented at the annual meeting of the International Communication Association, San Francisco (1999)
57. Weidlich, J.B., Bastiaens, T.: Explaining social presence and the quality of online learning with the SIPS model. *Comput. Hum. Behav.* **72**, 479–487 (2017)
58. Weinberger, A.: *Scripts for computer-supported collaborative learning: effects of social and epistemic scripts on collaborative knowledge construction*. Unpublished PhD dissertation. Ludwig-Maximilians-Universität, München (2003)
59. Weinberger, A., Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Comput. Educ.* **46**(1), 71–95 (2006)
60. Wiener, M., Mehrabian, A.: *Language within language: immediacy, a channel in verbal communication*. Apple-Century-Crofts, New York (1968)
61. Zhao, H., Sullivan, K.P.H., Mellenius, I.: Participation, interaction and social presence: an exploratory study of collaboration in online peer review groups. *Br. J. Edu. Technol.* **45**(5), 807–819 (2014)



# A Syllogism for Designing Collaborative Learning Technologies in the Age of AI and Multimodal Data

Mutlu Cukurova<sup>(✉)</sup>

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

**Abstract.** Different paradigms of research interpret the social reality in different ways and these differences are not always apparent in technology enhanced learning research. However a paradigm's visibility and its elements' internal consistency are fundamental to the quality of research. As a philosophical position, a paradigm guides researchers to understand the nature of reality (ontology); how we create, acquire and disseminate knowledge (epistemology); and a systematic set of research strategy (methodology). In this research paper, the relationship between ontology, epistemology, and methodology is defined within the context of designing multimodal, AI technologies for collaborative learning. Two case study examples of inductive and deductive research methodologies are presented with the purpose of clarifying their differences in research outputs. Moreover, based on a recent literature review, it is presented that most empirical research in the field (40 out of 46) falls under the inductive methodology. Although, both deductive and inductive approaches are valuable for the advancement of the field; it is argued that the apparent lack of deductive investigations may lead researchers falling into technological determinism.

**Keywords:** Deductive research · Inductive research  
Collaborative learning technologies · Artificial intelligence

## 1 Introduction

Social reality is the main subject of educational research, but it is interpreted in multiple ways by different paradigms of research. In his seminal work “the structure of scientific revolutions”, Thomas Kuhn conceptualises the term ‘paradigm’ in two different senses (1962, 1970). The first one defines a paradigm as “the entire constellation of beliefs, values, techniques, and so on, shared by the members of a given community” (p. 175). The second one is as exemplary of past achievements, “the concrete puzzle-solutions which, employed as models or examples, can replace explicit rules as a basis for the solution of the remaining puzzles of normal science.” (p. 175). This paper refers to the first conceptualisation of paradigm. More specifically here, a paradigm is referred as a set of beliefs, values, and assumptions that are shared by the members of the research community that undertakes research into the design of collaborative learning technologies with AI techniques and multimodal data.



Unfortunately, often the paradigms of the research we are undertaking are not explicit to other researchers in the field, and in some cases, it is not even explicit to us. Optimistically, this is due to such assumptions and values being deeply embedded in our thinking, and even though we do not explicitly reflect upon them, our research actions are internally consistent within our paradigm. Pessimistically, we are constantly shifting between the ontology, epistemology, and methodologies of different paradigms, lacking an internally consistent way of thinking about the research we are undertaking. The purpose of this paper is to remind us that *regardless of the technological development era we are operating in, it is not possible to undertake research without committing to a paradigm and their transparency and consistency is paramount to ensure research quality.*

Every good researcher's decision to reject one paradigm almost simultaneously means that they make the decision to accept another, and ideally, the judgment leading to that decision involves an informed comparison of different paradigms with nature and with each other (Kuhn 1970). However, whether it is conscious or not, in the research we undertake we take a decision to accept one paradigm which becomes apparent in our methodological decisions. These different methodologies we implement lead to different research products and outputs. This paper presents two recent research studies on the design of collaborative learning technologies within the context of multimodal learning analytics to exemplify the different research products created due to the use of different methodologies; even though both approaches have similar data sources, tools, research contexts, and purposes.

## 2 The Relationship Between Ontology, Epistemology, and Methodology

As discussed in the introduction section, the social reality is interpreted by multiple perspectives and differences can best be understood by an analysis of the assumptions that underpin research. One set of assumptions are the ontological ones, which mainly aims to provide answers to the key question of how the reality is defined. Ontology asks philosophical questions such as, is social reality external to an individual - imposing itself from without - or is it the product of individual's consciousness? Ontology is the study of being and ontological assumptions are concerned with what constitutes reality. For example, if a researcher believes that the social reality is external to an individual, and it exists outside the existence of the individual inquiring about it, their ontological position can be considered as a realist, assuming the facts are independent of mind. Alternatively, if a researcher considers that the reality is the product of individual consciousness and its pure existence is dependent upon the individual who enquires about it, they can be considered as an 'anti-realist'. For instance, in a constructivist paradigm it is argued that meaning does not exist in its own right; rather it is constructed by human beings as they interact and engage in interpretation. Constructivist ontological assumption is that the reality is socially constructed. The reality is perceived based on the very context in a given situation and can hardly be generalised into one common

reality. This perception directly challenges the typical positivist view that reality perceived in one context could be transferred to another with a similar setting (Table 1).

**Table 1.** Assumptions of different research paradigms of positivism and post-positivism about key concepts of research (Adapted from O’Leary (2004, p. 7)

Positivist assumptions	Assumptions about	Post-positivist assumptions
Knowable	←the world⇒	Ambiguous
Predictable	←the world⇒	Variable
Single truth	←the world⇒	Multiple reality
Empirical	←the nature of research⇒	Intuitive
Reductionist	←the nature of research⇒	Holistic
Objective	←the researcher⇒	Subjective
Removed expert	←the researcher⇒	Participatory & Collaborative
Deductive	←methodology⇒	Inductive
Hypothesis-driven	←methodology⇒	Exploratory
Reliable	←methodology⇒	Dependable
Reproducible	←methodology⇒	Auditable
Often quantitative	←findings⇒	Often qualitative
Statistically significant	←findings⇒	Valuable
Generalisable	←findings⇒	Idiographic or transferable

Another set of assumptions outlined are epistemological. These assumptions relate to the questions about the nature of knowledge, how is knowledge created, what are its forms, and, how can it be communicated. In constructivism epistemological assumptions indicate that knowledge is subjective because it is socially constructed. Therefore, knowledge generation is bound to the context where the acquisition of knowledge is happening. It is argued that therefore the epistemological positioning of the constructivist paradigm seeks to understand how social actors recognise, produce, and reproduce social actions and how they come to share an intersubjective understanding of specific life circumstances. On the other hand, post-positivists believe that knowledge about reality could be gained through an empirical evaluation. Nevertheless, because of the researcher limitations, any theories and knowledge well-established are only tentatively proven until disapproved by new evidence and findings (Mertens 2014).

Ontological and epistemological positions play a direct role determining the research methodology that will be implemented in research. A methodology is the entire set of research strategies. In other words, it is the summary of the research process that ensures the data collected and the chosen study context are in line with the knowledge that research question(s) intend to obtain. And finally, as a component of the methodology, methods are data collection and analysis techniques. For instance, a researcher taking a positivist paradigm tend to form an abstraction of reality through primarily quantitative models using an experimental or quasi-experimental design with deductive hypotheses (Mertler 2016). They will be taking an “outsider” position aiming to test their hypotheses. In contrast a constructivist position researcher will often use more flexible research methodologies in which the subjectivities are clearly defined and explained. They may

form part of the research, taking an “insider” position, acknowledging that the knowledge investigated is socially constructed and inductively investigated.

### 3 Two Case Studies with Different Research Paradigms to Design Collaborative Learning Technologies

Here, inductive and deductive research methodologies will be focussed to exemplify two different paradigms’ research approaches. The differences in the research products will be illustrated with two case studies both aiming to design collaborative learning technologies with multimodal data collected from the same project-based learning context.

#### 3.1 Case Study 1: Deductive Approach

The case study example for the deductive approach is a recent paper by Cukurova *et al.* (2018). The purpose of the paper is to identify students’ effective collaborative problem-solving (CPS) behaviours in real-world teaching environments so that technology that observes such behaviours can be designed and can be used to support skill development. The article starts with a definition of CPS and presents a literature review on the mechanisms through which CPS may influence cognition and support deeper learning. The researchers identified four key constructs from the learning sciences literature that are argued to be relevant to the process of CPS, namely synchrony, individual accountability, equality, and intra-individual variability and experimentally investigate their relation to CPS. Their results show that students in high competence CPS groups have member students who have high and equal scores for physical interactivity and low and equal scores for intra-individual variability. Moreover, high competence CPS groups appear to have high levels of student synchrony and individual accountability values. Based on these results, taking a deductive approach, the authors argue that the future research will involve attempts to design a piece of technology to automate this process of interpreting student behaviours using multimodal learning analytics in order to provide real-time feedback to students and teachers about learning processes. More recently, they designed a computer vision system based on multimodal data and deep neural networks that is able to detect those key constructs of CPS that are deductively created in NISPI framework (Landolfi *et al.*, under review).

#### 3.2 Case Study 2: Inductive Approach

With the same purpose of identifying and supporting students’ effective CPS behaviours in project-based learning environments, Spikol *et al.* (2018) investigates the potential of data collected from highfidelity synchronised multimodal recordings of small groups of learners interacting. As opposed to the NISPI paper’s approach of deductively identifying key constructs of CPS, in this article the authors process and extract different aspects of the students’ interactions to identify which features are representative of success in educational contexts of opened project work. Inductively exploring

multiple data sources with different machine learning approaches, the authors investigate the potential of number of the faces looking at the screen, the mean distance between learners, the mean distance between hands, the mean hand movement speed, the mean audio level, project complexity, active hardware and software blocks, and the students' work phases. They conclude that the distance between learners' hands and faces is a strong predictor of student collaboration, whereas other features do not predict the project outcomes and student collaboration.

## 4 Discussion and Conclusions

Case studies presented above have the same goal of designing a piece of technology that would support collaborative learning in project-based learning environments, yet the decisions that are taken during the research process are distinct from each other. More importantly, even though both research studies investigate almost identical research contexts with almost identical data sources and almost identical data intelligence tools, due to different paradigms underpinning two case studies, they produce completely different research products. At this stage, one could argue for the potential superiority of one approach over the other. However, this is not the point of this paper. Ontology and epistemology are axiological, that means they are related to values (Carter and Little 2007). Therefore their associated methodologies are bound to certain assumptions and values. The purpose here is not to argue for, or against, any of them as they are incommensurable. However, it is to argue that good quality research should clearly explain the methodological decisions taken and argue for the internal consistency of its elements (Mantzoukas 2004). As long as such internal consistency is provided, arguments that relate to ultimate superiority of a certain paradigm over other would require repression of existing useful logics (Carter and Little 2007), and the world of ideas does not call for one true 'logic' (Kaplan 1964).

Although, there are research papers emphasising on the value of considering epistemic beliefs in the design of learning analytics (c.f. Knight *et al.* 2014); epistemic transparency is often not present. For instance, in Worsley (2018)'s recent review of the field of multimodal learning analytics (MMLA), eighty-two papers (forty-six empirical) were identified and there is hardly any mention of the operated research paradigms in these identified papers. Considering the emerging nature of the field, perhaps, this is expected. However, based on considerations exemplified in two case studies above, it also becomes apparent that most research published in the field takes an inductive approach (fourty out of forty-six empirical studies). This might be problematic; not due to the inferiority of this approach over the other, but due to the potential monopoly of one particular research paradigm in the field. More varied approaches, such as the deductive approach presented in case study one, can lead to MMLA's richer contribution to Educational contexts. For instance, the prevalence of inductive approach might lead to prioritising the existing data sources and tools over designing new ones that are based on the requirements of existing research in learning sciences and may lead to the production of research outcomes that are not of great value to authentically situated

learning environments. This limits the ways we can challenge the technology, and might ultimately lead to technological determinism.

## References

- Carter, S.M., Little, M.: Justifying knowledge, justifying method, taking action: epistemologies, methodologies, and methods in qualitative research. *Qual. Health Res.* **17**(10), 1316–1328 (2007)
- Cukurova, M., Luckin, R., Milln, E., Mavrikis, M.: The NISPI framework: analysing collaborative problem-solving from students' physical interactions. *Comput. Educ.* **116**, 93–109 (2018)
- Kaplan, A.: *The Conduct of Inquiry: Methodology for Behavioral Science*. Chandler, San Francisco (1964)
- Knight, S., Shum, S.B., Littleton, K.: Epistemology, assessment, pedagogy: where learning meets analytics in the middle space. *J. Learn. Anal.* **1**(2), 23–47 (2014)
- Kuhn, T.S.: *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago (1962)
- Kuhn, T.S.: Logic of discovery or psychology of research. In: *Criticism and the Growth of Knowledge*, pp. 1–23 (1970)
- Landolfi, L., Ruffaldi, E., Cukurova, M., Spikol, D.: Collaboration analysis of students' physical interaction based on neural networks and body pose. *IEEE Trans. Learn. Technol.* (under review)
- Mantzoukas, S.: Issues of representation within qualitative inquiry. *Qual. Health Res.* **14**(7), 994–1007 (2004)
- Mertens, D.M.: *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods*. Sage, Thousand Oaks (2014)
- Mertler, C.A.: *Action Research: Improving Schools and Empowering Educators*. Sage Publications, Thousand Oaks (2016)
- 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.* (2018). <https://doi.org/10.1111/jcal.12263>
- O'leary, Z.: *The Essential Guide to Doing Research*. Sage, London (2004)
- Worsley, M.: Multimodal learning analytics past, present, and, potential futures. In: *Companion Proceedings of LAK 2018, CrossMMLA* (2018)



# “Make It Personal!” - Gathering Input from Stakeholders for a Learning Analytics-Supported Learning Design Tool

Marcel Schmitz<sup>1(✉)</sup>, Maren Scheffel<sup>2</sup>, Evelien van Limbeek<sup>1</sup>,  
Roger Bemelmans<sup>1</sup>, and Hendrik Drachsler<sup>2,3,4</sup>

<sup>1</sup> Zuyd University of Applied Sciences, Heerlen, Netherlands  
{marcel.schmitz,evelien.vanlimbeek,roger.bemelmans}@zuyd.nl

<sup>2</sup> Open Universiteit, Heerlen, Netherlands

{maren.scheffel,hendrik.drachsler}@ou.nl

<sup>3</sup> Goethe University, Frankfurt, Germany

<sup>4</sup> German Institute for International Educational Research (DIPF),  
Frankfurt, Germany

**Abstract.** Teachers design learning activities purposefully to improve student learning. However, the impact of this is usually only evaluated after a course has ended by making use of self-reported data and assessment results. Learning analytics offers the opportunity to collect, analyse and visualise feedback on activities using authentic data in real-time. Incorporating learning analytics into the learning design makes just-in-time interventions attainable. This paper presents the first steps of the development of a Learning Analytics for Learning Design (LA4LD) tool that is co-created with students and teachers, using a design-based research methodology. Both teachers and students express the need to personalise feedback on learning activities in order to increase the quality of the learning process and want that embedded in the tool.

**Keywords:** Learning analytics · Learning design  
User-centred design · Teachers · Students · Higher education

## 1 Introduction

Teachers design learning activities to improve students' learning. A learning activity can be any form of interaction between a student and either a teacher, other students, or content [1]. More and more of such learning activities nowadays take place in an online environment. This is of course true for online or distance education; but also in cases where classes are taught in a face-to-face setting do students engage with the course material outside of the class and participate in learning activities in an online learning environment [2]. The learning design connects learning activities (interactions) to certain goals (learning outcome) [8]. Learning designs are evaluated after modules are done. Institutes collect results,

conduct surveys, talk to student boards, analyse the data and –based on that input– (re)design learning activities. Learning analytics (LA) can improve these efforts [7]. Collecting and analysing student data can improve the quality of the feedback and also the opportunity of using on-demand indicators for evidence-informed decisions [9] on a course’s learning design (LD).

To our knowledge there are just a few learning analytics tools with regard to learning design that give real-time feedback to both students and teachers on learning activities studied. We looked at the review by Park et al. [11], the only tool they mention that reflects on learning activities is StepUP! [16] but it does so for students only. We also consulted the review by Schwendimann et al. [18] and the only tool with a strong connection to learning design that we found there is the AEEA Software Suite [3]. It has a design and a run-time part. In the design part a learning design is set up based on a set of competences, and in the run-time part analytics are used to give insight and feedback on learning activities. The set-up is promising and will be taken into consideration while designing our own tool. After the publication of the article in 2013 there was no follow-up for this tool that we know of. In their review Jivet et al. [5] argue that although there are learning dashboards available that give input on self-regulated learning processes, they only incorporate part(s) of the self-regulation cycle and none cover the whole process. The latest LAK conference had one further work interesting for our case: Nguyen et al. [10] present research on a dashboard which connects online activity, learning design and grades. In this case, however, the current stakeholders for using the dashboard seem to be researchers in stead of teachers and students during the course.

While there are a number of cases to be found in the literature that do provide LA during a course, none of them seem to use these results to analyse and evaluate the LD during the runtime of the course. If at all, they only do so afterwards. A solution where analyses and evaluations of the LD to improve the learning activities is done during a course could not be found. In order to close this gap, we decided to develop a Learning Analytics for Learning Design (LA4LD) tool that enables teachers and students to get on-demand feedback on learning activities during the run-time of a course. Such a tool allows teachers to improve the LD by adapting the learning activities and empowers students to adapt their learning processes to the learning activities.

Students in higher education need to have (or to develop) a high level of self-regulation [19]. Self-regulated learning [20] involves three features: usage of learning strategies, responsiveness to self-oriented feedback about learning effectiveness, and the motivational processes. It is a cycle of forethought, performance and self-reflection. Presenting the LA results during a course may help students to adapt their learning processes and teachers to adapt their LD during the run-time of a course. This in turn can lead to an increase in students’ learning outcomes and their satisfaction [14, 17].

Following the design science process by Hevner [4], we previously derived opportunities and challenges from the literature and presented a theoretical model for a LA4LD tool [17]. In addition to this theoretical grounding, however,

requirements and insights need to be gathered from the context and the involved stakeholders. We therefore conducted two studies. Study 1 collected requirements guided by research questions 1 and 2 while study 2 gathered insights on self-regulation in our context guided by research question 3:

- (RQ1) What do teachers want to see displayed in their LA4LD tool?
- (RQ2) What do students want to see displayed in their LA4LD tool?
- (RQ3) What is the current state of self-reflection from the students?

The rest of the paper is structured as follows: Sect. 2 describes the methods used in the studies while Sect. 3 presents the results of each study. Section 4 then combines these results in a discussion. Finally, Sect. 5 concludes the paper and provides an outlook on future work.

## 2 Methods

In study 1 (S1) we used focus groups with different set ups and different participants on the data and the dashboard perspective to get requirements for our LA4LD tool. There are participants and focus groups on data perspective – students (A1), teachers (A2) and TEL experts (A3)– as well as on the dashboard perspective: first year students (B1), post first year students (B2) and teachers (B3). In study 2 (S2) we used a survey to ask students about their self-regulated learning processes and how they perceive LA from a data perspective.

### 2.1 Study 1: Focus Groups

**Participants.** In order to gather insights into students’ and teachers’ perceptions about data collection, we conducted six focus groups. Three of them focused on the aspect of ‘data as an input mean’ (groups A1, A2 and A3), where two of them are populated by students and teachers and the third focus group was used to get an overview of the current data perspective of our context by inviting TEL experts. The other three focused on ‘dashboards as an output mean’ (B1, B2 and B3). Due to organisational and time constraints of all participants involved, i.e. students, teachers, TEL experts as well as the group moderators, we decided to do separate focus groups for the two investigated perspectives of data and dashboard. All participants were students or teachers at the faculty ICT of Zuyd University of Applied Sciences (Zuyd) in the Netherlands. They were invited to take part in the focus groups via personal invitation. Informed consent to participate in the study was obtained from all participants.

In group A1 there were five students: all male. The group had one first year student, one second year student, one third year student, one fourth year student and one alumnus. In group A2 there were five teachers: two female and three male. Two of the teachers also have the role of a study coach at the faculty and one of the teachers also has the role of team leader. In group A3 there were five technology-enhanced learning (TEL) experts affiliated to faculty ICT: one female and four male. The group consisted of a TEL advisor on the faculty



level, two TEL advisors on the institutional level, a TEL technical support staff member, and a TEL external consultant.

In group B1 there were six first year students: all male. In group B2 there were six students from post-first-year study years: all male. The group consisted of two students from the second year, two from the third year, one from the fourth year and one alumnus. In group B3 there were six teachers: one female and five male. One of them also had the role of a study coach at the faculty, one of them had a management role, two of them were also researchers and one was from the faculty ICT.

The reason to organise separate dashboard perspective focus groups for first year students and students in later years was that higher education institutions in the Netherlands are made accountable for efficiency of both the first year as well as the entire study. In the first year students have the opportunity for orientation and the institution has the opportunity for selection. The information students need about their learning activities and what they would like to see on dashboards is thus likely to differ between first year students and students from later study years.

**Material and Procedure.** All focus groups took place in a face-to-face setting, lasted about one hour and took place between 29-5-2017 and 31-5-2017. The participants gave their informed consent at the beginning of the session. The three focus groups on the data perspective (A1, A2, A3) were audio recorded. The three focus groups on the dashboard perspective (B1, B2, B3) were not recorded due to technical issues.

For all three focus groups on the data perspective the moderator used guiding questions (see Table 1) to start the discussion about data collection and related issues. Throughout the discussion participants were asked to take notes about those issues that are most important to them. These notes were then discussed by the moderator and the participants together and a ‘cloud of demands’ was created. After the sessions the moderator provided a summary of the session to the participants. The guiding questions for groups A1 and A2 were inspired by relevant literature on what tools/data indicators are used to get insight in learning processes while those for group A3 were based on the output of the previous groups.

For the dashboard perspective focus groups a design thinking set-up comparable to Stanford D-School [13] was chosen. The method consists of several steps: emphasise, define, ideate, prototype and test. To start the discussions, insight cards were used. The insight cards were defined after analysing relevant literature about descriptions of dashboard configurations that concern LA4LD aspects and making a strengths, weaknesses, opportunities and threats analysis of the descriptions (emphasise). All focus group participants were asked to rank the insight cards according to how interesting the mentioned aspects are (define). In a next step, participants were asked to compile post-its about the selected insight cards (ideate). Following this, participants co-created elements, solutions and visualisations for the topics and elements collected from the previous step

**Table 1.** Guiding questions for the participants of the data perspective focus groups

<b>Guiding questions for students and teachers</b>
What are your ideas on aspects of the study where learning analytics would be beneficial.
Is it only hard data (login, download and online presence data) that should/may be stored or also soft data like (emotions and learning styles).
Should the analyses be anonymous?
Do you see ways to use learning analytics in the modules you are in now?
Do you have sufficient insight in your own progress, now? (students only)
<b>Guiding questions for TEL experts</b>
Which learning tools are being used at UUU at the moment?
In which ways can those tools be connected?
If not, why aren't they connected at the moment?
Are there things to take into considerations when connecting the tools, like policy rules or legislation wise?
Are there analytics systems available at this moment?

(prototype). The final output of the focus groups were a set of configurations, wants and needs, and visualisation samples.

## 2.2 Study 2: Survey

**Participants.** For the survey 575 bachelor students from the faculty ICT of Zuyd were invited to participate. About 25% of these students, i.e. 143, responded positively and participated in the study. Informed consent was obtained from all 143 students.

**Material and Procedure.** The survey consisted of two parts: one about self-regulation and one about data collection. For the first part we chose a subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) [12]. The MSLQ is a commonly used questionnaire to measure the types of learning strategies and academic motivation of students [15] that consists of 81 questions in total. For this study, however, only the meta-cognitive self-regulation scale was used (the individual scales of the MSLQ can be used separately). The chosen scale describes the self-regulation competences of planning (4 items), monitoring (5 items) and regulating (3 items). All items are rated on a scale from 1 for not agree to 7 for totally agree.

The second part of the survey consisted of questions regarding the students' perception of the data collection about their learning processes. The questions were designed specifically for this study. Although they are not part of a validated questionnaire, they provide useful input to our study. The survey was created using Questback<sup>1</sup> and was sent to the students via email on May 23, 2017. Students were given two weeks to answer.

<sup>1</sup> <https://www.questback.com>.

## 3 Results

### 3.1 Study 1: Focus Groups

**Focus Group A1:** The discussion in the focus group with students on the data perspective had a lot of input. Students took the opportunity to reflect on the current learning design, on the way technology-enhanced learning was used in courses and gave ideas how to improve the design or use technology in their educational setting. The moderator had to redirect the discussion several times towards learning analytics and the questions: what data do you want to share and what in your opinion seems relevant to report on. Table 2 describes what information students want to get. Students also mentioned that they would like to use a learning analytics tool to give reflection on the learning activity. They would like to provide data on the difficulties they experiences, the task value, the quality of the learning material and performance of the teacher.

**Focus Group A2:** In the data perspective focus group for teachers one of the teachers quickly started the discussion: he wanted to use analytics to get insights in the way students learn. He quickly added that he wanted to see how learning design elements were used by students. The data needed in his opinion are the number of downloads, usage of material, personal information like living at home or on campus and success- or risk factors. Another teacher added to this that social and emotional aspects of students would be as interesting. Teachers want to know from the entire group of students when they study, how they study and on what aspects of the learning activity they have difficulties. A better view of group work and the achievement of the individual would be beneficial. Table 2 shows the elements teachers want to get information on.

**Focus Group A3:** The target audience of this focus group was the technology-enhanced learning experts. Goal of this focus group was getting insight in the current status of used systems concerning learning at Zuyd and the possibility of extracting data. The tooling is present to enable online activity and extracting data from systems that contain grades. At present there is no learning-related dashboard. There is a business dashboard containing information on how many students applied for Zuyd, the credits of the different cohorts and the efficiency of students per module (fail/pass ratio per cohort), per year and per study. However this system only is used on the macro and meso level of learning analytics (institute, management).

**Focus Group B1:** Students in the dashboard focus group for first year students mention several elements that are needed to help them in coping with the progress of their study. There are some aids available but these are static documents. A more dynamic way of getting feedback on their activities is one of the results of the brainstorming. The other one is the comparison of their own results and progress

with their peers. During the a co-creating part of the session students could contribute one element that they wanted to be in the learning dashboard. The students did this as a group, discussing the elements before they were added. Table 2 displays those elements in column B1. One thing the first year students asked for are functionalities that motivate students to do extra work.

**Focus Group B2:** The students that were part of the institute for more than one year had a similar but more extensive brainstorming step during their dashboard focus group. Students are more aware of the role a study coach has and reflect on that; they mention how learning analytics could help when the information provided about themselves is more personal. Students focus on guidance for themselves as individuals but also on group work. Besides guidance, students also mention valuing the individual effort in group work both in positive as well as in negative situations. In the co-creation part several elements were mentioned. They are illustrated in column B2 of Table 2.

**Focus Group B3:** The focus group of teachers on dashboards was different than the one students did. The teachers had a more diverse opinion on learning dashboards and their benefits. The question stated was if the learning dashboard should be used to get better study efficiency or just for the guidance of the students. The final element contributed during the brainstorming was to get more personal information about the student. Elements developed in the co-create part are added under B3 in Table 2.

### 3.2 Study 2: Survey

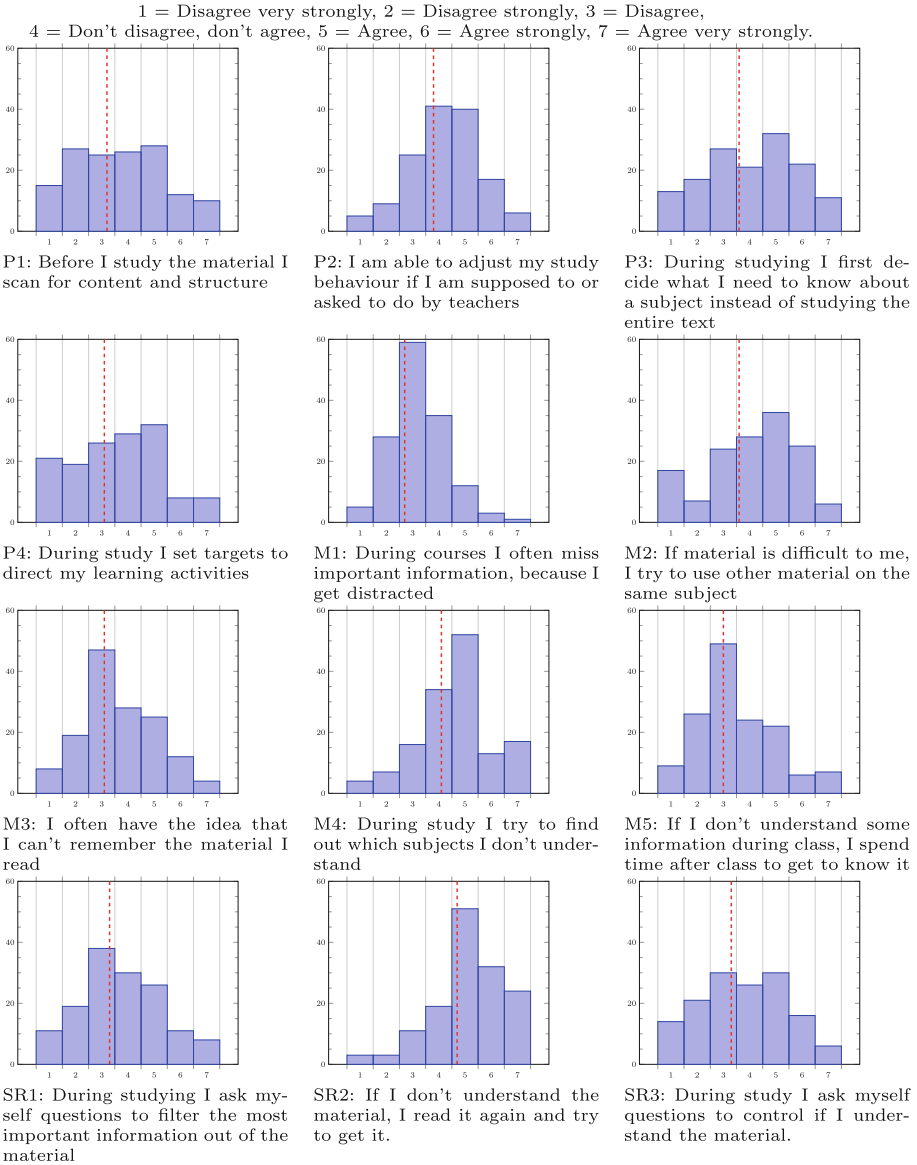
**Planning, Monitoring, Self-regulation.** Table 3 shows the results of the MSLQ’s meta-cognitive self-regulation scale survey. Cronbach’s Alpha for this part of the survey was 0.70. For question P1 we can see that 67 out of 143 students do not scan the material before studying (ratings 1–3) while 50 students do (ratings 5–7). Similarly, for question P4, 66 students do not set targets (ratings 1–3), while 48 students do (ratings 5–7). Looking at P2, however, shows that 63 students rate their ability to adjust their behaviour positively while only 39 students do not. Notable here is the fairly high number of students, i.e. 41, that are undecided. On the monitoring scales we can see that 92 out of 143 students are not easily distracted (ratings 1–3 for question M1) while only 16 students are. There are 82 students that during studying try to see what material they do not understand (score 5–7, question M4) and only 27 students do not (score 1–3). Half of the students (74/143) are sure that they can remember the things they learned (score 1–3, question M3) as 41 (score 5–7) have the idea that they can not. 67 of the 143 students say that they search for other resources if they do not understand the presented ones (score 5–7, question M2), 48 of the 143 students do not. That being said only 35 of 143 (score 5–7) say that they spend extra time trying to understand the things they do not understand (M5) and 84 students will not do so. In the answers from the self-regulation questions we see

**Table 2.** Results from focus groups A1, A2, B1, B2, B3

	A1	A2	B1	B2	B3
Info on what to prepare before a learning activity	X				
Expectations of the teacher per learning activity	X				
Which teacher is responsible for which learning activity	X				
(Comparison on) usage of learning material	X	X	X		X
(Comparison on) progress of learning activities	X		X	X	
Often made mistakes in the course	X		X		
Teacher performance	X		X		X
Quality of the learning material	X		X		X
Triggers for motivation (especially in first weeks)	X		X		
Personal, social, emotional conditions of student		X	X		X
What does a student need to pass the course		X			
Do students use competences within the course		X			
Do students participate in learning activities		X			
Performance of an individual in a group assignment		X	X	X	X
Moments (days, hours) that a student studies		X			
Activity plan of a student		X			
Own learning strategy and possible improvement				X	
(Comparison on) distance to goal credits 1st year			X		
Progress entire study			X	X	X
Learning goals (to pursue outside regular program)				X	X
Personal feedback assignments/grading (individual/group)			X	X	
Feedback on positive things in assignments				X	
Feedback connected to professional skills				X	
Peer feedback from group work				X	
Identification of possible peer assistance				X	X
(Comparison on) Test scores					X
Learning efficiency					X
Risk assessment					X
Students' existing skills before course					X
Classification and benchmark of students					X

a divers image. SR3 shows us that 52 students ask themselves question to see if they understand (score 5–7) and 65 students do not (score 1–3). SR1 shows us that questions are also not being asked by 68/143 students to retrieve the most important information. Only 45 students do. 107 out of 143 students do read material again if they do not understand it (SR2) opposed to 17 that do not.

**Table 3.** Answer distribution to the meta-cognitive self-regulation questions about planning (P), monitoring (M) and self-regulation (SR); average marked by red line.



**Perception of Collection of Data.** In Table 4 the results of the “perception of collection of data” questions are given. Cronbach’s Alpha for this part of the survey was 0.64. Students mostly have a positive mind (73.4%) on the collection of data needed to improve their learning process or the learning material.

Although the majority (58.7%) is in doubt or negative about using measurement instruments in the classroom, 66.4% is positive about sharing their online activity data. 55.2% would share data, information not anonymously to get personalised feedback through a learning dashboard which helps improving their study behaviour and study results. Students want to receive personal feedback on: their test results (69.2%), the way they learn (55.2%), their usage of learning material (46.2%) and the collaboration in group work (49%).

## 4 Discussion

The goal for our study as part of our design science process was to obtain information from the users and their context in order to determine desired tool functionalities [4]. From study 1, the focus groups, we got several insights (I1–I3). Furthermore, we derived requirements from our studies (Table 5) to build a prototype for the next phase of our design based research. Insights I1, I2 and I3 in combination with the requirements in Table 5 are thus the answers to RQ1 (What do teachers want to see displayed in their LA4LD tool?) and RQ2 (What do students want to see displayed in their LA4LD tool?). From study 2, the survey we got further insights (I4–I7), which give answers to RQ3 (What is the current state of self reflection from the students?).

**I1: Students and teachers do not make a specific distinction between learning systems, learning design and learning analytics.** Suggestions, additions and comments made by the students and teachers during focus groups often are on the current learning design or technology-enhanced learning tools to create more interactive activities instead of learning analytics.

**I2: Because students see learning analytics as an integrated part of their online learning environment, other functionalities can stimulate them using learning analytics.** Students are stimulated to use a learning analytics tool when tools that are beneficial for them and only indirectly connected to learning analytics are embedded. Students mention in the study the possibility to schedule learning activities.

**I3: Learning activities are a central element.** Both teachers as well as students show interest in information about their learning activities. One group (students) wants to see what they are supposed to do within them, how they perform, how they perform with respect to their classmates or comparable learners, when they are happening and if there are alternative learning activities to achieve the same goal. The other group (teachers) wants to get more direct input about students' behaviour to make better informed decisions on (re-)designing learning activities.

**I4: There is an opportunity for improvement in students' planning.** Large groups of students do not set targets, or scan material before a learning activity. So there is a lot of room in planning and as just a small group of students claims that they are not able/willing to change their behaviour, there is room for improvement there.

**Table 4.** Answers to the data-related questions of the survey.

Question	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I am satisfied with the insight my institution gives me regarding study progress	5.60%	17.50%	28.00%	43.40%	5.60%
I have no problems with measuring instruments in classrooms	9.80%	13.30%	35.70%	39.90%	1.40%
I want to make my data and information that I generate within an on-line environment available if I will be supported better	6.30%	8.40%	18.90%	54.50%	11.90%
I think it’s acceptable if my data, information is used to improve educational units	7.70%	7.70%	11.20%	56.60%	16.80%
<i>Question</i>	<i>Yes</i>	<i>No</i>			
I would share data not anonymously	55,2%	44,8%			
<i>Question</i>	<i>The way I learn</i>	<i>Usage of learning material</i>	<i>Tests results</i>	<i>Lectures</i>	<i>Collaboration</i>
I would like to receive personal feedback on	55.2%	46.2%	69.2%	55.2%	49.0%

**I5: The students claim to do more in less time.** We saw interesting contradictions in the monitoring questions. The group of students that states that they use other material if the presented material is too difficult is a lot bigger than the group of students that tells us that they are willing to put in extra time in their study. Further research should clarify if students are able to check other researches and use less time in the process.

**I6: Self-reflection is low.** The group of students that do not ask themselves whether they understand material and the group that ask questions to retrieve the most important information is small.

**I7: Students are prepared to share data.** The additional survey questions teach us that the majority of students are prepared to share personal online data under the condition that they can see the results and can get some input on their own process. A small majority is even willing to share this data non-anonymously. Students participating in the survey especially mention personalised feedback on their tests. Students are somewhat holding back on sharing in classroom data.



**Table 5.** Requirements for the LA4LD prototype

Requirements	Demanded by	Data source
Personal, social, emotional condition of students	Students and teachers	Ask students
Quality of learning material	students and teachers	Ask students
(Comparison of) usage -number of times-of learning material	Students and teachers	From systems
(Comparison of) usage -duration of task-of learning material	Students and teachers	From systems
Performance of individual in group assignments	Students and teachers	Ask students
Progress of study	Students and teachers	From systems
Progress of study	Students and teachers	From systems
Analysis/recommendation on study behaviour	Students and teachers	From systems
Identification of possible peer assistance	Students and teachers	From systems
Intention of learning activity	Students	Ask teacher
Estimated time to finish task	Students	Ask teacher
Own behaviour during learning activity	Students	Ask student
Quality of teacher	Students	Ask student
Alert systems group analysis	Teachers	From systems

## 5 Conclusion

In a design science research process the users are involved in every step. We thus asked students, teachers and TEL experts to participate in focus groups to define objectives and design a concept for a solution of a LA4LD tool and, in addition to that, asked students to fill in a survey on their meta-cognitive competences to get more input on the problem investigation and context part.

First, it is noteworthy that both students and teachers in their focus groups designed a more personal dashboard. Students are willing to let data be collected if they know what it is for and if it helps in improving the learning processes. Students want to give qualitative feedback on learning activities. Teachers want more personal information from students so that they are able to adjust learning activities or help students. Students mention personalisation of feedback with regards to assessments, group work and the performance of an individual in a group.

Secondly, we notice that students and teachers that were involved in our study do not care about the academic differences between learning tools and learning analytics or learning design and instructional design. Education and opportunities with analytics change rapidly, thus a data ecosystem is used as a

framework. Within that framework a class and student dashboard for teachers is designed and a study scheduler and learning activity dashboard for students.

From the students’ perspective we thirdly also notice that students in their first year differ a little from students in later years. First year students do not mention learning strategy as an element they want to see information on in their dashboard. Students that have finished their first year do. First year students mention credits, while students that are beyond that first year mention learning goals as elements that needed to be plotted.

Our next step will be to look at other research for known issues on the presented functionalities. For instance students state in this research that they want to be compared to peers while Jivet et al. [6] recommend to be very careful with social comparison in learning analytics as comparison with peers can be motivating for some students, i.e. those wanting to be at the top of the class, but de-motivating, disappointing and even stressful for others. As in our design science process the input of the user is very important, we are aware that it is not the only way to go. “Make it personal” in our opinion also means “make it sensible”.

## References

1. Beetham, H., Sharpe, R.: *Rethinking Pedagogy for a Digital Age: Designing for 21st Century Learning*. Routledge, New York (2013)
2. Dziuban, C., Graham, C.R., Moskal, P.D., Norberg, A., Sicilia, N.: Blended learning: the new normal and emerging technologies. *Int. J. Educ. Technol. High. Educ.* **15**(1), 3 (2018)
3. Florian-Gaviria, B., Glahn, C., Gesa, R.F.: A software suite for efficient use of the European qualifications framework in online and blended courses. *IEEE Trans. Learn. Technol.* **6**(3), 283–296 (2013)
4. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. *MIS Q.* **28**(1), 75–105 (2004)
5. Jivet, I., Scheffel, M., Drachslar, H., Specht, M.: Awareness is not enough: pitfalls of learning analytics dashboards in the educational practice. In: Lavoué, É., Drachslar, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017. LNCS*, vol. 10474, pp. 82–96. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_7](https://doi.org/10.1007/978-3-319-66610-5_7)
6. Jivet, I., Scheffel, M., Specht, M., Drachslar, H.: License to evaluate: preparing learning analytics dashboards for educational practice. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 31–40. ACM, New York (2018)
7. Lockyer, L., Heathcote, E., Dawson, S.: Informing pedagogical action: aligning learning analytics with learning design. *Am. Behav. Sci.* **57**(10), 1439–1459 (2013)
8. Mor, Y., Craft, B.: *Learning design: mapping the landscape* (2012)
9. Nelson, J., Campbell, C.: Evidence-informed practice in education: meanings and applications. *Educ. Res.* **59**(2), 127–135 (2017)
10. Nguyen, Q., Rienties, B., Toetel, L., Ferguson, R., Whitelock, D.: Examining the designs of computer-based assessment and its impact on student engagement, satisfaction, and pass rates. *Comput. Hum. Behav.* **76**, 703–714 (2017)

11. Park, Y., Jo, I.H.: Development of the learning analytics dashboard to support students' learning performance. *J. Univ. Comput. Sci.* **21**(1), 110–133 (2015)
12. Pintrich, P., Smith, D., Garcia, T., Wilbert, M.: A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ) (1991)
13. Rauth, I., Köppen, E., Jobst, B., Meinel, C.: Design thinking: an educational model towards creative confidence. In: Proceedings of the 1st International Conference on Design Creativity (ICDC) (2010)
14. Rienties, B., Toetenel, L.: The impact of 151 learning designs on student satisfaction and performance: social learning (analytics) matters. In: Proceedings of the 6th International Conference on Learning Analytics and Knowledge, pp. 339–343. ACM, New York (2016)
15. Roth, A., Ogrin, S., Schmitz, B.: Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educ. Assess. Eval. Account.* **28**(3), 225–250 (2016)
16. Santos, J.L., Verbert, K., Duval, E.: Empowering students to reflect on their activity with StepUp!: two case studies with engineering students. In: Proceedings of the 2nd workshop on Awareness and Reflection (ARTEL), pp. 73–86 (2012)
17. Schmitz, M., van Limbeek, E., Greller, W., Sloep, P., Drachsler, H.: Opportunities and challenges in using learning analytics in learning design. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 209–223. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_16](https://doi.org/10.1007/978-3-319-66610-5_16)
18. Schwendimann, B.A., et al.: Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Trans. Learn. Technol.* **10**(1), 30–41 (2017)
19. Van Laer, S., Elen, J.: In search of attributes that support self-regulation in blended learning environments. *Educ. Inf. Technol.* **22**(4), 1395–1454 (2017)
20. Zimmerman, B.J.: Self-regulated learning and academic achievement: an overview. *Educ. Psychol.* **25**(1), 3–17 (1990)



# Investigating the Relationships Between Online Activity, Learning Strategies and Grades to Create Learning Analytics-Supported Learning Designs

Marcel Schmitz<sup>1</sup>(✉), Maren Scheffel<sup>2</sup>, Evelien van Limbeek<sup>1</sup>,  
Nicolette van Halem<sup>3</sup>, Ilja Cornelisz<sup>3</sup>, Chris van Klaveren<sup>3</sup>,  
Roger Bemelmans<sup>1</sup>, and Hendrik Drachsler<sup>2,4,5</sup>

<sup>1</sup> Zuyd University of Applied Sciences, Heerlen, Netherlands  
{marcel.schmitz,evelien.van.limbeek,roger.bemelmans}@zuyd.nl

<sup>2</sup> Open Universiteit, Heerlen, Netherlands

{maren.scheffel,hendrik.drachsler}@ou.nl

<sup>3</sup> Vrije Universiteit, Amsterdam, Netherlands

{n.van.halem,i.cornelisz,c.p.b.j.van.klaveren}@vu.nl

<sup>4</sup> Goethe University, Frankfurt, Germany

<sup>5</sup> German Institute for International Educational Research (DIPF),  
Frankfurt, Germany

**Abstract.** Learning analytics offers the opportunity to collect, analyse and visualise feedback on learning activities using authentic data in real-time. The REFLECTOR project was used to investigate whether there are correlations between students learning strategies, their online activity and their grades. Information about the learning strategies was obtained using the Motivated Strategies for Learning Questionnaire. The grades and the online activity of students for two pilot courses was collected from the log data of the learning management system. Analysis of the collected data showed that there are moderate correlations to be found, for instance between metacognitive self-regulation, documents that are related to planning and grades. The pilot sessions taught us that there are practical issues with regards to data storage location as well as data security that need to be taken into account when learning analytics is integrated into existing learning designs. Overall, the project results show that a close relationship between learning analytics and the learning design of courses is urgently needed to make learning analytics effective.

**Keywords:** Learning analytics · Learning design  
Learning strategies · Online activity · Grades · Correlations  
Pilot study

## 1 Introduction

Learning analytics [6] is used for research, studies and applications that try to understand and support the behaviour of learners based on large sets of collected data. As introduced by Buckingham Shum [14], it can provide different

levels of insights, i.e. on the micro-, meso- and macro-levels. The micro-level addresses the needs of teachers and students and aims at a single course; the meso-level addresses a collection of courses and provides information for course managers; the macro-level takes a bird view on a directory of courses and can provide insights for a whole community by monitoring learning behaviour across courses and even across different scientific disciplines. The main opportunities for learning analytics as a domain are to unveil and contextualise so far hidden information out of the educational data and prepare it for the different stakeholders.

The current study investigates whether learning analytics can support individual learning or teaching processes on the micro-level. Teachers are able to make more evidence-based design decisions using learning analytics when running a course and students are enabled to change learning behaviour based on the insights they get to make their learning process more efficient, effective and fun [12]. Although there is a rather rich sample of learning analytics tools available, we rarely see educational concepts being used as the basis for those tools or any learning analytics indicators being embedded in a learning/instructional design as a measure point for educational interventions so that they can be used for reflection and feedback for students and teachers [15]. Also, in reviews like those by Jivet et al. [8], Schwendimann et al. [13] or Park et al. [10] many learning analytics tools are mentioned but only few of them work in real-time and none specifically cope with learning analytics-supported learning design.

Higher education institutes (HEIs) in the Netherlands had the opportunity to use the SURF Learning Analytics Dashboard (SURF-LAD) within some of their courses. The SURF-LAD gives insight in several online activities within the learning management system (LMS) of that institute. Around this SURF-LAD usage the REFLECTOR project was formed. Two institutes that were going to use the SURF-LAD participated in REFLECTOR: Vrije Universiteit Amsterdam (VU) and Zuyd University of Applied Sciences (Zuyd). The project analyses data from the students usage of online learning material, their learning strategies, and their grades and investigates whether there are correlations between these three data sets. The VU participated with one pilot course and combined the result from their SURF-LAD with those from the online practice platform IHS<sup>1</sup> and Blackboard. The online activity specifically reported the difference in used tools and a self-regulated learning model [5] is used to examine if students ability to self-regulate their learning is related with the actual learning behaviours that can be observed in the LMS [7].

The study presented here describes the two pilot courses of the faculty ICT at Zuyd, during the REFLECTOR project. Here the SURF-LAD results were used with the LMS Blackboard. We were especially interested in the connection between learning analytics and the currently available learning design. A learning design describes the development and purposeful compilation of learning activities, i.e. one interaction or a set of interactions between a student and

---

<sup>1</sup> <https://www.ihatestatistics.com/>.

student(s), teacher(s) or learning material [1]. A learning design also outlines the resources and technologies needed to support these interactions. The result of such an interaction (i.e. learning goal achievement) is also part of the learning design [3,4].

On Zuyd's side of the REFLECTOR project two pilot courses were used to retrieve data. Every pilot course started with several pre-pilot meetings between teachers and technical support to set up the learning analytics within the course. Once the courses started, students were asked to participate and to provide some information about themselves as well as their learning data to the study. The research questions that guided our analysis of the collected data were:

**RQ1:** Are there any practical challenges that need to be taken into account when using learning analytics within an existing learning design and if so which ones?

**RQ2:** Are there any significant correlations between the students' learning strategies, their online activity and their grades and if so which ones?

## 2 Methods

### 2.1 Participants and Materials

**The Pilot Courses.** The two pilot courses were conducted at faculty ICT of Zuyd. The faculty strongly supports the learning philosophy of learning-by-doing, a learning process where students learn within tasks recognisable from the professional practice. Feedup, feedback, feedforward and (self-)reflection are thus essential parts of the learning design and the courses therefore demand a high level of self-regulation from the students [9].

In its overall educational design, the faculty makes use of ten achievement indicators. For the learning design of each course three to five of these achievement indicators are chosen and formulated within the context of the course using measurable aspects per indicator as that course's specific focus. The chosen indicators can have different weights. The weighted average grade (AG) is calculated. It even is possible that an indicator is that important that a student will not pass the course if the student does not have a sufficient grade for that indicator. Therefore an overall course grade (OG) was introduced that either depends on the average grade or on an achievement indicator grade that has to be passed. The faculty ICT at Zuyd uses the tool Faculty ICT Information Engine (FICTIE) to store results of every achievement indicator from every student. Both pilot courses had a blended learning set-up, i.e. both employed face-to-face as well as online learning activities. The majority of activities were face-to-face ones that were, however, supported by documents stored in the online learning environment.

The first pilot course was a first-year bachelor degree level course on 'Communication'. This course has four achievement indicators, i.e. tasks students have to do and that are then graded: a written exam (AI1), two individual assignments (AI2 and AI3), and group work participation (AI4). The course ran from May

2017 to July 2017. 135 students –5 female, 130 male– were enrolled in the pilot course. 91 students were in their first year at Zuyd, 44 students had already been at the institute in the previous year(s). Six teachers –three female, three male– were involved in the course. Two female teachers were involved in the preparation and evaluation of the SURF-LAD that was used in this course.

The second pilot course was a bachelor degree level course on ‘Logics’. This course has three achievement indicators: a quiz (AI1), an individual assignment (AI2), and a group assignment (AI3). The course ran from September 2017 to November 2017. 177 students –14 female, 163 male– were enrolled in the pilot course. 131 students were in their first year at the institute, 46 students had already been at the institute in the previous year(s). Eight teachers –one female, seven male– were involved in the course.

**Online Activity.** The LAD provided by SURF is a teacher-facing dashboard that is meant to support teachers in their teaching processes. The dashboard is meant to raise awareness among teachers about what learning analytics and LADs can do. SURF pre-designed several possible scenarios and chose to add five of them for the REFLECTOR project. For every chosen event (e.g. click on a link, download of a file) the actions of every student are accumulated. This is displayed in several visualisations. There is a pie-chart which informs on the percentage of usage of that event for a user. Also there is a box-plot which shows the first, the last and the majority of times some type of learning material is used. There are several histograms to visualise usage of certain events. Another line graph shows how many students over time have accessed a specific event and how many students over time still had to.

Access to the SURF-LAD was embedded into the course’s LMS via a direct link. The data collected for the SURF-LAD is stored and processed using the xAPI protocol [2]. By placing indicators, e.g. an empty picture or javascript, on pages in the LMS, a data entry is made to the database whenever a page is accessed, i.e. whenever the indicator is loaded. The decision where to place the indicators was made by the teachers involved in the study. They chose those documents within the course that are of particular interest with regards to the learning design. Thus, every click on a document or menu-item was counted as one data entry. As a back-up, the LMS logs were queried for the same actions.

The documents that were selected by the teachers for further analysis were: a learning activity plan for every week (OA2); a case description for the group work of weeks 7–9 (OA3); the Modulebook with information about the course (OA4); the achievement indicator overview document (OA5); a practice quiz (OA6) and the document with the correct answers to that quiz (OA7); learning material such as articles and videos (OA8); knowledgebytes, e.g. short video clips (OA9); the presentations used during the lectures (OA10); and the attempts students do to submit assignments (OA11).

**The MSLQ.** There are several instruments to measure learning strategies [9]. The Motivated Strategies for Learning Questionnaire (MSLQ) was used as it is

a widely used, accepted and validated instrument [11]. The MSLQ consists of 81 items and is divided into fifteen sets (scales) that can be used separately. For each item, participants enter a rating from 1 for ‘totally not agree’ to 7 for ‘totally agree’. The fifteen scales are distributed among two categories: learning strategies and motivation. The learning strategy scales are: Rehearsal (M1), Elaboration (M2), Organisation (M3), Critical Thinking (M4), Metacognitive Self-regulation (M5), Time and Study Environment (M6), Effort Regulation (M7), Peer Learning (M8), and Help Seeking (M9). The motivational scales are: Intrinsic Motivation (M10), Extrinsic Motivation (M11), Task Value (M12), Control of Learning Beliefs (M13), Self-efficacy for Learning and Performance (M14) and Test Anxiety (M15). The whole questionnaire –but especially the set of nine learning strategies scales– can give insight in the students own perception of their learning strategies. In addition to the MSLQ items, some demographic information was also included in the questionnaire, i.e. age, highest educational level, gender, and study specialisation.

## 2.2 Procedure

Before the courses started, pre-pilot meetings between teachers and technical support staff took place to set up the learning analytics that was to be used in each course. In pilot course 1, two teachers were asked to regularly evaluate the SURF-LAD throughout the course. The teachers received an introduction to the dashboard at the beginning of the course and were later contacted again to provide their evaluations. There were no specific questions for the evaluation. Teachers were asked to report on their personal impression.

During the the first week of pilot course 1, all enrolled students were invited to participate in the study by mail. During the second lecture the research project was presented in class and students were reminded of the invitation to participate. In the LMS there was a link during the entire course called ‘Experiment’. By clicking the link students were presented information about the experiment and a button to give consent on storing, analysing and visualising their learning data for the study. The invitation to fill in the MSLQ was sent to students in the third week of the course by mail. The questionnaire contained an informed consent form where students could agree or disagree with the usage of their questionnaire answers and of their achievement indicator grades for the study. In week four students were reminded in class to fill in the questionnaire. It was distributed using Qualtrics<sup>2</sup>.

For pilot course 2, an invitation to participate in the study by filling in the MSLQ and by agreeing to the collection and analysis of the online activity as well as of the achievement indicator grades, was sent to all students by mail in the first week. In week 3 the research project was promoted by the teachers in class. A personalised mail was sent to the students in week five to remind them of the study and the questionnaire. The questionnaire contained an informed consent form where students could agree or disagree to the collection and analysis of the

<sup>2</sup> <https://www.qualtrics.com/>.



questionnaire data, the online activity data and the achievement indicator data. The questionnaire for the second pilot course was distributed using Questback<sup>3</sup>.

For both courses, the data from those students who gave their consent was exported from FICTIE. Only those achievement indicators used in the two courses were used. Answers to the two MSLQ runs were processed and the questionnaire results were calculated according to the MSLQ guidelines. For every scale the average of the items belonging to that scale were calculated. With regards to the data collected from the LMS, for each of the elements we stored the daily online activity per person, we calculated the accumulated online activity for that element per person and for all participants per course.

A Pearsons Correlation Matrix was used to compare the scores of the 15 MSLQ scales (M1-M15) with the students' eleven online activities (OA1-OA11), the MSLQ scores with the four achievement indicator grades (AI1-AI3 and OG), and finally, the online activities with the achievement indicator grades as well. The correlation coefficients were calculated to determine the strength of association between the different factors as well as the significance level. In order to examine the three sets of data further, a  $31 \times 31$  scatter plot matrix was created for all elements. The matrix was then visually checked by three members of the research team. IBMs SPSS Statistics 24<sup>4</sup> was used for calculating the correlation and scatter plot matrices.

### 3 Results

#### 3.1 Pilot Course 1: May 2017–July 2017

A few issues occurred during preparation and execution of the first pilot course. Some were related to the SURF-LAD, others to Zuyd's LMS. The first issue already occurred during the set-up of the SURF-LAD. Zuyd runs a local installation of Blackboard and due to security settings on the server SURF's preferred option of tracking the use of online resources with javascript was not possible. Empty pixels were used instead. This, however, also turned out to not fit all scenarios as one factor chosen by the teachers to be of interest was the weekly usage of several resources (i.e. presentations used in lectures). Faculty ICT, though, chose to combine all presentations of lectures into one document. Within this document there was an interactive menu to easily navigate through the content. The students' interaction within the document once they downloaded it could of course not be tracked and thus no xAPI statements could be generated in order to feed the visualisation of the SURF-LAD. Therefore, a specific measurement per presentation and lecture was not possible.

Another issue was that faculty ICT has set up their educational logistics in such a way that all content is stored in one part of the LMS while the learning analytics tools only worked on another part of the LMS. For the first pilot course

<sup>3</sup> <https://www.questback.com>.

<sup>4</sup> <https://www.ibm.com/products/spss-statistics>.

a work-around was thus created by redesigning the educational logistics specifically for this course. A third issue occurred with embedding the functionality of opt-in / opt-out of the study's data collection for the students due to local security settings on the Blackboard server. Another work-around was created by embedding the SURF-LAD as a website via an iframe. Due to these issues it was decided to also collect the required online activity by querying the Blackboard database as a backup.

Soon after pilot course 1 started and students had been told about REFLECTOR, however, it became clear that the second work-around had its own limitations. Students at the faculty ICT log in with their own device on a closed network. Because of security settings in the network the SURF-LAD system did not store the permission status of the student. This led to students being asked to opt-in to the data collection every time they accessed the online course and having to click through a number of items in order to give permission. In addition to this, both the SURF-LAD as well as the Blackboard installation at Zuyd suffered from technical issues. The LMS, for example, was down for an entire week.

During pilot course 1 the SURF-LAD was configured, used and evaluated by two teachers. The teachers were impressed with the ability to get insights in the usage of learning material. In the setup of the SURF-LAD insight could be given on how much and when material was used by a group of students, and what the percentage of usage was per anonymised user. The SURF-LAD had no specific student dashboard. The dashboard for the teachers did not have the possibility to track an individual student's usage. The teachers recommended this as an addition. Eventually, only the activity of 16 students was collected and visualised to the teachers in the SURF-LAD. Only two of those students filled in the MSLQ and made their grades available. We thus chose not to perform any analysis on this small sample size.

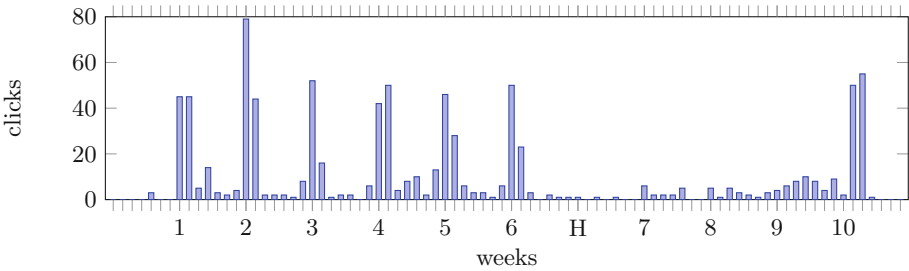
### 3.2 Pilot Course 2: September 2017–November 2017

The SURF-LAD environment was not used within this pilot course as the technical issues encountered in pilot course 1 could not be addressed in time. Online activity was measured by using the activity logs from Blackboard. The queries used on the Blackboard database provided the same information as the SURF-LAD tool did in pilot course 1. There were 52 students that filled in the MSLQ, seven of them did not agree to their data being used for the study when filling in the informed consent form. We were thus able to use the MSLQ results, the online activity and the achievement indicator grades from 45 students –1 female and 44 male– aged on average 20.13 years. Table 1 shows the descriptive statistics of the MSLQ results.

Figure 1 shows an overview of the online activity related to the presentation document used in the course. We distinguish the following sections: In the first six weeks students participated in lectures and did some assignments. Then there was a holiday week. After that there were three weeks (7–9) to conduct a group

**Table 1.** Descriptive statistics of results for the 15 MSLQ scales in pilot course 2

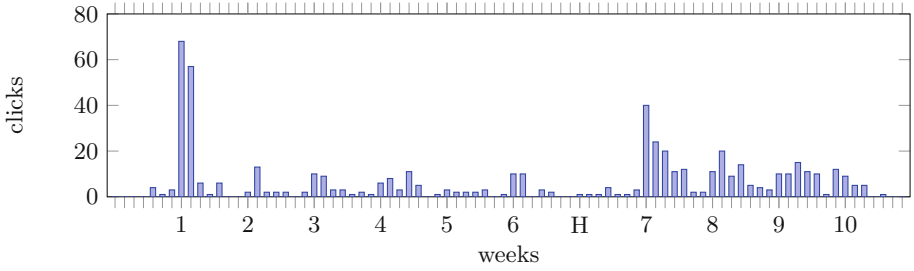
		N	Min.	Max.	Mean	Std.Dev.
Rehearsal	M1	44	1.00	6.50	4.74	1.09
Elaboration	M2	44	3.33	6.33	4.87	0.85
Organization	M3	43	1.00	6.25	4.26	1.15
Critical Thinking	M4	45	2.00	6.40	4.03	1.07
Metacognitive Self-regulation	M5	42	1.92	5.58	4.17	0.88
Time and Study Environment	M6	44	2.63	6.38	4.60	0.82
Effort Regulation	M7	44	2.25	7.00	4.85	1.02
Peer Learning	M8	44	1.33	6.67	4.33	1.20
Help Seeking	M9	42	2.25	6.75	4.86	1.03
Intrinsic Goal Orientation	M10	44	3.75	6.75	5.41	0.66
Extrinsic Goal Orientation	M11	44	1.00	7.00	4.83	1.10
Task Value	M12	45	3.67	6.83	5.54	0.71
Control of Learning Beliefs	M13	44	4.00	6.75	5.52	0.62
Self-Efficacy for Learn. & Perf.	M14	45	2.50	6.88	4.99	0.97
Test Anxiety	M15	45	1.80	7.00	4.00	1.36



**Fig. 1.** Usage (y-axis) of the presentations document throughout the weeks (x-axis) of the course; H = holiday week

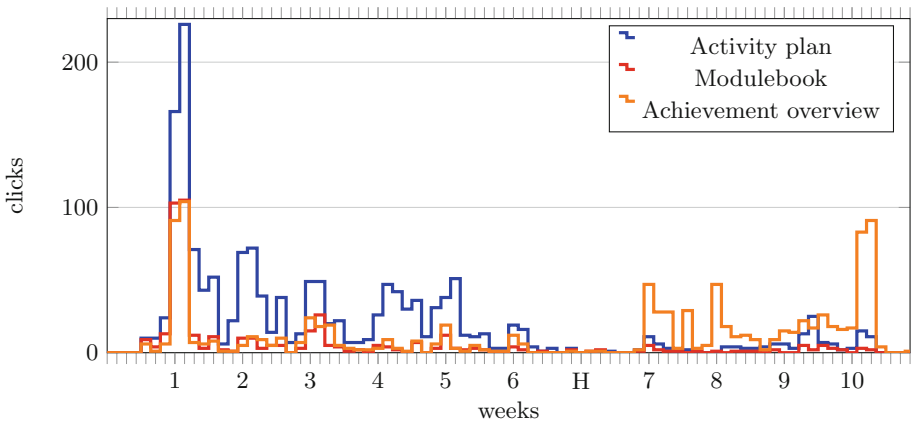
assignment. In the closing week (10) assignments had to be submitted and a final quiz was done on the 15th of November.

In Fig. 2 we see the usage of the case description document. In the beginning of the course (first days) there was a higher amount of students that wanted to know what the groupwork (i.e. the case) in weeks 7–9 is about. The three weeks when students were supposed to work on the case had higher online activity values. In the overview of planning documents (activity plan, modulebook and achievement plan) shown in Fig. 3 we can see a big spike for the activity plan in the first days of the course, especially with respect to the usage of it in the



**Fig. 2.** Usage (y-axis) of the case description document throughout the weeks (x-axis) of the course; H = holiday week

remaining weeks of the six week period. It might be the case that the activity plan is downloaded in the first week and then used on a local computer or copied into a personal agenda. We do not have the instruments at the moment to account for this. Also interesting is the second boost of usage of the achievement plan because it is almost as big in the three weeks of the case as it is in the first week of the course. The reason could be that deadlines for delivering assignments are planned in the beginning of week 7 and week 10.



**Fig. 3.** Usage (y-axis) of the activity plan, modulebook, and achievement overview documents throughout weeks (x-axis) of the course; H = holiday week

Table 2 shows the result of the Pearson correlation matrix used to investigate if there are any correlations between learning strategies and online activity. It needs to be noted that all of these correlations are just that: correlations. They are not to be seen as predictive. There are several significant correlations: The accumulated online activity (OA1) of a student and the activity plan (OA2) moderately negative correlate with test anxiety (M15). The usage of the modulebook (OA4) correlates moderately negatively with the scales intrinsic goal

**Table 2.** Correlation learning strategies (M1–M15) - online activity (OA1–OA11)

	OA1	OA2	OA3	OA4	OA5	OA6	OA7	OA8	OA9	OA10	OA11
M1	.142	.094	.040	.107	.189	-.181	.149	.153	.185	.279	-.375*
M2	.082	.120	.141	-.017	.011	-.301*	.115	-.003	.045	.151	-.339*
M3	.055	.006	.033	-.018	.104	-.480**	.121	.124	.164	.274	-.361*
M4	-.041	-.012	.124	.017	-.066	-.105	.181	-.154	-.078	-.063	-.117
M5	.185	.147	.201	.087	.209	-.427**	-.023	.177	.045	.297	-.430**
M6	.177	.194	.075	-.005	.096	-.154	.157	.121	-.043	.267	-.416**
M7	.199	.151	.149	-.251	.135	-.088	.259	.222	.143	.293	-.341*
M8	.011	.055	.075	-.007	-.016	-.186	.058	-.025	-.122	.062	-.113
M9	.122	.136	.097	-.263	.008	-.057	-.124	.145	.056	.014	.017
M10	-.195	-.171	-.156	-.310*	-.103	.077	.203	-.182	.036	-.032	-.099
M11	-.138	-.147	-.007	-.228	-.074	-.105	.200	-.165	-.083	-.055	-.237
M12	-.078	-.121	-.055	-.325*	.041	.207	.177	-.011	.171	.031	.226
M13	.040	.064	-.068	-.187	.182	-.043	.017	-.012	.107	-.110	.228
M14	.016	.130	-.133	-.374*	-.065	.356*	.168	-.096	-.030	-.008	.002
M15	-.298*	-.308*	-.040	.151	-.234	-.068	.000	-.256	-.161	-.213	.127

The correlations marked with \* have a significance at the 0.05 level, those marked with \*\* have a significance at the 0.01 level.

orientation (M10), task value (M12) and self-efficacy for learning performance (M14). The practice quiz (OA6) moderately negatively correlates with elaboration (M2), organisation (M3) and metacognitive self-regulation (M5). And the amount of attempts to send in assignments and portfolio material (OA11) moderately negatively correlates with rehearsal (M1), elaboration (M2), organisation (M3), metacognitive self-regulation (M5), time and study environment (M6) and effort regulation (M7). All these are negative correlations which means the higher the students perception of their learning strategy/motivation, the lower their usage of the online document. There is only one significant positive moderate correlation and that is between the practice quiz (OA6) and self-efficacy for learning and performance (M14). Thus, students that rank their self-efficacy for learning as high, tend to use the practice test often.

Table 3-(a) shows the results of the Pearson correlation calculation between learning strategies and grades. There were again several significant correlations: The metacognitive self-regulation scale (M5) had a negative moderate correlation with the grade of the written exam (A11), the grading from portfolio of the case (A13), the weighted average (AG) and the final grades (OG). Help seeking (M9) and Task value (M12) had a moderate positive correlation with AI2 (grading of the assignments during the first six weeks), Task value (M12) also had a positive moderate correlation on the final grade (AG). And the Self Efficacy for Learning Performance scale (M14) had a positive moderate correlation to AI1. To see if there is a relation between online activity and the grades, another set of Pearson correlation coefficients was calculated. The results are shown in Table 3-(b). There is a positive moderate correlation between the practice test (OA6), AI1 and the final grade (OG). There is a significant positive correlation

**Table 3.** Correlation between learning strategies and grades (a) and between online activity and grades (b)

	AI1	AI2	AI3	AG	OG		AI1	AI2	AI3	AG	OG
M1	-.190	.025	-.121	-.124	-.177	OA1	.102	.196	.135	.172	.204
M2	-.066	.033	-.065	-.048	-.112	OA2	.108	.113	.064	.109	.155
M3	-.074	.142	-.103	-.034	-.037	OA3	.084	.106	.128	.132	.134
M4	-.196	-.178	-.170	-.217	-.187	OA4	-.012	.002	.163	.084	.084
M5	-.394**	-.186	-.353*	-.385*	-.401**	OA5	.063	.255	.166	.194	.241
M6	-.047	-.030	-.128	-.094	-.122	OA6	.318*	.174	.166	.257	.320*
M7	.063	.091	-.011	.046	.019	OA7	.369*	.241	.048	.233	.269
M8	-.166	.199	-.148	-.072	-.158	OA8	.027	.233	.156	.169	.170
M9	.040	.307*	.024	.127	.057	OA9	-.066	.192	.144	.117	.123
M10	.094	.132	-.008	.071	.152	OA10	-.021	.169	-.016	.039	.095
M11	.025	-.019	.020	.013	.026	OA11	.213	.302*	.221	.291	.371*
M12	.237	.447**	.182	.325*	.364*						
M13	-.011	.256	.108	.138	.102						
M14	.343*	.231	.022	.207	.210						
M15	-.124	-.246	-.117	-.186	-.162						

(a)

The correlations with \* have a significance at the 0.05 level, those with \*\* have a significance at the 0,01 level.

(b)

**Table 4.** Planning related information

	n	AG	Avg.OA2	Avg.OA4	Avg.OA5
Total	45	6.8	30.4	2.9	6.8
M5 > 4.5	17	6.1	33.7	2.8	7.7
M5 < 3.5	11	7.8	29.1	2.1	5

between the amount of attempts to post material (OA11) and AI2 and the final grade (OG). And we see a significant moderate correlation between the solution of the practice test (OA7) and the written exam (AI1).

A relationship that draws attention is the moderate negative correlation between metacognitive self-regulation (M5) and most grades (AI1, AI3,AG, OG). This means that students that score high on that scale have low grades, and students that have low grades, score high on that scale. Table 4 shows that students that score high on their metacognitive self-regulation scale (M5) have a higher average usage of all planning documents. The low scoring metacognitive self-regulation scale (M5) students have a lower average usage of the planning documents.

To search even further for relationships the data mining technique of making a scatter plot matrix was used. Three researches did a visual search on the 31 × 31 matrix. Every cell is a scatterplot from two of the variables from MSLQ,

achievement indicators and online activity documents. Every cell was looked at. From the scatterplot matrix no leads for further investigations were found.

## 4 Discussion

While preparing and running the pilot courses we saw some practical issues. In order to answer the research questions posed at the beginning of this study, we have compiled several recommendations and lessons learned. Even though not all of them are to be seen as new to the research community in general, we compile them here as an overall output from what was encountered at faculty ICT of Zuyd as they most likely will also apply to many other institution. Recommendations R1–R4 are presented for future experiments when using learning analytics in existing set ups of learning design, educational logistics and security of servers and networks. The second pilot course and analysis from the data led to insights in the learning design of faculty ICT. Lessons learned L1–L5 are defined based on that.

**R1: Learning design should have elements that can be measured.** At faculty ICT there is a distinction between several achievement indicators and the aspects with which the indicators can be graded. There also is a clear connection between the learning activities and the achievement indicators. This provides a measurable learning design.

**R2: Take measurement of efficiency and effectiveness of learning in consideration while connecting learning activities and achievement indicators.** In the design phase of the pilot courses there were connections made between learning activities and achievement indicators but the efficiency and effectiveness of learning and how it can be measured was not taken into consideration at design time. Doing this may improve the indicators and thereby better learning analytics for learning design.

**R3: Store learning material in a way it can be measured.** A very specific issue we encountered is the way that learning material was stored in the LMS for our courses. This was problematic because the tool used to collect and visualise the learning data did not work due to the originally envisioned method of collection and storing. Location, security settings on the server level and security settings on the network level have presented themselves as problematic during the REFLECTOR project.

**R4: Further investigate if and how students want to share learning data from their own devices.** Looking at the activity in Figs. 2 and 3 we see almost double the amount of activity in the first days. The reason for this could be that part of the students download the learning material on the first day onto their own device and then never go to that specific material in the LMS again. To be sure that this is the case more information is needed either on what is downloaded or what is used on the device of a student. Questions to answer are how this can be done and under which conditions students are prepared/willing to do this?

- L1: Students do not prepare for lectures.** It is by design that at faculty ICT all presentations are made available to students before a lecture in order for students to be able to prepare themselves for the lecture. Table 1 shows that the presentations are most used on the day of the lecture.
- L2: Students use the presentations most during the lecture.** Our analysis on the usage of the presentations shows that the majority of usage is during the lecture. Students use their laptop during the course to look at the presentation on their screen while the teacher is presenting it on the stage. The amount of usage of the documents before the lecture is minimal.
- L3: Practice test and solution are hardly used.** The two learning activities of taking an example quiz and reviewing the example quiz are designed in order for students to be prepared optimally for the quiz in the last week of the course. Usage, however, is minimal, just one or two students in our sample group made use of the test.
- L4: There is a negative moderate correlation between metacognitive self-regulation and grades.** Interesting to see is that the group of students that score “high” (>4.5) on the metacognitive self-regulation scale (M5) have a lower average grade (AG) (Table 4). This observation is in line with the moderate negative correlation of this scale with the grades from Table 3. When we look at how online material is used, then we see that this is in line with the learning strategy scale value. Further research is needed to see whether the learning material used has to be improved or whether these students have too high an esteem of their self-regulating capabilities.
- L5: Significant correlations can be found.** The example of the negative moderate correlation between metacognitive self-regulation (M5) and grades (AI1, AI3, AG, OG) shows us that significant correlations can be found, but more specific questioning and research is needed. More potential relationships can be searched this way, but specific research questions or hypotheses are needed.

Overall, statistically there were moderate relationships to be found between learning strategies, online activity and grades. It is interesting to further explore –with more data than the population of 45 we had now– if relationships based on choices in the learning design between learning strategies, online activity and grades exists. It will also be interesting to see how the addition from self-reports from students activity will define those relationships. More specified questions based on the learning design and the population are needed to get a clearer view on the relationships.

## 5 Conclusion

This paper describes our experiences of using learning analytics during two courses at a HEI. Data about online activity, students perception about their learning strategy based on the MSLQ and their grades were collected and analysed from a learning analytics-supported learning design perspective.



We observed that the chosen HEI has a learning design that has potential to be supported by learning analytics. Also, we observed that practical and technical issues still have to be resolved to get a big enough data set. From the relatively small dataset now we can already see the potential of statistical analysis. More specific questions such as “Do students with a high score on the rehearsal scale benefit from using the practice test often” or “Can we see the group work achievement indicator grade rise when students with a low score on peer learning read the collaboration article” rather than simply checking for correlations between certain factors can then be taken into account as well in order to investigate if valuable information for changing the behaviour of students in their learning processes or for improving the quality of learning activities by teachers can be obtained. Also, further statistical analyses like structural equation modelling to learn something about predictive relations between the observed factors will be interesting. The REFLECTOR project has shown us that learning analytics should be a talking point while designing learning activities and when deciding how learning material is supplied to students.

## References

1. Beetham, H., Sharpe, R.: *Rethinking Pedagogy For a Digital Age: Designing for 21st Century Learning*. Routledge, New York (2013)
2. Berg, A., Scheffel, M., Drachslers, H., Ternier, S., Specht, M.: The Dutch xAPI experience. In: *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, pp. 544–545. ACM (2016)
3. Celik, D., Magoulas, G.D.: Approaches to design for learning. In: Chiu, D.K.W., Marenzi, I., Nanni, U., Spaniol, M., Temperini, M. (eds.) *ICWL 2016*. LNCS, vol. 10013, pp. 14–19. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-47440-3\\_2](https://doi.org/10.1007/978-3-319-47440-3_2)
4. Cross, S., Conole, G., Clark, P., Brasher, A., Weller, M.: Mapping a landscape of learning design: identifying key trends in current practice at the open university. In: *European LAMS Conference*, pp. 98–103 (2008)
5. Efklides, A.: Interactions of metacognition with motivation and affect in self-regulated learning: the MASRL model. *Educ. Psychol.* **46**(1), 6–25 (2011)
6. Greller, W., Drachslers, H.: Translating learning into numbers: a generic framework for learning analytics. *J. Educ. Technol. Soc.* **15**(3), 42–57 (2012)
7. van Halem, N., Schmitz, M., Drachslers, H., Cornelisz, I., van Klaveren, C.: Tracking patterns in self-regulated learning behavior in online learning environments: a case study. (to be submitted)
8. Jivet, I., Scheffel, M., Drachslers, H., Specht, M.: Awareness is not enough: pitfalls of learning analytics dashboards in the educational practice. In: Lavoué, É., Drachslers, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017*. LNCS, vol. 10474, pp. 82–96. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_7](https://doi.org/10.1007/978-3-319-66610-5_7)
9. Panadero, E.: A review of self-regulated learning: six models and four directions for research. *Front. Psychol.* **8**, 422 (2017)
10. Park, Y., Jo, I.H.: Development of the learning analytics dashboard to support students’ learning performance. *J. UCS* **21**(1), 110–133 (2015)

11. Roth, A., Ogrin, S., Schmitz, B.: Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educ. Assess. Eval. Account.* **28**(3), 225–250 (2016)
12. Schmitz, M., van Limbeek, E., Greller, W., Sloep, P., Drachler, H.: Opportunities and challenges in using learning analytics in learning design. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017. LNCS*, vol. 10474, pp. 209–223. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_16](https://doi.org/10.1007/978-3-319-66610-5_16)
13. Schwendimann, B.A., et al.: Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Trans. Learn. Technol.* **10**(1), 30–41 (2017)
14. Shum, S.B., Knight, S., Littleton, K.: Learning analytics. In: UNESCO Institute for Information Technologies in Education. Policy Brief (2012)
15. Wise, A.F., Shaffer, D.W.: Why theory matters more than ever in the age of big data. *J. Learn. Anal.* **2**(2), 5–13 (2015)



# Evidence for Programming Strategies in University Coding Exercises

Kshitij Sharma<sup>(✉)</sup>, Katerina Mangaroska, Halvard Trætteberg,  
Serena Lee-Cultura, and Michail Giannakos

Norwegian University of Science and Technology, Trondheim, Norway  
{kshitij.sharma,katerina.mangaroska,hal,serena.leecultura,  
michailg}@ntnu.no

**Abstract.** Success in coding exercises is deeply related to the strategy employed by the students to solve coding tasks. In this contribution, we analyze the programming assignments of 600 students from an introductory university course in object-oriented programming. The students were provided unit tests for the assessment of their code, and their editing and testing actions were recorded using an Eclipse plug-in. The primary motivation for this study is to discover the programming strategies used by students for coding exercises with different difficulty levels, and find out if any relation exists between these strategies and the success in solving the coding tasks. More insights into this process will enable educators to provide future students timely, appropriate and constructive feedback on their coding process. Thus, to predict success in the coding exercises, we used indicators from students' testing behaviour reflecting the time and effort differences between two successive unit test runs. The results show a clear difference in the strategies employed by students within different success levels. The results also highlight ways of providing actionable feedback to the students in a timely and appropriate manner.

**Keywords:** Programming strategies · Personalized feedback  
Computer science education

## 1 Introduction

Programming involves the process of generating a solution to a problem, thus one of the main learning outcomes of a programming course is to develop a student's ability to solve problems [31]. Therefore, it is important for educators to be responsive to “the problem-solving skills students bring to programming, and to those required by programming” because students are influenced by the facilitated strategies [33]. Soloway et al. managed to show that students' sensitivity to strategies while learning to program has significant effect on their performance [33]. However, first year students have a small skill set and the ability to read code [22]. Therefore, besides choosing the most appropriate programming approach,

programming environment and tools, the educators should consider conveying and teaching problem-solving strategies (e.g. hill climbing, trial and error, divide and conquer, top down, and bottom up) that students could exploit and apply while learning coding [2]. In addition, Felder says, that students “should be given the freedom to devise their own methods of solving problems rather than being forced to adopt the teacher’s strategy” (p. 679) [16]. But all strategies are not equally good, thus students need feedback from educators in order to learn and improve. Moreover, the strategies that students employ to solve coding problems cannot be observed directly and must be inferred. Therefore, this study aims to analyze the program assignments of 600 students from an introductory Java university course. Consequently, we aim to investigate the programming behavior of freshmen while learning how to program, by utilizing data generated when solving their programming assignments. This allows us to ascertain the strategies students employ during coding activities and understand the efficiency of these different strategies, so educators can offer actionable feedback to nurture good programming habits and strategies [4]. Enhancing the learning experience of students with carefully designed coding exercises and support in assessing the required knowledge, should assist freshmen when faced with the difficulties of syntax and semantics, as well as understand error messages and control flow.

To capture students’ programming behavior and identify their strategies, the authors extended the Eclipse programming tool with a plug-in for data collection. The goal of this study is to identify successful students’ programming strategies. This will allow educators to provide meaningful personalized feedback promoting reflection and support, allowing students to improve the way they program. Consequently, the study addresses the following research questions:

**RQ1:** What programming strategies do freshmen employ to succeed in their assignments?

**RQ2:** Which actions can predict students’ programming behavior and support educators in early detection of difficulties and misconceptions?

## 2 Related Work

Previous research has shown a multitude of individual factors influencing academic achievement at various educational levels (e.g. primary, secondary, university). Some of these factors include self-efficacy [14,35], personality traits (e.g. conscientiousness) [3,28], cognitive ability [6], prior knowledge and experience [14,35], and motivational and strategic (e.g. learning strategies) aspects [30].

Consciousness has been shown to be the personality trait that is most influential on academic achievement according to past studies [3,8,13,28]. Moreover it is the dimension most closely linked to the will to achieve [13]. Another key predictor of student learning and academic performance is self-regulated learning (SRL) [11,12,23,27]. SRL leads to deep cognitive engagement with the learning resources [11] which in turn transitions the extrinsic motivational behavior to behavior that is driven by intrinsic motivation [12]. This path from deep cognitive engagement to high levels of intrinsic motivation was found to be correlated

with student learning and academic achievement [40]. Another behavioral factor correlated with student learning (e.g. mastering the content) and academic achievement is performance approach [14] or deep strategy [30]. Deep learning strategies (when the student's focus is to attain understanding of the content and not merely obtaining a higher grade) result in mastering the content [14] which may lead to higher examination success [30]. In past studies, researchers show the difference between strategies (deep vs. surface) and their relation to academic achievement, and concluded that deep and surface strategies were positively and negatively correlated with academic achievement [7], respectively. Finally, previous research has shown that intellectual (cognitive/mental) ability influences academic performance. Intellectual abilities can be measured in different ways such as IQ [1], general mental ability (American College Test scores) [35] and logical reasoning [9]. Although several different factors can influence student academic achievement, when it comes to programming, problem solving ability demonstrates the most significant correlation with student performance in solving coding tasks [21]. In this contribution we will focus on the behaviour of the students rather than the above mentioned constructs. These previous contribution are to give reader a brief summary of which factors affect the academic achievement.

In computer science education, student assessment still abides by traditional outcome-based assessment [10]. However, programming is more than just the capability to generate code. It is a problem solving skill. Past research has shown that this assumption has been neglected, leading to a gap in students' ability to apply core programming concepts to real-world problems [32,37]. To address this issue, educators must be able to guide students in determining correct strategy, and identifying the appropriate time to abandon an inefficient approach [17]. Thus, researchers need to collect more authentic data and explore the processes by which students arrive at their final solutions [34]. This idea has become reality with the increase in popularity and usage of automated code testing and assessment in computer science education. Existent systems aid educators in assessing various features of coding assignments and scale the assessment up for large courses [15]. For instance, Jadud introduced the idea of researching students' compilation behaviour (i.e. "the programming behaviour students engage in while repeatedly editing and compiling their programs"), to better understand how students progress through a programming task, so that appropriate interventions can be applied [19]. Following this idea, Blikstein et al. utilized code snapshots to uncover differences between novices and experts' programming strategies [4]. Expanding on these past research studies, we extended the Eclipse tool to collect data portraying students' programming behaviour; with a goal to explore students strategies when solving coding tasks and their success in doing so.

Feedback is one of the most powerful variables influencing learning [18]. However, feedback is of little use if it only conveys a message of right or wrong. Feedback must be meaningful and actionable in order to help the learning process. Traditionally, in computer science education, students receive basic level

of feedback presented by the compiler [29]. Compiler messages are not always helpful, as they do not allow students to understand why they fail in solving the coding task. In most cases, coding tasks have multiple ways of achieving multiple solutions. To complete programming tasks, students apply strategies that build on their previous knowledge [20]. This led researchers to categorize students based on their programming behavior and employed strategies. Perkins et al. classify novice programmers as “stoppers” and “movers” based on the strategy they choose when facing a problem [25]. Turkle and Papert proposed two categories, “tinkerers” and “planners” [36], while Bruce et al. identified five: “followers”, “coders”, “understanders”, “problem solvers”, and “participants” [5]. Turkle and Papert’s idea was not only related to categorizing the novice programmers, but also conveying epistemological pluralism. Epistemological pluralism highlights that students can have separate approaches to the same problem and communicate different behavior (e.g. “tinkerer” or “planner”) while achieving similar results. Consequently, educators recognized the importance of the students learning process when learning how to program, and developed tools and systems to support this progress [24, 29, 39]. This study contributes to a data-driven development of personalized feedback in programming by using the writing and testing behavioral indicators of the students as they attempt to solve coding exercises. Our aim for this contribution is to keep the behavioral indicators as semantic-less as possible to attain greater generalizability and reproducibility of results.

### 3 Methodology

#### 3.1 Research Objectives

The context of this research is a compulsory course in object-oriented programming (OOP). This course is offered to second semester CS-majors (600 students) in Java. As an introductory to OOP, there is a substantial variation in motivation and skills. This course is the basis for later software development courses, thus, it is important to identify struggling students early, provide appropriate feedback and help them develop good strategies for solving programming problems. Hence, the goal of the research is twofold: (1) identify programming strategies that lead to success in solving coding exercises; and (2) find ways to quickly detect student difficulties and misconceptions.

#### 3.2 Assignment Structure

The course has 10 assignments with a reward of 100 points for completing each successfully. A student needs 750 points to qualify for the exam. Seven of the assignments (1-3, 5-6 and 8-9) are composed of smaller coding exercises with specific requirements indicating what to code. This allows us to use unit tests for automatic grading, as well as collect rich data regarding student progression. Students are encouraged to test by writing and launching their own testing

code. Due to the open nature of the remaining assignments (4, 7, 10), they have been excluded from this part of the study. The size (number of Java classes and methods) and difficulty level of exercises vary; thus, the students are granted a certain degree of freedom in selecting exercises based on their (self-assessed) skill level. Statistics indicate that exercise choice is evenly spread. As well, exercises use approximately the same amount of time each week.

### 3.3 Data Collection

We focus our data collection to the last 4 assignments, as the first three assignments were relatively basic for students to develop concrete strategy. For each of these exercises we provided Eclipse with detailed instructions about which files and activities to track. In particular, we collected the following data: (1) snapshots of files when they are saved, with compiler errors and warnings (2) student programs that are launched, typically for testing their own code (3) unit tests that are run, with information as to whether they pass or fail, and (4) the use of certain commands and panels, typically those used for debugging

All data is time-stamped and most are limited to the relevant files of a specific exercise, for both practical and privacy reasons. A special “Exercise panel” shows the details of which data has been collected, allowing the students to track their progress and review their process. The data is anonymized, but with identifiers corresponding to exam result, prior to its use in our research such that it can be correlated at a later stage.

### 3.4 Measurements

To analyze the behavior and predict the outcome of each assignment, we captured the following measures:

1. **Number of test runs:** is the total number of times a student ran the unit tests to check their code. This is counted for each exercise in every assignment.
2. **Improvement in unit test success:** each time a student ran the unit tests, they passed and/or failed a specific number of tests. The score they obtained is the number of passed tests divided by the total number of tests. As a result, the authors computed the improvement (or lack thereof) in this score between two consecutive test runs.

To predict and analyze a student’s programming behavior in terms of the above mentioned measures, the authors also computed the following variables from the student’s unit test running time series:

1. **Time difference launch:** is the average time difference between two consecutive launches of their own test code, before the students runs another unit test.
2. **Time difference edit:** is the average time difference between two consecutive logs of saving the file(s).

3. **Size difference:** is the difference in the number of lines of code between two consecutive unit test runs, i.e. code growth.
4. **Improvement in errors:** is the reduction in number of errors and warnings between two consecutive unit test runs.
5. **First test run score:** is the unit test success score of the first time a student ran a unit test for each exercise in every assignment.

## 4 Results

In this section, we present the prediction results followed by the behavioral analysis based on student categorization using an explanatory model.

**Prediction Results.** To predict the dependent variables: (1) improvement in unit test success and (2) the number of test runs, we used four different independent (also termed predictor) variables: (1) time difference launch, (2) time difference edit, (3) size difference, and (4) improvement in errors fitting a Generalized Additive Model (GAM). We divided the data set into 80% training and 20% testing set. We performed 5-fold cross-validation for both the training and testing. On one side, considering the improvements in the unit test success, in Table 1 we can see that the overall prediction error using the combined data of the four assignments is 0.11; and the average prediction error using data from each assignment separately is 0.18 (SD = 0.03). On the other side, in the same table, considering the number of test runs, we can see that the overall prediction error is 0.18 and the average prediction is 0.24 (SD = 0.04). Table 2 show the coefficients of the explanatory variables.

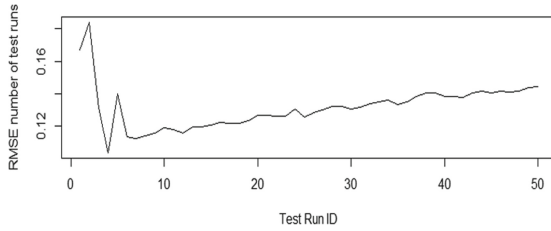
**Table 1.** Prediction results for the final score in a given assignment and the total number of test runs using data from individual assignments and the complete data sets.

Assigment ID	5	6	8	9	Overall
RMSE improvement in score	0.13	0.20	0.20	0.18	0.11
RMSE number of attempts	0.21	0.26	0.21	0.28	0.18

Relative to the number of test runs per individual assignment, we explore the question *how early can we predict?* Figure 1 demonstrates Root Mean Square Error (RMSE) of 0.10 from as early as the fourth test run. We can see that most of RMSE values are between 0.12 and 0.16, however the lowest value is observed at the 4<sup>th</sup> test run. This facts can be seen as a “proof of concept” for the hypothesis regarding early prediction of the total number of test runs.

**Explanatory Models.** Table 2 shows the linear model fitted over the complete data set for the improvement of unit test success. We observe that the time difference launch and the difference in size are positively correlated with the improvement in unit tests success. These results support the assumption that students





**Fig. 1.** RMSE values for predicting the total number of test runs using the data up to a given test runs ID.

**Table 2.** Linear model for score improvement and total number of tests run, all the exercises combined in one data set, bold t-values are significant ( $p < 0.01$ ). Unbiased risk estimation for score improvement = 0.01 and for number of attempts = 0.03

	Improvement in score			Number of test runs		
	Estimate	Std. err.	t-val	Estimate	Std. err.	t-val
Intercept	1.78e-01	1.520e-02	<b>11.75</b>	4.764e+01	4.279e-01	<b>11.33</b>
Time diff launch	1.737e-06	2.958e-07	<b>5.82</b>	-2.945e-05	8.320e-06	<b>-3.54</b>
Time diff edit	1.928e-04	1.797e-07	0.13	-4.511e-05	5.063e-06	<b>-8.91</b>
Diff size	5.300e-02	1.415e-03	<b>2.95</b>	-2.975e-01	3.990e-02	<b>-7.45</b>
Diff error	-3.740e-02	2.089e-02	-1.79	13.694e-01	5.890e-01	0.62
Diff warning	-4.743e-02	5.008e-02	-0.94	1.491e+00	1.413e+00	1.05

who made larger and less frequent changes in their code showed greater improvement in unit test success. Furthermore, Table 2 also shows the linear model fitted over the complete data set for the number of tests run. Here we observe that the time difference launch and the difference in code size are negatively correlated to the number of test run. These results support the assumption that students who made larger and less frequent changes in code had fewer number of test runs. The average marginal effects are shown in Table 3.

**Table 3.** Average marginal effects for the models shown in Table 2

Dependent variable	Time diff launch	Time diff edit	Diff size	Diff error	Diff warning
Score improvement	1.701e-06	5.3e-06	0.0009	-0.03	-0.04
Number of test runs	-2.945e-05	-4.511e-05	-0.29	0.36	1.49

### 4.1 Categorization

In order to explain the coding behavior of the students in more details, we categorized the student population into three categories (i.e. intellects, thinkers,

and probers) based on the total number of unit test runs by each student. Table 4 presents the number of students belonging to each category for every assignment and Fig. 3 shows the change in category between two consecutive assignments. Assumptions for the suggested three categories of students, we would like to point out here that the pragmatic sense of the category labels might be different from our interpretation in the paper:

1. **Intellects:** run tests less frequently, as they are skilled and confident.
2. **Thinkers:** run tests more frequently, to receive early feedback regarding progress.
3. **Probers:** run tests most frequently, as they experience difficulty.

We would like to point out here that the categories are for each assignment and could change student to student and even for one student from one assignment to other.

**Table 4.** Number of students in the different categories for the separate assignments.

Data used	Thresholds	Intellects	Thinkers	Probers
Assignment 5	5, 14	163	131	160
Assignment 6	5, 10	173	140	141
Assignment 8	8, 19	138	132	126
Assignment 9	7, 13	88	85	62

**The Difference from the Perspective of the Three Categories.** We present the differences between the three categories with respect to the explanatory and dependent variables (Table 6). These results hold for individual assignments as well (barring a few exceptions) as shown in Table 5.

1. Significant difference on time between two student program launches ( $F [2,383] = 70.27, p = .00001$ ): post-hoc pairwise comparisons show that intellects have higher time difference than thinkers; and thinkers have higher time difference than probers.
2. Significant difference on change in code between two tests ( $F [2,383] = 198.85, p = .00001$ ): post-hoc pairwise comparisons show that intellects have greater code change than thinkers; and thinkers have greater code change than probers.
3. Significant difference on the average improvement in success ( $F [2,383] = 121.51, p = .00001$ ): post-hoc pairwise comparisons show that intellects have greater success improvements than thinkers; while thinkers have greater success improvements than probers.
4. Significant difference on average change in number of errors and warnings ( $F [2,383] = 5.79, p = .01$ ): post-hoc pairwise comparisons depict intellects reduce more errors than thinkers; while thinkers and probers have no significant difference based on reducing the number of errors in the code.

- 5. Significant difference on average success in first test run ( $F [2,383] = 16.60, p = .001$ ): post-hoc pairwise comparisons show that intellects score higher in the first attempt than thinkers; while thinkers and probers have no significant difference based on first test run scores.

**Table 5.** ANOVA results for difference measures for the three categories.

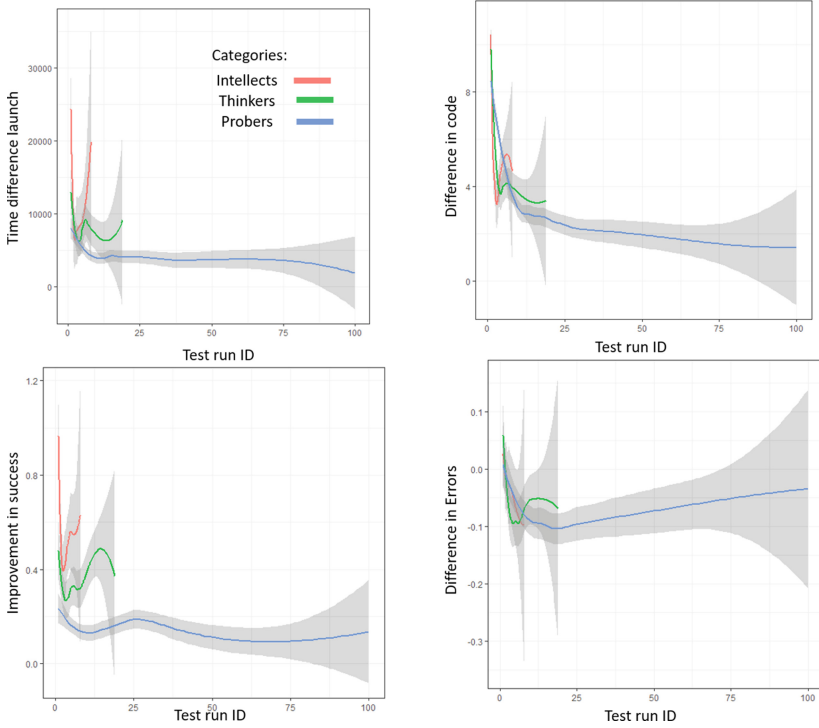
	Assignment 5		Assignment 6		Assignment 8		Assignment 9	
	F	p	F	p	F	p	F	p
Time diff launch	37.95	.0001	24.41	.0001	66.28	.0001	2.6	.10
Diff size	17.95	.0001	56.00	.0001	50.01	.0001	45.41	.0001
Diff success	94.87	.0001	39.99	.0001	60.93	.0001	31.00	.0001
Diff error	4.7	.03	2.13	.14	0.61	.43	0.65	.41
Score 1st attempt	2.4	.11	4.65	.03	10.46	.001	5.07	.02

Figure 2 shows the explanatory variables corresponding to the three categories with progress based on the number of test runs. Upon inspection of Fig. 2, (left panels) it is evident that there exists a clear difference in the time between two student program launches and the average improvement between the intellects (shown with red) and the remaining two categories for the test runs 5–10 (i.e. time between main method launches) and 15–25 (i.e. improvement). However, the other differences are not as pronounced.

From the explanatory models for each category (Table 6), we observe that the behavior of the students in each category is subtly different than the other two categories. The intellects have two positively significant coefficients: the wait between two student program launches and the change in code size. This indicates that intellects take their time to alter the code and remove errors and bugs. The thinkers have only one positively significant coefficient: the wait between two student program launches. That means the thinkers take time to test, but nothing clearly can be said about the other parameters. The probers have change in code as a negative and significant coefficient, meaning that they make smaller code changes between two successive unit tests runs.

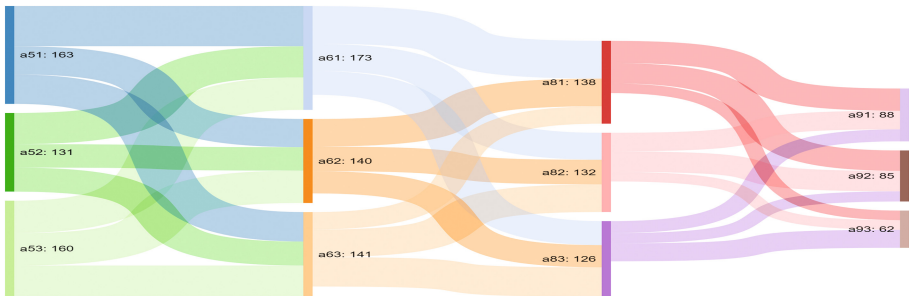
**Table 6.** Linear model for improvement with all the exercises combined in three data sets, one each for intellects, thinkers, probers, bold t-values are significant ( $p < 0.01$ ).

	Intellects			Thinkers			Probers		
	Estimate	std. err.	t-val	Estimate	std.err	t.val	Estimate	std.err	t-val
Intercept	2.9e-01	2.8e-02	<b>10.29</b>	1.6e-01	2.6e-02	<b>6.36</b>	8.5e-02	2.2e-02	<b>3.70</b>
Time diff launch	1.7e-06	4.5e-07	<b>3.76</b>	1.4e-06	6.4e-07	<b>2.19</b>	1.5e-06	4.8e-07	<b>3.13</b>
Time diff edit	8.7e-07	2.8e-07	<b>3.07</b>	-7.0e-08	3.3e-07	-0.21	-4.8e-08	3.3e-07	-0.14
Diff size	-2.3e-03	2.2e-03	-1.02	-1.2e-03	2.4e-03	-0.52	-6.2e-03	3.0e-03	<b>2.07</b>
Diff error	-6.9e-02	3.7e-02	-1.83	-7.6e-04	3.1e-02	-0.02	-5.2e-02	3.9e-02	-1.30
Diff warning	-1.1e-02	9.5e-02	-0.12	4.5e-02	8.6e-02	0.52	-1.5e-01	7.5e-02	<b>-2.00</b>



**Fig. 2.** Different measures for the three categories for each test run ID. (Color figure online)

Finally, it could be expected that students belong to more than one category while attempting to solve programming assignments. Figure 3 shows students changing across the categories intellects, thinkers and probers, for different assignments. For example, the intellects are a larger group (163) than the



**Fig. 3.** Students changing their strategies across the different assignments.  $a_{51}$  : 233 shows that in assignment a5, there were 233 students in category 1. Category labels: 1 = intellects; 2 = thinkers; 3 = probers.

thinkers (131) for assignment 5 (*a5*); for the next assignment (i.e., *a6*) we see that similar to *a5*, the largest category is intellects followed by similar numbers of thinkers and probers. Also, a large majority of intellects did not change category, while most thinkers and probers either stayed the same or interchanged categories.

## 5 Conclusion and Discussion

In this study we analyzed the programming patterns of 600 students from an introductory university course in object-oriented programming using an Eclipse plug-in to collect data. Results from the analyses supported our two assumptions: (1) there are different programming strategies that lead students to success when attempting to solve coding exercises, and (2) we can early identify low performers. Using semantic-less measures from students' coding and debugging behavior (e.g. time difference launch, time difference edit) and one code-base measure (i.e. growth in size), we managed very early (fourth attempt) to predict improvement in unit test success at a low granularity level of one student with one assignment. Our focus on semantic-less-ness lead to better reproducibility and generalizability of the results, because we can not, at least with current state-of-art, know without explicitly asking students if they are experiencing difficulty with the coding constructs (e.g. loops, recursion) or in the domain (e.g. Fibonacci numbers). Moreover, our study also adds to the growing body of research utilizing low granularity data compared to previous studies that have successfully provided predictive models that either looked at the students' level as a whole class, or focused only on code-based variables [4, 26, 38]. In addition, none of the previous studies attempted early prediction.

Furthermore, we also presented behavioral analysis of students practicing different programming strategies. Thus, we can say that *intellects* as a group are characterized by having the highest first test run score; the highest improvement in unit test success; the lowest total number of test runs among the three categories; the longest wait time between two student program launches; and finally, the most changes in the code between two unit tests. *Thinkers* are characterized as follows: a low first unit test score; a short wait time between two successive student program launches; a lower change in code size than the intellects but higher than the probers; and unit test success that is higher than the probers but lower than the intellects. Finally, *probers* are characterized by having low first unit test score; the shortest wait time between two successive student program launches; the least code size change between two successive tests; and finally, the least improvement in unit test success. The key difference between thinkers and probers is the modifications they make to the code in a similar duration of time. The thinkers appear to have a strategy to fix errors and bugs in the code, while the probers appear to employ a trial and error approach. This is also evident from Fig. 2 (bottom-left), where we can see that for a large number of attempts, the probers have slow growth (close to 0.25, that is, 4 unit test runs for passing one unit test); where as, after certain test runs students from the remaining two

categories require one or two test runs to pass one unit test. This exponential improvement is demonstrated earlier by the intellects than the thinkers, indicating that intellects initially make fewer mistakes and hence require fewer test runs to pass the complete set of unit tests. However, thinkers show more regulated and informed behaviour of testing the code than probers, and this might be a plausible explanation for why probers require more tests run to pass all of the unit tests. Consequently, from past studies we know that the weaker students have less understanding of what is tested by each test, and that makes them more likely to use a trial and error approach [25].

Finally, the prediction results presented in this study could support educators in providing motivational feedback to act as incentive to students to test their code a few more times before giving up. For example, we can predict the number of tests run a student would carryout at an early stage and we can also predict their projected improvement in unit test success at each test run. Given the current *TestRunID* and unit test score of the student, we could provide him/her with a target number of test runs at his/her given pace of improvement which might motivate the student to change their strategy (from probing to thinking) or to continue testing the code (if he/she is relatively close to the target number of tests run).

*Limitations and Future Work.* Our approach carries a few limitations that we plan to overcome in the next studies. First, this is a “black box” approach because we do not examine the code, instead we look into behavioral patterns when coding. In future work, we plan to analyze the mistakes made by the students and observe the corresponding strategic category. Next, we also did not consider any semantic features computed from the code; incorporating code metrics into the analysis could improve the prediction results. Finally, we do not gather or utilize data about students (e.g. consciousness, SRL, exam performance) or their motivation during the course, which hinders us in providing personalized feedback at this stage. Thus, we plan to incorporate this information in future studies in order to provide feedback that is not only timely and actionable, but personalized and adaptive as well.

## References

1. Alloway, T.P., Alloway, R.G.: Investigating the predictive roles of working memory and IQ in academic attainment. *J. Exp. Child Psychol.* **106**(1), 20–29 (2010)
2. Barnes, D.J., Fincher, S., Thompson, S.: Introductory problem solving in computer science. In: 5th Annual Conference on the Teaching of Computing, pp. 36–39 (1997)
3. Barrick, M.R., Mount, M.K., Strauss, J.P.: Conscientiousness and performance of sales representatives: test of the mediating effects of goal setting. *J. Appl. Psychol.* **78**(5), 715 (1993)
4. 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)


5. Bruce, C., Buckingham, L., Hynd, J., McMahon, C., Roggenkamp, M., Stoodley, I.: Ways of experiencing the act of learning to program: a phenomenographic study of introductory programming students at university. In: *Transforming IT Education: Promoting a Culture of Excellence*, pp. 301–325 (2006)
6. Busato, V.V., Prins, F.J., Elshout, J.J., Hamaker, C.: Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Pers. Individ. Differ.* **29**(6), 1057–1068 (2000)
7. Cano, F.: Epistemological beliefs and approaches to learning: their change through secondary school and their influence on academic performance. *Br. J. Educ. Psychol.* **75**(2), 203–221 (2005)
8. Chamorro-Premuzic, T., Furnham, A.: Personality traits and academic examination performance. *Eur. J. Pers.* **17**(3), 237–250 (2003)
9. Chamorro-Premuzic, T., Furnham, A.: Personality, intelligence and approaches to learning as predictors of academic performance. *Pers. Individ. Differ.* **44**(7), 1596–1603 (2008)
10. Cooper, S., Cassel, L., Moskal, B., Cunningham, S.: Outcomes-based computer science education. In: *ACM SIGCSE Bulletin*, vol. 37, pp. 260–261. ACM (2005)
11. Corno, L., Mandinach, E.B.: The role of cognitive engagement in classroom learning and motivation. *Educ. Psychol.* **18**(2), 88–108 (1983)
12. Corno, L., Rohrkemper, M.: The intrinsic motivation to learn in classrooms. *Res. Motiv. Educ.* **2**, 53–90 (1985)
13. Digman, J.M.: Five robust trait dimensions: development, stability, and utility. *J. Pers.* **57**(2), 195–214 (1989)
14. Diseth, Å.: Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. *Learn. Individ. Differ.* **21**(2), 191–195 (2011)
15. Edwards, S.H., Perez-Quinones, M.A.: Web-CAT: automatically grading programming assignments. In: *ACM SIGCSE Bulletin*, vol. 40, pp. 328–328. ACM (2008)
16. Felder, R.M., Silverman, L.K., et al.: Learning and teaching styles in engineering education. *Eng. Educ.* **78**(7), 674–681 (1988)
17. Fitzgerald, S., McCauley, R., Hanks, B., Murphy, L., Simon, B., Zander, C.: Debugging from the student perspective. *IEEE Trans. Educ.* **53**(3), 390–396 (2010)
18. Hattie, J., Timperley, H.: The power of feedback. *Rev. Educ. Res.* **77**(1), 81–112 (2007)
19. Jadud, M.C.: Methods and tools for exploring novice compilation behaviour. In: *Proceedings of the Second International Workshop on Computing Education Research*, pp. 73–84. ACM (2006)
20. Kiesmüller, U.: Diagnosing learners problem-solving strategies using learning environments with algorithmic problems in secondary education. *ACM Trans. Comput. Educ.* **9**(3), 17 (2009)
21. Lishinski, A., Yadav, A., Enbody, R., Good, J.: The influence of problem solving abilities on students' performance on different assessment tasks in CS1. In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 329–334. ACM (2016)
22. Lister, R., et al.: A multi-national study of reading and tracing skills in novice programmers. In: *ACM SIGCSE Bulletin*, vol. 36, pp. 119–150. ACM (2004)
23. Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R.F., Morales, N., Muñoz-Gama, J.: Mining theory-based patterns from big data: identifying self-regulated learning strategies in massive open online courses. *Comput. Hum. Behav.* **80**, 179–196 (2018)

24. Mitchell, C.M., Boyer, K.E., Lester, J.C.: When to intervene: toward a Markov decision process dialogue policy for computer science tutoring. In: *The First Workshop on AI-supported Education for Computer Science*, p. 40 (2013)
25. Perkins, D.N., Hancock, C., Hobbs, R., Martin, F., Simmons, R.: Conditions of learning in novice programmers. *J. Educ. Comput. Res.* **2**(1), 37–55 (1986)
26. Piech, C., Sahami, M., Koller, D., Cooper, S., Blikstein, P.: Modeling how students learn to program. In: *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, pp. 153–160. ACM (2012)
27. Pintrich, P.R.: A conceptual framework for assessing motivation and self-regulated learning in college students. *Educ. Psychol. Rev.* **16**(4), 385–407 (2004)
28. Poropat, A.E.: A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* **135**(2), 322 (2009)
29. Rivers, K., Koedinger, K.R.: Automatic generation of programming feedback: a data-driven approach. In: *The First Workshop on AI-Supported Education for Computer Science*, vol. 50 (2013)
30. Rodriguez, C.M.: The impact of academic self-concept, expectations and the choice of learning strategy on academic achievement: the case of business students. *High. Educ. Res. Dev.* **28**(5), 523–539 (2009)
31. Saeli, M., Perrenet, J., Jochems, W.M., Zwaneveld, B.: Teaching programming in secondary school: a pedagogical content knowledge perspective. *Inform. Educ.* **10**(1), 73–88 (2011)
32. Simon, B., Chen, T.Y., Lewandowski, G., McCartney, R., Sanders, K.: Commonsense computing: what students know before we teach (episode 1: sorting). In: *Proceedings of the Second International Workshop on Computing Education Research*, pp. 29–40. ACM (2006)
33. Soloway, E., Bonar, J., Ehrlich, K.: Cognitive strategies and looping constructs: an empirical study. *Commun. ACM* **26**(11), 853–860 (1983)
34. Soloway, E., Ehrlich, K.: Empirical studies of programming knowledge. In: *Readings in Artificial Intelligence and Software Engineering*, pp. 507–521. Elsevier (1986)
35. Stajkovic, A.D., Bandura, A., Locke, E.A., Lee, D., Sergent, K.: Test of three conceptual models of influence of the big five personality traits and self-efficacy on academic performance: a meta-analytic path-analysis. *Pers. Individ. Differ.* **120**, 238–245 (2018)
36. Turkle, S., Papert, S.: Epistemological pluralism and the revaluation of the concrete. *J. Math. Behav.* **11**(1), 3–33 (1992)
37. VanDeGrift, T., Bouvier, D., Chen, T.Y., Lewandowski, G., McCartney, R., Simon, B.: Commonsense computing (episode 6): logic is harder than pie. In: *Proceedings of the 10th Koli Calling International Conference on Computing Education Research*, pp. 76–85. ACM (2010)
38. Vee, M., Meyer, B., Mannock, K.L.: Understanding novice errors and error paths in object-oriented programming through log analysis. In: *Proceedings of Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems*, pp. 13–20 (2006)
39. Vihavainen, A., Vikberg, T., Luukkainen, M., Pärtel, M.: Scaffolding students' learning using test my code. In: *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education*, pp. 117–122. ACM (2013)
40. Zimmerman, B.J., Schunk, D.H.: Reflections on theories of self-regulated learning and academic achievement. In: *Self-Regulated Learning and Academic Achievement*, pp. 282–301. Routledge (2013)





# Incorporating Blended Learning Processes in K12 Mathematics Education Through BA-Khan Platform

Valeria Henríquez, Eliana Scheihing<sup>(✉)</sup> , and Marta Silva

Universidad Austral de Chile, Valdivia, Chile  
{valeria.henriquez, escheihi}@inf.uach.cl,  
marta.silva@uach.cl

**Abstract.** This study explored the impact of the use of Khan Academy as an instructional and practical teaching material in primary and secondary school classrooms. The goal was to measure the platform's effects on Chilean public school students' performance in mathematics. We used blended learning methodologies consisting of a mixture of individual work and workshops guided by teachers according to the progress of each student. This process was supported by the Khan Academy and BA-Khan Academy platforms. The latter tool, BA-Khan Academy, facilitates teachers' engagement with the blended learning process. Our analysis indicates that in the grade levels involved in the study, the students who received instruction with blended learning methodologies showed better results than control groups, both in mathematics assessments in class and on national standardized tests.

**Keywords:** B-learning · Khan Academy · K-12 · Orchestration of learning

## 1 Introduction

In Chile, annual measurements of learning outcomes are conducted through national standardized tests known as SIMCE (System of Measure of Education Quality), which evaluates achievement in the subjects of language and communication, mathematics, natural sciences, history, geography and social sciences, and English for students in years two, four, six, and eight of primary school and years two and three of secondary school. During the past decade, these measurements have indicated low performance in mathematics, showing a weak increase of only 14 points in the national average during that time [21, 22]. In addition, Chile, as a member of the OECD (Organization for Economic Cooperation and Development), has participated in the PISA exams (Program for International Student Assessment), which evaluate fundamental knowledge and abilities required for full participation in modern societies. This evaluation, applied to students at age 15, centers on basic school materials in sciences, reading, and mathematics [1]. The results of the PISA exam show performance below the OECD average in all areas assessed, and only 3% of students have a performance level of "excellent" in at least one of the exam areas. By contrast, 23.3% demonstrate low performance in all three areas. These performance levels directly impact students'

perception of the sciences and therefore their future vocations in STEM fields, given that performance in the sciences is one of the three factors that influence the choice of a STEM career. The other two factors are exposure to extra-curricular courses in mathematics and science and beliefs about self-efficacy in mathematics [2].

The traditional instructor-centered teaching model, which is still implemented in Chile, makes it impossible to give adequate attention to the diversity of student learning rhythms, capacities, and interests due to teachers' high workloads. They have little time left over for preparing classes or attending to the needs of specific groups.

The implementation of a b-learning methodology, supported by innovative technologies, reduces the amount of time that a teacher must dedicate exclusively to presenting contents [3]. As such, it allows more time for the management and planning of classes based on contents in accordance with the effort and achievement of each student, enabling informed, objective, and facilitative feedback on teaching-learning processes that attend to student diversity. In this context, the application of this methodology would produce better results than the traditional teaching methods in place. Research such as that of Cargile and Harkness has demonstrated that the b-learning model generates better results than situations in which students are learning with traditional approaches [4].

The b-learning methodology presents a combination of Internet and digital media use with formal classes that require the physical presence of the teacher and students. Numerous adaptations have arisen combining these two elements, and according to Horn, Gu, and Evans there are currently four models of K-12 b-learning: Rotation Model, Flex Model, Self-Blend Model, and Enriched-Virtual Model. Under this classification system, the Inverted Classroom Model is a submodel of the Rotation Model, which in recent years has been the object of increased attention for both research and teaching practice [5], unlike the station and laboratory rotation models, which have not been studied extensively. This study specifically addressed the station and lab rotation model since not all students in Chilean public schools have Internet access or adequate devices to access digital contents from home. Khan Academy was chosen as an LMS because it is a web platform allowing free and unlimited access to academic content validated by scores of professors specialized in mathematics and other disciplines. It was developed by a non-profit organization whose mission is to offer free, world-class education to any person in any location [6]. Since its creation, it has been implemented by diverse learning institutions including universities such as MIT and Stanford, museums, and the College Board. Its contents have been translated into more than 40 languages, and its videos and exercises have been accessed by more than 10 million students worldwide.

During the development of this study, the suitability of the use of Khan Academy was validated; this tool was further enriched by the BA-Khan Academy platform, an adaptation created by the research team that incorporates support functionalities for the professor. In this context, the questions guiding the investigation were the following: (i) Is there improvement in learning when students access the proposed b-learning modality? (ii) Is there a difference in the level of mastery acquired by students who are accompanied in the classroom by the research team in addition to the teacher, compared to those with whom the teacher works independently? (iii) Is there an impact in the medium term on the learning of students who have been involved in the b-learning experience?

## 2 Related Work

A review of studies on the impact of ICT on students' achievements reports that positive impact on students' learning has not been proven despite the increasing amount of research being done [7]. Moreover, this impact is in constant debate due to the difficulty of measuring learning. However, Trucano's report indicates that a positive impact is more likely to occur when teachers have adequate pedagogical skills and clear goals in the use of ICT, and when students' and teachers' levels of access to ICT both in and out of school are higher. In the same line, [8] point out that most of the research on the use of ICT and its impact on students' performance comes from OECD countries. Nevertheless, research from developing countries supports similar findings, including (1) that ICT helps to promote change in teaching methods in school and community improvement programs; (2) that the implementation of computers in schools is not enough when impacting students' learning outcomes, but when ICT is used for specific subjects or applications it is more likely to positively impact students' knowledge, attitudes, and skills; and (3) that ICT has an impact on specific classroom practices such as working on research projects or collaborating with students from other countries. However, in developing countries, the widespread use of ICT has barriers such as lack of infrastructure, time in the curriculum, and personnel skill levels.

There are various studies that have recognized the benefits and controversies of using the Khan Academy platform to aid student learning [9–12]. The principal benefits include the following: more dynamic classes owing to the use of instructional videos; access to teaching materials from students' homes, allowing further advancement in class contents; and the fact that assignments traditionally given as homework can be done in the classroom with the help of the instructor who in turn, thanks to the platform, can keep better track of the progress of each of his or her students. This latter benefit allows the teacher to give personalized attention to the diversity of students in the classroom because of the constant feedback that the platform provides [13]. By contrast, some of the controversial aspects of Khan Academy arise from the concept of authentic learning, which may not occur in such a platform because it depends on hierarchically-organized knowledge. Authentic learning is based on direct experience, is consolidated in practice, and requires formative feedback, all of which are characteristics that online implementations of b-learning models, including those that use Khan Academy, might not possess [11]. However, this discussion has yet to be supported by a large number of studies; the impact of the use of Khan Academy has not been studied extensively [9]. The existing research on Khan Academy has not, focused on measuring the impact of learning at the quantitative and systematic level; rather, studies have examined its use by teachers and students with qualitative and quantitative methodologies [3, 15–17, 21], and in some cases the creation and use of extensions and/or adaptations of the platform [14, 18, 20]. Regarding the use of Khan Academy by teachers and students, the results of recent studies, including one carried out in Chile [15], point to the fact that the instructor has a dominant role in the adoption of the platform and that it cannot be used effectively without a guide. They also suggest a change in classroom dynamics to allow for greater learning among peers. In addition to increasing student participation and motivation in learning mathematics, the platform

promotes autonomous learning and self-efficacy [3, 16, 17, 19]. However, the use of Khan would not necessarily change instructors' teaching methods altogether.

In terms of extensions of Khan Academy, existing studies have likewise focused on the use and adoption of these adaptations. For example, the ALAS-KA platform advances a model of visual analytics that implements the same types of graphics for each of the indicators for the 6 areas (total use of the platform, correct progress in the platform, distribution of time using the platform, gamification habits, problem-solving habits, and affective state). The objective of this extension was to facilitate the user's comprehension of his or her performance [20]. The study, though, was centered on providing guidelines and examples to help university professors make methodological decisions based on data provided by ALAS-KA.

One of the farthest-reaching investigations carried out to date has been the pilot study run by the Stanford Research Institute-Education (SRI-E), which, for two years, involved over 20 US schools in the investigation of the use of Khan Academy as a complementary resource in the classroom. The study did not measure learning impacts, but it did arrive at the following conclusions: (1) professors highly valued the use of Khan Academy as a pedagogical resource, but they consistently wanted to maintain their roles of being responsible for the contents presented and methodologies used in the classroom as well as the degree of control their students had in the use of the platform; (2) the majority of students, while they demonstrated a liking for Khan, were not ready to assume the role of autonomous learners, thus teachers had to foment the habits and learning practices necessary to work on the platform; (3) teachers valued the use of Khan as a support in their classes since it helped students understand contents—the majority expressed, therefore, that they would like to continue experimenting with the integration of the platform in their classrooms. Regarding the impact on learning, the study in question suggests that, based on the evidence gathered from one establishment, the teaching of mathematics using Khan during extended periods of time, together with close guidance by a teacher, could improve students' learning [19].

Having reviewed this group of studies addressing the use of Khan Academy in classroom contexts, it is clear that there is a lack of studies measuring its impact on learning in mathematics. Accordingly, this study—in addition to presenting the design and classroom implementation of the BA-Khan extension as a supporting tool for professors—exhibits the results of the measurement of learning outcomes following the use of this extension with Chilean students, which was done using the Rotation Model. This innovation permits the use of Khan Academy in the classroom in contexts in which digital technology is often lacking in students' homes.

### 3 Materials and Methods

This pilot study was carried out in two phases, the first of which was developed during 2015 with students in years 1 and 2 of secondary education in a public high school (School 1L). The objective of this phase was to measure the impact of the incorporation of the b-learning methodology, both for the review of previous contents and for the introduction of new contents, using a quasi-experimental design with two control groups and two treatment groups. The second phase was carried out during the first

semester of 2016 in three educational establishments, including both primary and secondary schools in various localities in the same region of Chile. This phase saw the incorporation of the BA-Khan Academy module, which was developed by the research team as a tool to complement Khan Academy, facilitating the adoption of the innovation by teachers. The objective of this second phase was to foment the continuous use of the b-learning methodology. The use of data for this study was authorized by the principal of each educational establishment in a collaborative agreement.

### 3.1 First Phase

During the first semester of 2015 at School 1L, 5 of the 7 weekly hours of mathematics instruction were carried out with the b-learning methodology, and in the second semester the number of hours was reduced to 3. The participating courses, which had 90-min class periods, utilized rotation among 4 stations: (1) autonomous work on the Khan Academy platform; (2) advanced group instruction; (3) reinforcement group instruction; and (4) tutorials among peers. For each course, students had a defined schedule indicating the classes that would be in the b-learning modality, for which they were directed to the computer lab instead of their normal classrooms. Teachers determined which contents to teach prior to each class, setting them as recommendations with date limits on the platform. These contents depended on the level of progress and mastery achieved by each student, so not all students were working on the same activities during a given class. During the first 45 min of the class, students worked at their own pace on the materials recommended by the instructor, who acted as a guide responding to questions as well as monitoring work. The “real-time” functionality of Khan Academy was utilized so the whole course could observe the “energy points” obtained (energy points measure effort, e.g., for each video completed, for practice completed, or for mastery of a task). It was recommended that students register these points at the beginning and end of each class in order to continually demonstrate their progress. At the end of this period, the professor would review the levels of mastery in the assigned activities, identifying 3 levels: advanced students, intermediate students, and those who needed reinforcement. The professor summoned each of the advanced level students to be the tutor of a group of 2–4 intermediate students. While the peer tutorials were carried out, the professor was able attend to the students who needed reinforcement and, using the board, could review the contents associated with the exercises that had been recommended at the beginning of class. Later, students could return to the exercises, with the professor present to help with questions. On other occasions the professor could opt to organize tutorial groups in which the intermediate students helped the reinforcement group, allowing personalized instruction for the advanced group to further strengthen their abilities. Five minutes before the end of class, the professor would review the “energy points” generated by all students during the session and create two rankings utilizing the “student progress” functionality, which is based on mastery of abilities (performance) and time dedicated (effort). Students in first place were called to the front of the class and applauded by their peers. The instructors who worked in the b-learning courses were the same ones who carried out classes for the control groups, teaching the latter using traditional methodologies. They participated in weekly meetings with the research team in order to receive small

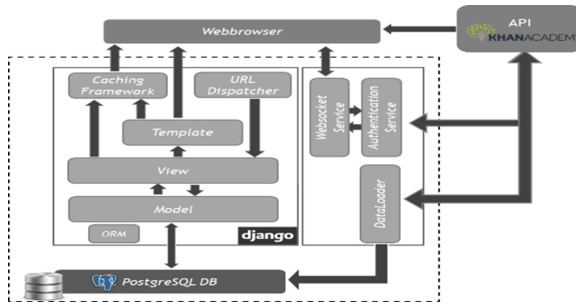
training sessions, contribute to the search for contents related to the national curriculum, and generate marks using the application. In terms of the challenges encountered, teachers demonstrated that it was somewhat difficult for them to grade activities. Grading was carried out by analyzing the use of the platform through the exportation of Excel sheets showing student progress, a process whose complexity acted as a barrier to comprehension for some teachers. In addition, the identification of the advanced, intermediate, and reinforcement groups was a complicated task for instructors since it was necessary to revise the activities assigned and level of progress attained for each student. This was particularly difficult because it was carried out during class, making it necessary to decide quickly who would participate in each group.

**Quasi-experimental design:** To study the impact on the improvement of learning outcomes during this first phase, the research team arrived at a quasi-experimental design with pre- and post-measurements for two groups, one receiving the intervention (treatment group) and the other a control. The sample considered all enrolled students in School 1L, without a selection process. It consisted of 125 students in year 1 of secondary school (courses A and B) and 102 students in year 2 (courses C and D), with the courses A and C receiving the intervention and B and D serving as controls. To measure the levels of learning achieved, 2 evaluations were carried out in all courses. These included a diagnostic exam to determine students' initial ability levels and a progress exam to determine the effectiveness of the methodology and the tools provided for reviewing and catching up on previous contents. To quantify the impact of the intervention, the differences between the diagnostic (pre-test) and progress (post-test) exams were analyzed for the treatment and control groups, and the statistical significance was measured with the help of Student's t-test for the difference of paired samples and the non-parametric Wilcoxon rank-sum test for the case of samples that did not have a behavior close to the normal distribution.

### 3.2 Second Phase

In the first semester of 2016, 3 educational establishments including both primary and secondary schools were included in the pilot study. During this period the new establishments imparted 2 of the 7 weekly pedagogical hours of mathematics with the b-learning modality. Prior to the second phase, the suggestions of the teachers who had participated in the first phase were incorporated into the BA-Khan Academy tool, which was developed and improved by the research team to support the work of teachers. In terms of technological attributes, BA-Khan Academy was developed in the Python language using the Django web framework, so it follows an MVC architecture. The data archived by the application is stored in a PostgreSQL database. The application utilizes the authentication service of Khan Academy, implementing the OAuth 2.0 protocol so that teachers can log in with the same credentials they use to access Khan Academy. Khan Academy offers a service called Khan Academy API Explorer, which presents various REST services that allow the obtention of data on students' use of the platform. BA-Khan Academy, through a batch process, repeatedly utilizes the services of Khan Academy API Explorer to collect information about the work completed by each student, storing it in the BA-Khan Academy database. This is updated

every 24 h so teachers can use the tool with the most current data. Figure 1 shows an outline of the architecture of BA-Khan Academy.



**Fig. 1.** Outline of the Architecture of BA-Khan Academy

The BA-Khan Academy tool was developed with the intention of maintaining the look and feel of Khan Academy from the user's perspective, incorporating the imagery used on the platform in 2016. It has two main macro functionalities that support teaching activity:

- (1) **Support for evaluation and grading:** Teachers can establish evaluation guidelines based on the contents reviewed in their classes with Khan Academy and configure the importance that they attribute to the level of mastery achieved in a given content area and the amount of practice completed in that area. It also allows students to be given credit based on effort indicators such as video time, review of hints, and quantity of attempts to solve problems, all of which are shown in the application. Students' scores are calculated daily, which allows teachers to offer support to students in a more opportune manner since they can monitor the level of achievement before the completion of the evaluation. Figure 2 shows the grade panel, where teachers can observe the indicators involved in the calculation of students' marks. For each student, the following is shown: an indicator of progress in the evaluated contents, that is, how much of the evaluated contents have been practiced; the quantity of exercises completed, differentiating between correct and incorrect responses; the time devoted to exercises; the time devoted to videos related to the evaluation; and the level of mastery of evaluated contents. The grade display has two characteristics: color and size. The color represents how close a student is to the maximum mark (a blue color grade) or to the minimum achievement level (shown in red). The function of effort is not part of this indicator—it is shown separately by the size of the circle containing the mark, which represents the effort associated with the mark.
- (2) **Support for classroom management:** Teachers are able to form groups based on student characteristics (abilities, effort and performance) in order to identify the aforementioned levels of reinforcement, intermediate, and advanced. In addition, the tool allows the students who have acted as peer tutors to be registered.

As shown in Fig. 3, the teacher can select the list of Khan Academy contents to automatically group based on the students' level of mastery in each of the abilities she has selected, classifying as “reinforcement” all students who are having difficulty with at least one content area, “advanced” those who have a high level of achievement (mastery 2 or mastery 3) in all content areas, and “intermediate” the remaining cases. Additionally, the instructor can register the tutors of each group in order to maintain a record of activities carried out.

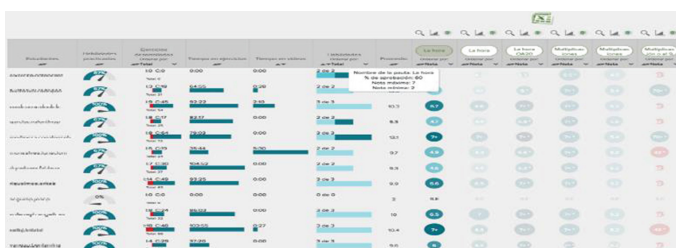


Fig. 2. Grade Panel in BA-Khan Academy



Fig. 3. Grouping panel on BA-Khan Academy

**Use Statistics:** The use of the Khan Academy platform in the establishments involved in the second phase is summarized in Table 1. During the study a sample of 118 students was examined, organized in 3 treatment groups totaling 77 students and 3 control groups totaling 41 students with one pair for each participating educational establishment (the sample was selected by convenience). In the case of the groups receiving intervention, the mathematics teachers were trained and accompanied weekly in their classrooms during the 2 pedagogical hours developed in the b-learning

Table 1. Use of the Khan Academy platform by establishments in the second phase.

School	Quantity of students	Quantity of exercises	Quantity of hints	Practice time (min.)	Average exercises per student	Average hints per student	Average time per ex. (min.)
2C	125	87226	13278	1090	698	106	9
2E	56	82729	13376	873	1477	239	16
2R	306	308179	106131	4934	1007	347	16
1L	258	285126	77411	3007	1105	300	12



modality. The same teachers carried out classes with b-learning methodology in the control groups in an autonomous manner, without accompaniment in the classroom.

To measure the impact of this accompaniment in the improvement of student learning, two evaluations were carried out in each experimental group: a diagnostic exam, which covered previous contents, and an exam evaluating progress, which was applied at the close of the first semester to measure the same expected learning outcomes evaluated on the diagnostic test. To quantify the impact of the intervention, the differences between the diagnostic (pre-test) and progress (post-test) exams in the control and treatment groups were analyzed, and the statistical significance was measured with the help of Student's t-test for the difference of paired samples. The non-parametric Wilcoxon rank-sum test was also considered in the case of samples that did not meet the condition of normality.

## 4 Results

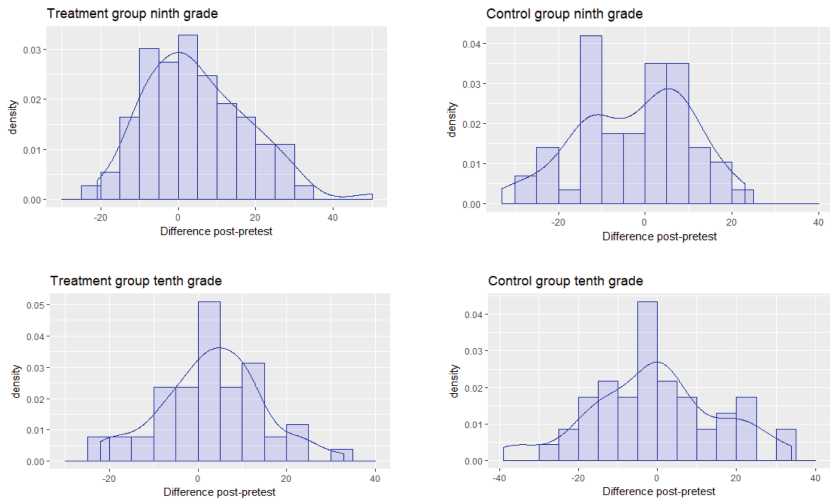
Below the results are presented from the diagnostic (pre-test) and progress (post-test) evaluations for all study groups. The scores are listed using the Chilean grading scale (1–7) multiplied by 10; in other words, the values are listed in a range of 10 to 70.

### 4.1 First Phase

In Table 2 the results of the pre- and post-tests are summarized for the two treatment groups and the corresponding control groups from the first phase of the experiment, which was carried out with students in the first and second year of secondary education at School 1L. Figure 4 complements Table 2, showing histograms of the differences between post-test and pre-test scores in the same groups. It can be observed that in the treatment groups, the mean differences are greater than 0, while in the control groups these values are negative, suggesting that the effect of advancement in learning only occurred in the case of the students in the treatment group. The results of the Student's t-test of paired samples are presented in Table 3 for each group. These results suggest that in the two treatment groups, there was a significant increase in the results of the post-test with respect to those of the pre-test, with a level of significance of 95%. In the case of the control groups, there is not a significant difference between pre- and post-test results. However, examining the normality of the differences in question with the Shapiro-Wilk test, only the treatment group from first year can be considered normal, with a confidence level of 90%. The values of the differences between the pre- and post-test results of the remaining groups do not meet the hypothesis of normality required by Student's t-test. Considering these results, the non-parametric Wilcoxon rank-sum test was calculated, the results of which are presented in Table 3. Based on the results of this test, it can be concluded that only in the case of the first year treatment group does there exist a significant increase in the results of the post-test versus the pre-test, with a confidence level of 95%. In the other cases, there are not significant differences.

**Table 2.** Statistical summary of the results of the pre- and post-test for the treatment and control groups of students in ninth and tenth grade at School 1L in 2015

		Treatment groups			Control groups		
		Pre-test	Post-test	Difference	Pre-test	Post-test	Difference
Ninth grade	Min.	20.00	23.00	-21.00	20.00	20.00	-33.00
	1st Qu.	31.00	41.00	-5.00	37.50	34.25	-13.00
	Median	43.00	47.50	2.50	47.50	44.50	0.00
	Mean	43.33	48.76	5.11	46.64	44.24	-2.40
	3rd Qu.	54.50	57.75	13.75	56.00	53.75	7.75
	Max.	70.00	70.00	50.00	70.00	70.00	23.00
	N	74	74	74	58	58	58
Tenth grade	Min.	20.00	20.00	-22.00	25.00	20.00	-39.00
	1st Qu.	36.50	37.50	-3.00	45.00	40.00	-12.00
	Median	44.00	48.00	4.00	52.00	55.00	0.00
	Mean	44.82	48.06	3.24	52.12	51.31	-0.82
	3rd Qu.	53.00	57.50	11.00	60.00	66.00	7.00
	Max.	69.00	70.00	33.00	70.00	70.00	34.00
	N	51	51	51	49	49	49



**Fig. 4.** Histograms of the differences in post- and pre-test results for treatment and control groups of students from ninth and tenth grade at School 1L in 2015.

**Table 3.** Student's t-test for paired samples and non-parametric Wilcoxon rank-sum test to establish whether there are significant differences between post- and pre-tests of the treatment and control groups of students from ninth and tenth grades at School 1L in 2015.

	Paired t test					Wilcoxon rank sum test	
	Group	Mean of Diff.	t	df	p-value	W	p-value
Ninth grade	Treatment	5.1081	3.2909	73	0.0008	2197.0	0.0282
	Control	2.3965	-1.3958	57	0.9159	1862.5	0.3199
Tenth grade	Treatment	3.2352	1.9807	50	0.0266	1119.0	0.2254
	Control	-0.8163	-0.3421	48	0.6331	1156.0	0.7541

## 4.2 Second Phase

In Table 4 the pre-test and post-test results of the three treatment groups and the corresponding control groups from the second phase of the experiment are summarized. These were carried out with students from Schools 2R, 2E, and 2C in 2016. It can be observed that in the treatment and control groups the mean differences are greater than 0, except in the case of School 2C, in which the control group exhibits a negative mean difference. The results of the Student's t-test for paired samples are presented in Table 5 for each group considered. These results suggest that for the treatment groups in Schools 2E and 2C there is a significant increase in the result of the post-test with respect to the pre-test with a 95% level of significance, while in the case of the control groups there is no significant difference between the post-test and pre-test results. In the case of School 2R, significant differences were not established in either case (with or without accompaniment in the classroom). Despite these results, when examining the normality of the differences in question with the Shapiro-Wilk test, none of the groups present values that allow the hypothesis of normality to be accepted for the data. Considering these results, the non-parametric Wilcoxon rank-sum test was calculated for the data, the results of which are listed in Table 5. Based on these results, it can be concluded that there are significant differences, with a confidence level of 90%, in all cases except for the control groups from Schools 2E and 2C. However, the validity of this test is debatable given the small sample sizes, which ranged from 8 to 46 students. In summary, all of the participating courses with training and accompaniment in this second phase improved their learning during the use of the methodology if we observe the mean values. However, these results can only be considered in the descriptive scope, and they do not allow for inferences to be made due to their lack of statistical significance. Moreover, also in the descriptive scope, we can observe that, on average, the courses in which weekly accompaniment in the classroom was not carried out had primarily positive results with the exception of School 2C, in which it is not possible to determine whether or not differences exist between pre and post-test results.

## 4.3 Standardized Test Results

In this section the evolution of the results of the SIMCE standardized test are presented for the educational establishments that participated in this study, considering the results

**Table 4.** Statistical summary of pre- and post-test results in Schools 2R, 2E, and 2C in 2016.

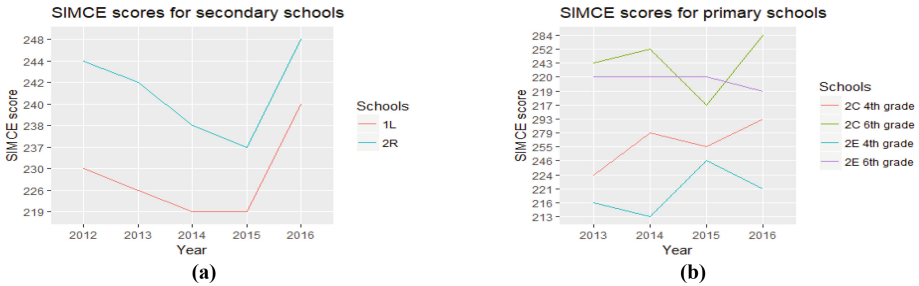
		Treatment groups			Control groups		
		Pre-test	Post-test	Difference	Pre-test	Post-test	Difference
School 2R	Min.	11.00	13.00	-12.00	9.00	7.00	-18.00
	1st Qu.	15.00	23.00	-3.00	15.00	18.00	-4.00
	Median	21.00	25.00	0.00	19.00	21.00	2.00
	Mean	20.88	24.53	2.80	17.71	20.13	1.71
	3rd Qu	25.00	29.00	9.00	21.00	23.00	4.00
	Max.	37.00	33.00	20.00	27.00	29.00	16.00
	N	15	15	15	21	21	21
School 2E	Min.	13.00	23.00	-2.00	21.00	25.00	-4.00
	1st Qu.	24.00	38.50	8.00	28.50	34.00	-1.00
	Median	33.00	42.00	12.00	32.00	38.00	3.00
	Mean	32.19	40.12	12.88	31.50	37.00	5.50
	3rd Qu	39.00	45.50	18.50	35.50	41.50	10.50
	Max.	53.00	49.00	22.00	39.00	45.00	20.00
	N	16	16	16	8	8	8
School 2C	Min.	20.00	18.00	-14.00	15.00	15.00	-16.00
	1st Qu.	30.00	30.00	-3.50	17.00	17.00	-4.50
	Median	40.00	48.00	4.00	21.00	21.00	0.00
	Mean	39.11	44.24	4.26	21.59	21.00	-0.83
	3rd Qu	46.00	56.00	11.50	25.00	23.50	4.00
	Max.	68.00	66.00	28.00	31.00	29.00	8.00
	N	46	46	46	12	12	12

**Table 5.** Student’s t-test for paired samples and non-parametric Wilcoxon rank-sum test

		Paired t test				Wilcoxon rank sum test		
		Group	Mean of differences	t	df	p-value	W	p-value
2R	Treatment		2.8000	1.2047	14	0.2483	380.5	0.0220
	Control		1.7143	1.0000	20	0.3293	469.0	0.0478
2E	Treatment		12.875	7.6057	15	0.0000	315.0	0.0131
	Control		5.500	1.6831	7	0.1362	48.0	0.1018
2C	Treatment		4.2609	2.6006	45	0.0126	2056.0	0.0266
	Control		-0.8333	-0.4033	11	0.6945	96.0	0.8044

available from 2012–2013 until 2016. In Fig. 5a, a clear increase in the performance of students from the year 2 of secondary school in School 1L can be observed in 2016 compared to previous years, and this performance corresponds precisely to those students who were exposed to the use of the b-learning methodology beginning in the first year of secondary school (2015 and 2016). In the case of School 2R, an increase can also be observed, but of lesser magnitude, when considering the previous values of the

same establishment. Figure 5b shows the performance of Schools 2E and 2C (primary schools) on the SIMCE exams for 4th and 6th grades. Improvements are not present at either level in the case of School 2E, but for School 2C an improvement can be noted at both levels in 2016.



**Fig. 5.** Results of the SIMCE exam between 2012 and 2016, obtained by the schools involved in the study: (5a, left) Schools 1L and 2R; (5b, right) Schools 2E and 2C

## 5 Discussion and Conclusions

Although sample size does not allow us to arrive at conclusions generalizable to other contexts, according to the research questions posed earlier, we can conclude the following. (i) Is there improvement in learning when students access the proposed b-learning modality? If we consider the results from the first phase, in the case of the first year secondary school students at School 1L it can be concluded that the b-learning intervention improved students' learning. In the case of the second year students, it is not possible to establish that improvements occurred because of the use of the b-learning modality due to the fact that the participating professors in the experiment were not sufficiently involved in the project. (ii) Is there a difference in the levels of mastery acquired by students who are accompanied in the classroom by the research team in addition to the teacher, compared to those with whom the teacher works independently? In this case, the results from the second phase do not indicate that there was a significant difference between the two situations. In descriptive terms, it can be established that 5 of the 6 groups showed learning improvements in terms of mean values, but this cannot be correlated to the presence or lack of accompaniment in the classroom. This result could be explained by the fact that accompaniment was not accomplished as regularly as scheduled. (iii) Is there an impact in the medium term on the learning of students who have been involved in the b-learning experience? The improvement in performance results on the SIMCE exam in the case of School 1L, where a group of students was exposed to the b-learning methodology for two consecutive years, represents a positive indication of the effect of the proposed methodology in the medium term. It is important to mention that there were also improvements in the SIMCE results in 2 of the 3 educational establishments in which the impact of the use of the b-learning methodology with BA-Khan and Khan Academy were studied (second phase), in both a primary school (School 2C) and a secondary school (School

2R). In this manner, and in accordance with what has been described in the literature, this work contributes information on an aspect of this topic that has not been studied in depth previously, that being the impact of the use of Khan Academy at different learning levels in formal schooling contexts. On the one hand, based on this study we can affirm that the use of Khan Academy in a b-learning framework, utilizing the Rotation Model, does have a positive impact on students' learning in a context in which the appropriation of the tool occurs on the part of the teachers in charge. In addition to this positive effect, it also has an impact in the medium term, as the SIMCE exam results indicate. Furthermore, the BA-Khan module developed by the research team contributes here by supporting the observations of previous studies regarding the crucial role of the professor in the adoption of Khan Academy in a b-learning framework [15, 21]. In addition, the use of this module improves the efficacy of the teacher in terms of evaluation processes and classroom management. Although the sizes of the samples in the second phase of the study do not allow us to generalize our results, they do allow us to conclude that in the study context, improvements can be seen in the levels of learning in mathematics that occurred using the BA-Khan platform to complement Khan Academy, and this independently of whether or not there was classroom accompaniment for the teacher.

**Acknowledgements.** This work was partially funded under FIC15-10 "Learning and teaching mathematics in the 21st century" and DID-UACH.

## References

1. <https://www.oecd.org/pisa/pisa-2015-results-in-focus-ESP.pdf>
2. Wang, F., Hannafin, M.: Design-based research and technology-enhanced learning environments. *Educ. Technol. Res. Dev.* **53**(4), 5–23 (2005)
3. Cargile, L., Harkness, S.: Flip or flop: are math teachers using Khan Academy as Envisioned by Sal Khan? *Trehtrends* **59**, 21 (2015)
4. Bergmann, J., Sams, A.: *Flip Your Classroom: Reach Every Student in Every Class Every Day*. International Society for Technology in Education (2012)
5. Horn, M.B., Gu, A., Evans, M.: *Knocking down barriers: How California superintendents are implementing blended learning*. Clayton Christensen Institute for Disruptive Innovation (2014)
6. Khan, S.: *The One World Schoolhouse: Education Reimagined*. Twelve, New York (2012)
7. Trucano, M.: *Knowledge Maps: ICTs in Education-What Do We Know about the Effective Uses of Information and Communication Technologies in Education in Developing Countries?* (2005). Online Submission
8. Wagner, D., Day, B., James, T., Kozma, R.B., Miller, J., Unwin, T.: *Monitoring and Evaluation of ICT in Education Projects: A Handbook for Developing Countries*. InfoDev/World Bank, Washington DC (2005)
9. Cargile, L.: Blending instruction with Khan Academy. *Math. Teach.* **1**, 34–39 (2015)
10. Severance, C.: Khan Academy and computer science. *Computer* **48**(1), 14–15 (2015)
11. Schwartz, M.: Khan Academy: the illusion of understanding. *J. Asynchronous Learning Netw.* **17**(4), 1–4 (2013)
12. Reich, J.: Rebooting MOOC research. *Science* **347**(6217), 34–35 (2015)
13. Tucker, B.: The flipped classroom. *Educ. Next* **12**(1), 82–83 (2012)

14. Cunningham J.: Reimagining Khan analytics for student coaches. In: EDM, pp. 651–652 (2015)
15. Light, D., Pierson, E.: Increasing student engagement in Math: the use of Khan Academy in Chilean classrooms. *Int. J. Educ. Dev. Inf. Commun. Technol.* **10**(2), 103–119 (2014)
16. Morrison, B., DiSalvo, B.: Khan Academy gamifies Computer Science. In: Proceedings of the 45th ACM Technical Symposium on Computer Science Education (2014)
17. Muir, T.: Google, Mathletics and Khan Academy: students' self-initiated use of online mathematical resources. *Math. Educ. Res. J.* **26**(4), 833–852 (2014)
18. Muñoz-Merino, P., Valiente, J., Kloos, C.: Inferring higher level learning information from low level data for the Khan Academy platform. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge (2013)
19. Murphy, R., Gallagher, L., Krumm, A., Mislavy, J., Hafter, A.: Research on the use of Khan Academy in schools research brief. SRI (2014)
20. Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., Leony, D., Kloos, C.D.: ALAS-KA: a learning analytics extension for better understanding the learning process in the Khan Academy platform. *Comput. Hum. Behav.* **47**(1), 139–148 (2015)
21. Ministerio de Educación: Resultados Nacionales SIMCE 2016. Unidad de Currículum y Evaluación. Santiago de Chile (2007)
22. Ministerio de Educación: Resultados Educativos 2016. Sistema de Medición de la Calidad de la Educación, Chile (2017)



# Predicting Learners' Success in a Self-paced MOOC Through Sequence Patterns of Self-regulated Learning

Jorge Maldonado-Mahauad<sup>1,3</sup>(✉), Mar Pérez-Sanagustín<sup>1,4</sup>,  
Pedro Manuel Moreno-Marcos<sup>2</sup>, Carlos Alario-Hoyos<sup>2</sup>,  
Pedro J. Muñoz-Merino<sup>2</sup>, and Carlos Delgado-Kloos<sup>2</sup>

<sup>1</sup> Department of Computer Science,  
Pontificia Universidad Católica de Chile, Santiago, Chile  
{j.maldonado, mar.perez}@uc.cl

<sup>2</sup> Department of Telematics Engineering,  
Universidad Carlos III de Madrid, Madrid, Spain  
{p.moreno, calario, pedmume, cdk}@it.uc3m.es

<sup>3</sup> Department of Computer Science, University of Cuenca, Cuenca, Ecuador

<sup>4</sup> Institut de Recherche en Informatique de Toulouse,  
Université Toulouse III Paul Sabatier, Toulouse, France

**Abstract.** In the past years, predictive models in Massive Open Online Courses (MOOCs) have focused on forecasting learners' success through their grades. The prediction of these grades is useful to identify problems that might lead to dropouts. However, most models in prior work predict categorical and continuous variables using low-level data. This paper contributes to extend current predictive models in the literature by considering coarse-grained variables related to Self-Regulated Learning (SRL). That is, using learners' self-reported SRL strategies and MOOC activity sequence patterns as predictors. Lineal and logistic regression modelling were used as a first approach of prediction with data collected from  $N = 2,035$  learners who took a self-paced MOOC in Coursera. We identified two groups of learners: (1) Comprehensive, who follow the course path designed by the teacher; and (2) Targeting, who seek for the information required to pass assessments. For both type of learners, we found a group of variables as the most predictive: (1) the self-reported SRL strategies 'goal setting', 'strategic planning', 'elaboration' and 'help seeking'; (2) the activity sequences patterns 'only assessment', 'complete a video-lecture and try an assessment', 'explore the content' and 'try an assessment followed by a video-lecture'; and (3) learners' prior experience, together with the self-reported interest in course assessments, and the number of active days and time spent in the platform. These results show how to predict with more accuracy when students reach a certain status taking in to consideration not only low-level data, but complex data such as their SRL strategies.

**Keywords:** Self-regulated learning · Prediction  
Massive Open Online Courses · Sequence patterns · Achievement  
Success



## 1 Introduction

The massive and open nature of Massive Open Online Courses (MOOCs) contribute to attract a great diversity of learners, who have seen in MOOCs an opportunity for their personal growth. Most of the learners who enroll in a MOOC decide which parts of the course content they choose to engage with, and eventually only a small proportion of these enrollees complete the course (typically less than the 10%) [8]. This has aroused the interest on studying the causes why learners complete or drop out a MOOC.

Prior research shows that self-regulation is one of the critical skills needed to achieve personal learning goals in a MOOC [19]. Self-regulated learners are characterized by their ability to initiate cognitive, metacognitive, affective and motivational processes [4]. Moreover, recent research in self-regulated Learning (SRL) suggests that successful learning and academic achievement are associated with the deployment of regulatory activities such as goal-setting, planning or monitoring [2].

MOOC enrollees present a diversity of behaviours depending on: learner's previous knowledge, prior experience, intentions and motivations [18, 24]. In a MOOC platform, this behaviour is recorded as the interactions of the learners with the course content, generating a great deal of information that offers an opportunity for identifying patterns and predict trends [11]. Actually, using all these data to run predictions about learner's success in a MOOC is of special relevance. Understanding enrollees' learning behaviour can help to detect learners who "probably" will not pass the course [28]. Moreover, this analysis could be used to better understand how learners work in the course and what kind of support he/she may need, anticipating problems which may lead to learners' dropouts.

Several studies have tried to predict attrition, retention and completion in MOOCs. Most of these studies have been carried out in cohort MOOC settings (e.g., instructor based), where time is typically structured, learners follow a fixed schedule, and course materials are released at specific times. However, in self-paced MOOCs, this prediction models may be more critical. On the one hand, the success in self-paced courses, without the support of an instructor, depends on the ability of enrollees to be able to self-regulate their behaviour [20]. On the other hand, learners' behaviour could be more variable, since students do not follow a strict schedule, all materials are released when the course starts, and dates for assessments are flexible [15].

As a consequence, to detect and predict trends in self-pace MOOCs is still a challenge that have been addressed in prior works with different approaches. For example, authors in [26] developed a grade predictive method that uses learner activity features to forecast whether or not a learner may get a certificate. Authors in [5] developed a predicting model to understand when learners will answer a question correctly. In [25], authors analysed the relationship between interactions and the number of days in which learners interact with the content.

Despite of the predictive power of the models proposed, these models raised some discussions in the community. On the one hand, some researchers argue that frequency and events count are not the best metrics to obtain practical indicators to explain individual differences in online learning [27]. On the other hand, existing models are based on the use of low-level indicators of learners' interaction with the course, but this

makes it difficult to obtain meaningful patterns of more complex behaviours, such the use of SRL strategies [27]. Therefore, there is an opportunity to improve these predictive models by considering both, data informing about the heterogeneity of learners (e.g. self-reported data about learning strategies) and more complex behaviours represented by activity sequences instead of individual events.

As a first proposal in this line, we present an exploratory study that uses SRL behavioural patterns related with learners' success as coarse-grained data to predict their behaviour in a self-paced MOOC. Specifically, we investigate whether or not learners pass the course based on these patterns together with demographic variables, SRL self-reported strategies and learners' intentions. As a result, we identified new factors to improve predictive models of learners' success in self-paced MOOCs.

## 2 Prior Work

### 2.1 Prediction in MOOCs and Self-regulated Learning

MOOCs have special features that differentiate them from other online courses. First, the big amount of global data that can be collected about learners' activity with the course content. Second, the variety of this data, in which we can identify heterogeneous profiles in terms of personality, learning preferences, education, etc. And third, the number of the interactions related to intensive use of video-lectures and assessments, less frequent in traditional online courses [22]. All these data have been used to discover predictive patterns of persistence or attrition through MOOC success and completion. Specifically, the data sources used in previous work is usually: (1) learners' demographic data, (2) learners' self-reports data (as intentions regarding the course), (3) clickstream data, (4) forums and social media data and (5) other clickstream traces [14].

In the past years, recent studies started considering not only learners' demographic data for predicting behaviour, but also self-reported data related with more complex students' learning strategies. For example, studies [6, 7] found positive relationship between learners' self-reported SRL strategies and academic achievement. According to these studies, the use of SRL strategies affects the learning outcomes achieved and is typically associated with better academic performance in both traditional and online learning situations. In study [10, 19] authors found 15 learning strategies were correlate with learners' academic performance (final grades) in online environments, and 5 were found to predict learners' grades. In another example with 50,000 learners [13], authors found significant differences in the scores obtained by learners who were already familiar or working in fields related with the MOOC content, with higher self-efficacy, than their counterparts. In another study with 4,831 learners [14], authors found that goal setting and strategic planning predicted attainment of personal course goals. Further, in [9] in a study with 2,439 learners, authors found that having a particular help seeking strategy predicts better performance in the course.

Regarding clickstream, data with video-lectures, assessments and forums have been used in predictive models. For example, studies [14, 19] use video-lectures actions related to pause, play, stop video, watch, complete or review as a method for measuring

learners' engagement the course content. Results of these studies showed that the amount of video-lectures intended and completed are predictors of course completion and showed that it is not necessary for learners to watch video-lectures from the beginning to the end to demonstrate its predictive effect [25]. In relation to assessment, different types of clickstream such as trying or completing an assessment, have been found to be predictors of course completion [19]. Researchers in [3], for example, found that the number of assessments' attempts is predictor of course completion; even more, those who try the first assessment were 30% less likely to drop out the course. Regarding the activity recorded in forums, the study [25] found that the number of forum pages viewed, or activities within the forum, such as voting up or down, were found as predictors of MOOC completion and persistence. Finally, some others clickstream traces have been found as predictors to MOOC persistence and completion, such as the number of active days that learners spent in a MOOC and the learners' pace through the contents [22].

Despite of their demonstrated predictive power, these models have some limitations. On the one hand, the use of these data sources as indicators for predict success in a MOOC are not always the more adequate. Learners' self-reported data captures only the intentions of the learners regarding the course, but not their actual behaviour. Since SRL is a continuous process rather than a single picture in time, considering indicators that come from the learners' activity within the course could be a better potential indicator. On the other hand, frequency counts of events from clickstream data and other clickstream traces that are obtained directly from low-level data are limited for detecting learners more complex behaviour in a MOOC for suggesting learning guidance. Moreover, as other studies already demonstrated, clickstream data in isolation do not necessarily build better predictive models [28]. Therefore, predictive models could be improved by adding variables built on longer activity sequences resulting from learners' interaction with the course content. That is, to propose new indicators that represent how learners adhere to the designed paths of the course, such as activity sequences extracted from coarse-grained data. This idea is built upon previous studies, which investigated the relationships between interaction sequences and learning outcomes using methods such as transition graphs, process mining, sequential pattern analysis, and Markov models [14, 19, 23].

Therefore, and based on prior work, this paper tackles the following research question: *Which indicators of SRL obtained from self-reported questionnaires and activity sequence extracted from trace data can predict course success in self-paced MOOCs?*

### 3 Methodology

#### 3.1 Context: Sample and MOOC

This study uses data from one MOOC on Electronics<sup>1</sup> offered by Pontifical University Catholica of Chile in Coursera. The course was taught in Spanish and the materials

---

<sup>1</sup> Coursera MOOC: Electrones en acción

were organised in four modules. In total the course included 17 lessons, 83 video-lectures and 16 summative assessments. The course followed a self-paced delivery mode in which course materials were available all at once, and without specific pre-defined deadlines. Data collection occurred between April and December of 2015.

A total of 25,706 learners registered for the MOOC, but the study sample is  $N = 2,035$  which corresponds with those learners who answered a self-reported SRL questionnaire that was introduced at the beginning of the course to define SRL learners' profile. Learners' average age was 30.7 years ( $SD = 11.06$ ); and the 11% were women.

### 3.2 Measures

The instrument used to define learners' SRL profile was already validated in previous studies. It contains 35 questions about learners' intentions with the MOOC content (e.g., hours expected to be dedicated to the MOOC, interest in the topic, etc.), demography (e.g., age, gender, employment status, etc.) and a measure of SRL [14]<sup>2</sup>. The SRL measure consisted of 24 statements related to six SRL strategies: goal-setting strategies (4 statements), strategic planning (4), self-evaluation (3), task strategies (6), elaboration (3) and help seeking (4). Learners rated statements using a 5-point scale (coded from 0 to 4), where a total average of 4 means a high SRL profile. The SRL measure exhibited high reliability for all strategy subscales with Cronbach's alpha of at least 0.70.

For this study, we also defined success in a self-paced MOOC based on the grades that learners achieve in the course. Therefore, success learners include any enrollee who meets one of the following two conditions:

1. obtains at least the minimum score to pass the course (80%) independently if he/she tackle most of the course materials (most common form of success),
2. obtains at least the minimum score to pass the course attempting at least 50% of the videos in the course materials

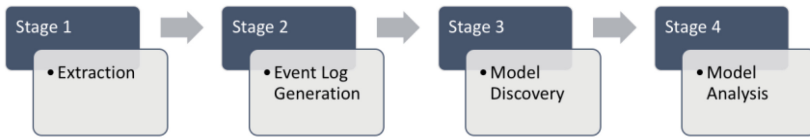
This choice is based on the common patterns that learners follow in a MOOC that were found in a previous work [19].

### 3.3 Procedure

In order to extract sequence patterns from a self-paced MOOC, we used the Process Mining method that was reported in [19]. This process is structured into four stages (see Fig. 1):

- (1) *Extraction stage*. In this stage, the data is extracted from the Information System databases (Coursera in our case). We obtained the trace data from Coursera database in order to study the interaction sequences of learners in the MOOC. This raw data is organised into three categories: (a) general data, (b) forums, and (c) personal data that contain relevant information about learners' behaviour.

<sup>2</sup> SRL measure questionnaire in Spanish and English are available at <https://doi.org/10.6084/m9.figshare.1581491>.



**Fig. 1.** Stages for extracting sequence patterns using process mining method.

- (2) *Event log generation stage.* In this stage gathered data is modeled in terms of event logs, defining the concepts of case (execution of a process), activities (steps of the process), and temporal order of the activities. We defined the main event log file including the learners' interactions in the MOOC within a session, their SRL scores, as well as information required to perform the analysis, such as the case id, time stamp and other resources. In this stage, we defined the concepts of (1) session and (2) interaction.

A *session* is defined as a period of time in which the Coursera trace data registers continuous activity of a learner within the course, with intervals of inactivity no greater than 45 min; this definition of study session has been already adopted in prior works [16].

An *interaction* is defined as an action recorded in the Coursera trace data that registers the interaction of a learner with a MOOC content. We defined six types of interactions depending on the content that learners interact with (video-lectures/assessments):

- **Video-lectures:** (1) *start a video-lecture* (begin to watch a video-lecture for the first time without completing it), (2) *complete a video-lecture* (watch a video-lecture entirely for the first time), (3) *review a video-lecture already completed* (go back to a video-lecture which was already completed)
- **Assessments:** (1) *try an assessment* (attempt to solve an assessment), (2) *pass an assessment* (successful attempt to solve an assessment for the first time), (3) *review an assessment already passed* (go back to an assessment that was previously completed successfully).

After defining these key concepts, we extracted the study sessions and coded as consecutive learning actions (interactive sequences) performed by learners when interacting with MOOC resources, such as video-lectures and assessments. Finally, we defined an event log that included a label to identify the first (begin session) and last interaction of the learner with the course (end session). Besides the interactions with the course, the event log also included learners' SRL scores obtained from the self-report questionnaire. The Table 1 shows an example of the event log generated.

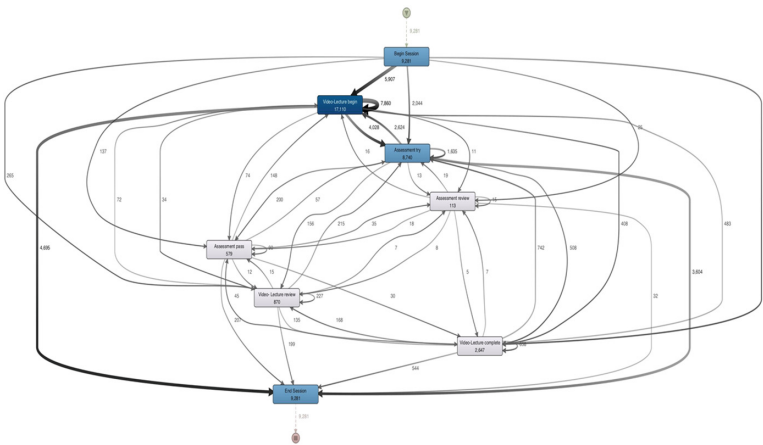
- (3) *Model discovery stage.* In this stage, Process Mining (PM) discovery algorithms are applied to the event log to obtain a process model (process map). This model represents the behaviour of the learners in the MOOC as a result of its interaction with the video-lectures and assessments. We selected the Disco algorithm and their implementation in the Disco commercial tool [12]. This algorithm is based on the Fuzzy algorithm concept combined with some features from the Heuristic

algorithm family [1]. We use this algorithm given that the exploratory context of this study in which is necessary to handle complex processes and the resulting models can be understood by experts in the domain without experience in PM [10].

- (4) *Model analysis stage.* In this stage, the discovered process models are analysed in order to understand the observed behaviour (see Fig. 2). Once the process model was generated, we identified learners' most frequent interaction sequences that characterize each session for a learner (an interaction sequence is defined as a set of concatenated interactions, from one interaction to another one, of the same learner within a session). That is the learner's path followed in the MOOC within a session (see Fig. 3).

**Table 1.** Example of the event log generated.

Case ID	Time stamp	Interaction	SRL Scores
1acc92cf40b27c8a36ea9d	1451023929	Begin session	3,162
1acc92cf40b27c8a36ea9d	1448567431	Video-Lecture.begin	3,162
1acc92cf40b27c8a36ea9d	1448567737	Video-Lecture.complete	3.162
1acc92cf40b27c8a36ea9d	1448568139	Assessment.try	3.162
1acc92cf40b27c8a36ea9d	1449105157	End session	3,162



**Fig. 2.** Process model obtained containing all interaction sequences by sessions.

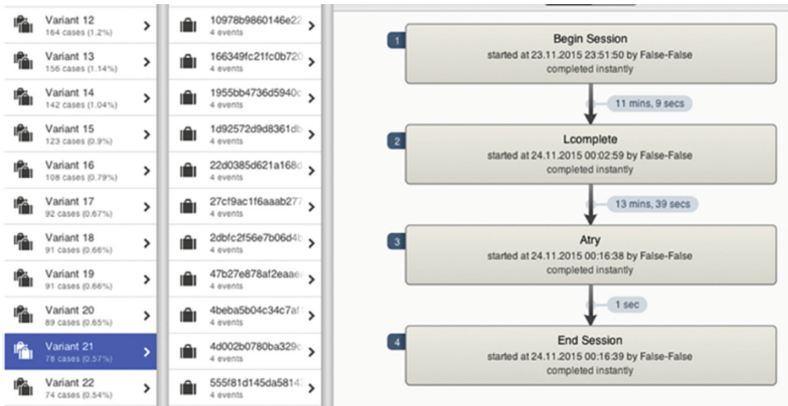


Fig. 3. List of the 1366 sessions obtained using Disco software. Session 21 shows the begin and the end of the session and 2 interactions (events) with 3 interaction sequences and the time associated with the duration of the session (variant 21).

### 3.4 Proposed Approach

Once the process model was generated and in order to answer the research question, we set up the proposed approach in two steps: (1) extracting meaningful SRL patterns, (2) applying predictive models.

(1) **Extracting SRL Patterns.** We used Process Mining techniques following the PM<sup>2</sup> method used in [19] to identify the most frequent interaction sequences of learners. As a result, six interaction sequences patterns were identified: (1) *only video-lectures*, (2) *only assessment*, (3) *explore*, (4) *assessment-try to video-lecture*, (5) *video-lecture-complete to assessment-try*, and (6) *video-lecture to assessment-complete*. Then, the interaction sequences patterns extracted were used as input for grouping learners with similar behaviour. This was done through agglomerative hierarchical clustering based on Ward’s method. This clustering technique is advisable for detecting learner groups in online contexts [16]. To select the optimal number of clusters, we inspected the resulting dendrogram and looked for different ways of cutting the tree structure, in order to obtain a minimal number of interpretable cluster explaining user behaviour (also the number of clusters were confirmed using the Silhouette method). As a result, the cluster indicates different kinds of learning strategies that learners deploy when they are facing the MOOC. Three clusters that classify learners according to their interaction sequences patterns and SRL profile were obtained. These clusters are:

- **Sampling Learners (cluster 1):** They have a low activity in the course. Generally, learners in this group “sample” the course materials and then, leave the course (n = 1,530). Only 7 learners complete the course.

- **Comprehensive Learners (cluster 2):** These learners usually follow the path designed by the instructor. They also invest more time watching video-lectures and then try assessments for deeply learning (n = 85). Only 30 learners complete the course.
- **Targeting Learners (cluster 3):** They watch fewer video-lectures than comprehensive learners, and focus on completing the assessments, thus being more strategic or goal oriented (n = 420). Only 143 learners complete the course.

We look for statistically differences between clusters 1, 2 and 3 based only in the SRL profile (mean) running t-tests. As a result, no statistically significant differences between comprehensive (cluster 2) and targeting (cluster 3) learners were observed based on the SRL profile. Consequently, we selected these two as groups of interests to explore if we can find differences in the predictors of the grades between them.

(2) **Applying Predictive Models.** Once we identified the mined sequence patterns, we combined these with self-reported SRL strategies, other traditional self-reported variables such as demographics, intentions, and variables that result from the activity of the learner within the platform, in order to identify which of these variables (fine- and coarse grained) are predictors of learners' success in self-paced MOOCs.

In order to assess whether the variables in Table 2 had statistically significant and independent effects for predicting learners' success, we conducted multiple linear regression analyses and logistic regression analysis. Variables used in the predictive

**Table 2.** Predictors classified by categories

Category	Predictors
SRL Strategies	(1a) Goal setting (1b) Strategic planning (1c) Self-evaluation (1d) Task strategies (1e) Elaboration (1f) Help-seeking
Sequence patterns	(2a) Only video-lectures (2b) Only Assessment (2c) Explore (2d) Assessment-try to video-lecture (2e) Video-lecture-complete to assessment-try (2f) Video-lecture to assessment-complete
Demographics	(3a) Age (3b) Gender (3c) Employment status (student) (3d) Employment status (job)
Intentions	(4a) Time commitment (4b) Interest in topic (4c) Interest in assessment (4d) Prior experience
Activity	(5a) Active days (5b) Time spent (minutes) (5c) Number of sessions



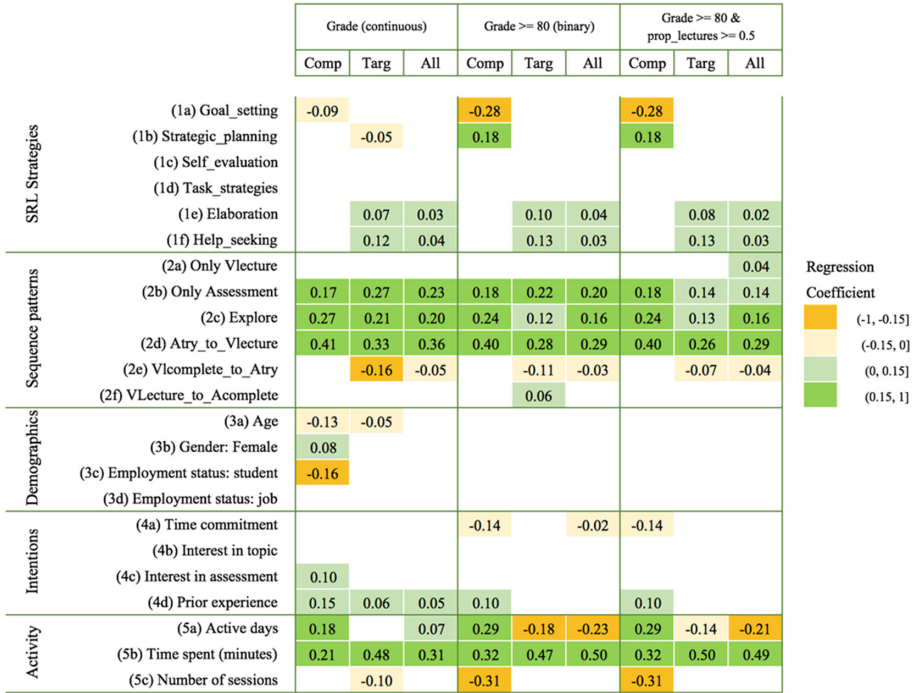
model were selected by means of a stepwise regression, using the 23 predictors. Stepwise regression uses an algorithm to select the best grouping of predictor variables that account for the most variance in the outcome ( $R^2$ ); this technique is useful in exploratory studies or when testing for associations.

All the predictors are continuous except for gender, employment status, interest in topic, interest in assessment and prior experience, which are dummy-coded binary predictors. Finally, with the self-reported data on SRL strategies as well as the patterns extracted, demographic data about learners, intentions towards the course and activity registered in the course, we built a dataset containing 23 variables that were considered as possible predictors of success. These predictors are presented in Table 2.

## 4 Results

### 4.1 Regression Analysis of Course Success

We assessed individual differences between three groups: (1) Comprehensive learners as a group (cluster 2), (2) Targeting learners as a group (cluster 3) and (3) all learners as one group (cluster 1, cluster 2 and cluster 3). For this assessment, we used 23 individual characteristics, encompassing SRL strategies, sequence patterns extracted from the behaviour of the learner with the course content, demographics, intentions and activity with the course resources. Figure 4 illustrates the results of the regressions, one for each group, with estimated standardized coefficients (sign and magnitude) from each model in each column. Blank entries in Fig. 4 indicate that the corresponding predictor was excluded from the model. These standardized coefficients were obtained after running multiple linear regression and logistic regression. For each group, we have considered grades as a dependent variable. For multiple linear regression, the grades were considered as a continuous variable. For logistic regression, the grades were considered as a binary variable (grade  $\geq 80$ ; grade  $\geq 80$  & proportions of video-lectures  $\geq 50\%$ ). A number of individual differences emerged for learners who succeed in a MOOC across different set of indicators and depending on the group in which they were classified. For comprehensive learners, the *strategic planning* strategy was associated with success in the course, while *elaboration and help seeking* were the strategies associated with success for targeting learners (grade  $\geq 80$ ; grade  $\geq 80$  & proportions of video-lectures  $\geq 50\%$ ). Comprehensive learners who performed the sequence patterns *only assessment*, *explore*, and *assessment try to video-lecture* while they were facing the course, were more successful (grade  $\geq 80$ ; grade  $\geq 80$  & proportions of video-lectures  $\geq 50\%$ ). Targeting learners who performed the sequence patterns *only assessment* and *assessment try to video-lecture* were more successful (grade  $\geq 80$ ), while for the same group the strategy *assessment try to video-lecture* was associated only with success (proportions of video-lectures  $\geq 50\%$ ) if learners passed the course and attempted, at least, 50% of video-lectures. Regarding activity indicators, comprehensive learners who spent more *active days and time* in the MOOC were more successful, while targeting learners only *time spent* was associated with success.



**Fig. 4.** Individual differences between 3 groups of learners (comprehensive, targeting, all) considering the grade as a continuous and binary variable (grade >= 80; grade >= 80 & proportions of video-lectures >= 50%), examined by SRL strategies, sequence patterns, demographics, intentions and activity. Blank boxes indicate predictor variables that were excluded by variable selection. Colors indicate the sign and magnitude of standardized coefficients. All regression coefficients are significant (p < .001).

To predict the final grade (as continuous), we run a stepwise method. As a result, we obtained 3 models for (1) Comprehensive learners as a group, (2) Targeting learners as a group, and (3) all learners as one group. Table 3 describes the regression models obtained for each group.

**Table 3.** Summary of the models using multiple linear regressions for the three groups (grade continuous)

Group	R <sup>2</sup>	adj. R <sup>2</sup>	df	F	p
(1) Comprehensive	0.8296	0.8039	73	32.31	<0.001
(2) Targeting	0.7249	0.7175	408	97.73	<0.001
(3) All	0.8559	0.8552	2026	1202	<0.001

For group (1) Comprehensive learners, the self-reported variable *goal setting*, the sequences patterns *only assessment*, *explore* and *assessment try to video-lecture*, the reported demographics as *young learners*, *be women* and *employment status as student*, the learners’ *prior experience* and *interest in assessment* reported, the *active days* and the *time spent* were significant predictors of the final grade. These variables explained 80.39% of the variance in the final grade ( $R^2 = .8039$ ,  $F = 32.31$ ,  $p < .001$ ).

For group (2) Targeting learners the self-reported variables *strategic planning*, *elaboration* and *help seeking*, the sequences patterns *only assessment*, *video-lecture complete to assessment try*, *explore* and *assessment try to video-lecture*, the reported demographics as *young learners*, the learners’ *prior experience*, the *time spent*, and the *number of sessions* were significant predictors of the final grade. These variables explained 72.49% of the variance in the final grade ( $R^2 = .7249$ ,  $F = 97.73$ ,  $p < .001$ ).

For group (3) “All learners as one group”, the self-reported variables *elaboration*, and *help seeking*, the sequences patterns *only assessment*, *video-lecture complete to assessment try*, *explore*, and *assessment try to video-lecture*, and the learners’ *prior experience* reported, the *active days* and the *time spent* were significant predictors of the final grade. These variables explained 85.5% of the variance in the final grade ( $R^2 = .855$ ,  $F = 1,202$ ,  $p < .001$ ).

The sequence patterns *only assessment*, *explore* and *assessment try to video-lecture*, and the *time spent* were significant positive predictor for the three groups. The magnitude of the standardized coefficient for the predictor *assessment try to video-lecture* for group “Comprehensive” and “All”, and the magnitude of the standardized coefficient for the predictor *time spent* for “Targeting” were the highest. It is also worth noting that *video-lecture complete to assessment try* and *employment status as student* were significant negative predictors for “Targeting” and “Comprehensive” respectively.

Finally, an evaluation of the models was performed to analyze the predictive power. The dataset was split in train and test sets (80% for training and 20% for testing) and 10-fold Cross Validation (CV) was used within the training set. The first model to predict continuous grades was evaluated through the Root Mean Square Error (RMSE), while the other models to forecast binary variables were assessed through the accuracy, kappa and the Area Under the Curve (AUC) (see Table 4).

**Table 4.** Evaluation of the predictive models

Cluster	Set	Grade (continuous)	Grade > = 80 (binary)			Grade > = 80 & prop_lectures > = 0.5 (binary)		
		RMSE	Accuracy	Kappa	AUC	Accuracy	Kappa	AUC
All	CV	11.30	0.95	0.74	0.98	0.96	0.77	0.98
	Test	11.85	0.95	0.70	0.98	0.95	0.70	0.98
Comprehensive	CV	16.62	0.82	0.63	0.84	0.82	0.63	0.84
	Test	11.66	0.94	0.86	0.92	0.94	0.86	0.92
Targeting	CV	17.22	0.86	0.70	0.92	0.83	0.63	0.92
	Test	17.86	0.80	0.57	0.90	0.90	0.79	0.92

\* CV – Cross Validation; AUC – Area Under the Curve

Results show that the predictive power is higher with all learners. This is normal because sampler learners are also included, and their grade is easier to predict given that sampler learners do not do the activities and they fail. As for comprehensive, some differences are encountered between the train and test set. The reason is that there are very few comprehensive learners and data limitations may suppose generalization issues. Nevertheless, the kappa values indicate at least substantial agreement [17] in all cases (in all groups) and AUC values are excellent [21] (excepting the AUC value for comprehensive learners in CV, which can be considered good). These results entail that the new variables related to self-regulated learning and sequence patterns can be useful for predicting grades, together with the well-known activity variables.

## 5 Conclusions

This paper has presented an exploratory study on the variables that are good predictors of the success (grades) for three groups of learners in a self-paced MOOC: “Comprehensive”, “Targeting” and “All” learners. Comprehensive learners are those who follow the course path designed by the teacher. Targeting learners are those who seek for the information required to pass assessments. For both type of learners, we found a group of variables as the most predictive: (1) the self-reported SRL strategies ‘goal setting’, ‘strategic planning’, ‘elaboration’ and ‘help seeking’; (2) the activity sequences patterns ‘only assessment’, ‘complete a video-lecture and try an assessment’, ‘explore the content’ and ‘try an assessment followed by a video-lecture’; and (3) learners’ prior experience, together with the self-reported interest in course assessments, and the number of active days and time spent in the platform.

The variables analysed in these groups were extracted from self-reported SRL strategies, mined interaction sequence patterns, traditional self-reported variables such as demographics, intentions, and variables that result from the activity of the learner within the platform. Multiple linear regression models were obtained for each of the three groups of learners, which are statistically significant at 99,9% level of confidence.

The findings of this study are subject to some limitations due to the nature of data, and methodological choices. First, the study is based on learners’ behavioural data automatically collected by the platform, and self-reported data collected from an optional survey. Second, the study sessions are computed considering an inactivity threshold of 45 min, and only the interactions of learners with video-lectures and assessment were used to extract interaction sequence patterns.

Future work will expand the study considering (1) week by week analysis instead of per sessions, and (2) considering interaction sequence patterns mined by using other MOOC resources such as forum messages, readings, use of dashboard, access to external resources outside the MOOC, and formative activities. We will also consider exploring different types of courses, those that have a defined start and end date. This, with the aim of finding other factors that affect the predictive power when forecasting grades. The final aim is to better understand how a student reaches the status of comprehensive or targeting.

**Acknowledgments.** This work was supported by FONDECYT (Chile) under project initiation grant No.11150231, the MOOC-Maker Project (561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP), the LALA Project (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), and CONICYT/DOCTOR ADO NACIONAL 2016/21160081, the Spanish Ministry of Education, Culture and Sport, under an FPU fellowship (FPU016/00526) and the Spanish Ministry of Economy and Competitiveness (Smartlet project, grant number TIN2017-85179-C3-1-R) funded by the Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER).

## References

1. Van der Aalst, W.M.P.: Process mining: data science in action. Springer, Heidelberg (2016)
2. Bannert, M.: Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychol.* **23**(2), 139–145 (2009)
3. de Barba, P.G., et al.: The role of students' motivation and participation in predicting performance in a MOOC. *J. Comput. Assist. Learn.* **32**(3), 218–231 (2016)
4. Boekaerts, M.: Self-regulated learning: a new concept embraced by researchers, policy makers, educators, teachers, and students. *Learn. Instr.* **7**(2), 161–186 (1997)
5. Brinton, C.G., et al.: Mining MOOC clickstreams: video-watching behavior vs. in-video quiz performance. *IEEE Trans. Signal Process.* **64**(14), 3677–3692 (2016)
6. Broadbent, J.: Comparing online and blended learner's self-regulated learning strategies and academic performance. *Internet High. Educ.* **33**, 24–32 (2017)
7. Broadbent, J., Poon, W.L.: Self-regulated learning strategies & academic achievement in online higher education learning environments: a systematic review. *Internet High. Educ.* **27**, 1–13 (2015)
8. Chuang, I., Ho, A.D.: HarvardX and MITx: Four Years of Open Online Courses—Fall 2012–Summer 2016 (2016)
9. Corrin, L., et al.: Using learning analytics to explore help-seeking learner profiles in MOOCs. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 424–428 (2017)
10. Davis, D., et al.: Activating learning at scale: a review of innovations in online learning strategies. *Comput. Educ.* **125**, 327–344 (2018)
11. Grainger, B.: Massive open online course (MOOC) report 2013. University of London. (2013)
12. Günther, C.W., Rozinat, A.: Disco: discover your processes. *Bus. Process Manag.* **940**, 40–44 (2012)
13. Hood, N., et al.: Context counts: how learners' contexts influence learning in a MOOC. *Comput. Educ.* **91**, 83–91 (2015)
14. Kizilcec, R.F., et al.: Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Comput. Educ.* **104**, 18–33 (2017)
15. Kocdar, S., et al.: Measuring self-regulation in self-paced open and distance learning environments. *Int. Rev. Res. Open Distrib. Learn.* **19**, 1 (2018)
16. Kovanović, V. et al.: Penetrating the black box of time-on-task estimation. In: *Proceedings of the Fifth International Conference Learning Analytics and Knowledge - LAK 2015*. October, pp. 184–193 (2015)
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics.* **33**(1), 159–174 (1977)
18. Littlejohn, A., et al.: Learning in MOOCs: motivations and self-regulated learning in MOOCs. *Internet High. Educ.* **29**, 40–48 (2016)

19. Maldonado-Mahauad, J., et al.: Mining theory-based patterns from Big data: identifying self-regulated learning strategies in Massive Open Online Courses. *Comput. Hum. Behav.* **80**, 179–196 (2018)
20. Maldonado, J.J., et al.: Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach. In: *Computing Conference (CLEI), 2016 XLII Latin American*, pp. 1–12 (2016)
21. Mezaour, A.-D.: Filtering web documents for a thematic warehouse case study: eDot a food risk data warehouse (extended). In: Kłopotek, M.A., Wierchoń, S.T., Trojanowski, K. (eds.) *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, vol. 31, pp. 269–278. Springer, Berlin, Heidelberg (2005). [https://doi.org/10.1007/3-540-32392-9\\_28](https://doi.org/10.1007/3-540-32392-9_28)
22. Moreno-Marcos, P.M., et al.: Analysing the predictive power for anticipating assignment grades in a massive open online course. *Behav. Inf. Technol.* **37**(5), 1–16 (2018)
23. Pardo, A. et al.: Generating actionable predictive models of academic performance. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 474–478 (2016)
24. Reich, J.: Rebooting MOOC research. *Science* **347**(6217), 34–35 (2015)
25. Sinha, T. et al.: Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv Prepr. arXiv1407.7131*. (2014)
26. Xu, B., Yang, D.: Motivation classification and grade prediction for MOOCs learners. *Comput. Intell. Neurosci.* **2016**, 4 (2016)
27. You, J.W.: Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet High. Educ.* **29**, 23–30 (2016)
28. Zhao, C., et al.: Discover learning behavior patterns to predict certification. In: *2016 11th International Conference on Computer Science & Education (ICCSE)*, pp. 69–73 (2016)



# Semantically Meaningful Cohorts Enable Specialized Knowledge Sharing in a Collaborative MOOC

Stian Håklev<sup>1</sup>(✉), Kshitij Sharma<sup>2</sup>, Jim Slotta<sup>3</sup>, and Pierre Dillenbourg<sup>1</sup>

<sup>1</sup> École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland  
stian.haklev@epfl.ch

<sup>2</sup> Norwegian University of Science and Technology, Trondheim, Norway

<sup>3</sup> University of Toronto, Toronto, Canada

**Abstract.** This study presents an analysis of a MOOC on inquiry and technology for in-service teachers, which was designed to scaffold multiple disciplinary knowledge communities through common weekly themes, and course-long collaboration scripts happening at different social planes. Using our course design to inform the design of the analysis, we examine how the discourse in each semantically meaningful cohort (Special Interest Groups, SIGs) is indexed to the weekly themes, and develops these themes in areas informed by the discipline, and by the group dynamics. We show that SIG membership influences individual contributions, and that more cohesive disciplinary SIGs are correlated with higher quality student work.

**Keywords:** Inquiry-based learning · MOOCs  
Massive Open Online Courses · Learning analytics  
Multi-level analysis · CSCL

## 1 Introduction

Massive Open Online Courses (MOOCs) attract large numbers of students with very diverse backgrounds and interests. The experiences, ideas, and collective energy of these students could potentially contribute a large amount to the learning experience, however the very number of students also represents an almost insurmountable challenge for teachers wishing to implement a knowledge-community approach in their courses.

Some MOOC platforms offer course cohorts as a solution—assigning students to random groups, and making forums local to each group, as a way to avoid information overload. We posit that grouping students based on their specific interests, and giving them access to rich and diverse knowledge tools, not just forums, can significantly improve the quality and relevance of their discussions.

In this study, we will present an analysis of a MOOC for in-service teachers which ran on the EdX platform. The course, which attracted around 8,500

registrations, and around 2,200 active users, focused on integrating technology and inquiry into the lesson design process, and used a large amount of custom activities to enable both crowd-sourcing and small group collaboration, with the goal to support transfer from theoretical concepts to students' professional lives.

Too much learning analytics research on MOOCs has treated every course as interchangeable, whereas we argue that taking into account the instructional design and structure of the MOOC is key to understanding the individual and collaborative processes of students [12]. In this paper, we attempt to analyze a MOOC with nested social structures, and complex interactions between multiple pedagogical scripts.

Our main goal is to use learning analytic approaches to explore how this intentional theoretically informed course design actually contributed to structure students' conversations and collaborative work. We also begin to explore factors contributing to higher quality artefacts, although we are not making the argument that this is a valid indicator for individual student learning.

Below, we will present some of the design features relevant to the subsequent analysis (for a more in-depth exploration of the course design, see [6]).

## 1.1 Course Design

The course design was an attempt at mapping the Knowledge, Community and Inquiry framework [16] to a large-scale setting, which meant ensuring that students' knowledge production was indexed to a knowledge structure representing the learning goals of the course. We conceived of the course design as a matrix, combining specific weekly content themes with course-long collaboration scripts. Students joined a Special Interest Group (SIG), for example "secondary science" with a few hundred others, and within the SIG, had the option of engaging in a lesson design project with up to six others.

As Fig. 2 shows, students engaged in collaborative scripts on multiple levels of granularity (whole class, SIG, and design group) beginning in the pre-course lounge, and continuing throughout the course. These scripts were then tied together through the weekly themes, which permeated all scripted activities during a given week, with the scaffolded design of a lesson in small groups of 3–6 students providing the organizing principle throughout the course.

For a given week on the theme of "collaborative learning", a student would begin by watching videos (lectures and mini-documentaries) about collaborative learning, followed by a personal reflection about collaboration (related to their own teaching practice), before responding to several prompts related to collaboration in their SIG. He or she would then look at the in-progress lesson designs in their SIG, and add a review comment addressing how this team could incorporate more collaboration into their design. Finally, a student who was a member of a Lesson Design group would log into their Collaborative Workbench, see the weekly prompt (related to collaboration, see Fig. 2), as well as the peer review comments from all their peers, and continue work on improving their lesson design document, informed by all the preceding activities (Fig. 1).



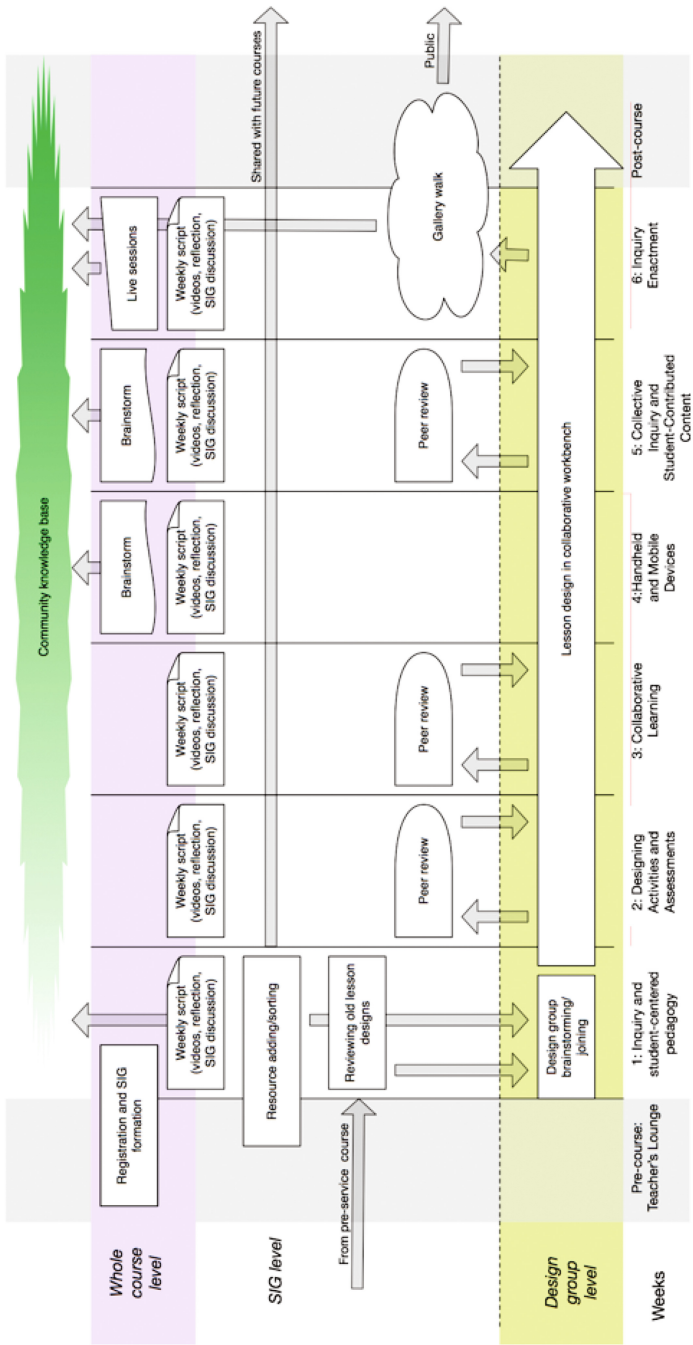


Fig. 1. Graphical depiction of interconnecting MOOC scripts.

1. Describe a typical classroom where this lesson might be enacted.
2. Describe the major theme of the lesson.
3. What are the learning goals of the technology-enhanced lesson?
4. Some aspects of the design (complete any that are relevant)
  - (a) Student-Centered Design
  - (b) Peer Collaboration
  - (c) Use of Handheld or Mobile Computers
  - (d) Supporting Equity and Diversity
5. What is the activity structure of the lesson?
6. Assessment notes.
7. Enactment notes.
  - (a) Ethics or enactment concerns

**Fig. 2.** Lesson design outline

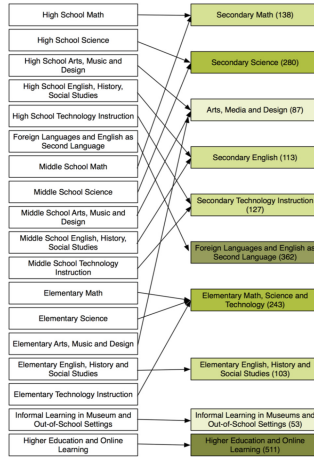
The collaborative workbench interface where students worked on their lesson designs featured weekly prompts, incoming information from the community (for example review comments), an Etherpad (collaborative scratchpad) for notes and ideas private to the group, and a wiki where the group authored the actual lesson design. The wiki page was gradually seeded with template headers which increased in sophistication each week (from learning goals, and activity structure, to technology integration, and assessment), and was exposed to the rest of the class for weekly reviews (see Fig. 2).

In this way, students began each week by receiving abstract and general ideas from the MOOC videos, and continued to engage with these ideas individually and in large and small groups, progressively making them more concrete, and more applied to a specific discipline and a specific lesson design, in the process increasing both the relevance and transferability of abstract concepts.

## 1.2 Special Interest Groups

Several weeks before the course had officially begun, we opened the “teachers’ lounge” — a virtual site for teachers to congregate, fill out a survey about their professional interests, and begin contributing resources to the course community. Based on this information, we produced a list of 18 suggested Special Interest Groups, designed to balance homogeneity (teacher discipline/age level taught) and number of participants. MOOC participants were invited to choose one of these SIGs, and based on their actual choices, we then combined a number of SIGs to rebalance the number of group participants (see Fig. 3).

The main focus of the course was science and technology in K-12 classrooms, which led to some very specialized SIGs, such as “Secondary Maths”, and some more general ones, like “Elementary Math, Science and Technology”, but also non-science SIGs, like “Arts, Media and Design”, and two non-K-12 SIGs: “Informal learning and museums”, and “Higher education and online learning”.



**Fig. 3.** Special Interest Groups: Initial and final configurations.

Our SIGs were different from most MOOC cohorts in two ways. First, they were semantically meaningful, i.e. designed based on actual data on student interests and professional contexts, and then actively chosen by the students. Second, in traditional MOOCs, cohorts are just applied to forum participation, but in this MOOC the integration between the EdX forum cohorts and our external activities, such as group peer review, and lesson design collaborative workbench, provided students with a rich variety of forms of engagement, knowledge exchange, and collaboration.

### 1.3 Previous Research

In an attempt to understand how this nested social structure, and the semantically meaningful SIGs, contributed to structuring student interaction and discourse, and impacted the quality of the final artefacts (lesson design documents), we have previously analyzed individual student activity traces (video watching and forum access behavior), collaborative actions within design groups, and the social network structures within SIGs [5]. We have presented a coding scheme for the quality of the lesson design documents across five dimensions (Table 1), and correlated these quality metrics with individual and SIG characteristics.

We have found evidence of different SIGs “making the MOOC their own”, with significant differences in video-watching behavior between K-12 SIGs, for whom the MOOC was originally planned, watching more of the K-12 focused videos, and higher education SIGs focusing more on theoretical and conceptual videos. We found strong correlations between SIG reviews and design document quality, but only for the early formative weeks. High network centrality in the SIG discussion forum social networks was also correlated with higher design document quality.

## 1.4 Research Questions

To continue our analysis of how the design of our course impacted student learning and behavior, and explore how we can conduct an analysis of learning data that corresponds with the learning design, this study will examine the semantic flow of ideas and concepts between the different social levels, with SIGs as our primary unit of analysis. All students began each week with the same set of new ideas delivered through the videos. We will examine how these common ideas became applied to each disciplinary area in the different SIGs, and how this influenced the knowledge work in the lesson design groups, through individual student uptake from SIG-specific forum discussions and the reviews they received from other SIG group members.

A key question will be whether the sub-community in a SIG adds something beyond what could be expected based on a simple correlation between individual student disciplinary interests, and that student's contributions. Our goal is to understand how the nested social structure, and the sub-communities students formed in SIGs, influenced students' discourse and work in Lesson Design groups. We will also look at the difference between SIGs in terms of disciplinary focus and cohesion, and whether this contributed to the quality of the design documents.

## 2 Literature Review

### 2.1 Knowledge Community and Inquiry

Knowledge Community and Inquiry (KCI) is a pragmatic framework for curriculum development to foster knowledge communities, which advocates scripting and coordinated grouping to assure comprehensive distribution across a targeted domain, but adds a layer of collective knowledge building, where students engage with Web 2.0 technologies to develop a shared knowledge base that serves as a resource for their subsequent inquiry [16].

KCI projects are designed explicitly to include inquiry activities that lead to the production of artifacts that allow for assessment of learning on a set of pre-specified goals or expectations. Typically, artifacts are evaluated for coherence (presence of mutually conflicting ideas), and completeness [14]. Many KCI designs feature a group project in which students collaborate throughout the term, with new elements or dimensions added as the students gain access to a larger individual and community knowledge base, and become more conceptually sophisticated [13]. Recent examples include students creating a wiki about human disease and body systems, researching Canada's biodiversity [15], or drafting proposals on how to remedy climate change issues [19].

### 2.2 Grouping and Cohorts in MOOCs

Researchers have looked at forming small groups in MOOCs based on criteria like study habits, time zones, language, learning goals, and collaboration method [20], often aiming to match these characteristics, but in other cases aiming to create

culturally heterogeneous groups [10]. Apart from intrinsic student attributes, researchers have also used data about previous student interactions in a course to form more effective groups [18].

Some unique aspects of our study are the nested social structure, with Lesson Design groups that operate within the social context of a Special Interest Group (co-hort), and also that the students could be stratified very naturally based on teaching interest and age group targeted. Because of the number of collaborative elements that we custom-designed and integrated into the course, the social stratification was also much more wide-reaching than in typical studies, where they have often focused on forum discussions or short video meetings.

### 2.3 Analysis of Text in xMOOC/cMOOC

Forums have been a key focus both in xMOOCs and cMOOCs. In the context of xMOOCs, most of the researchers have used social network analysis (SNA) based variables [4, 8], forum usage statistics [1, 11], and timing patterns [9] to predict grades of the MOOC learners. These methods often use clustering/classification algorithms to cluster/predict the learners' grades. One drawback of such methods is that these methods are used as "black boxes".

On the other hand, in the context of cMOOCs, the main focus is on how learners define their own roles [3], sentiments in the forum posts [2], topic analysis [7], and interaction patterns in the forums [17] to predict/explain the engagement within the MOOC. The primary drawback of these efforts is lack of a universal definition of engagement/dropout, which makes the findings difficult to generalize. In this paper, we present a simple text analysis from a collaborative MOOC to show the relation between the information flow at different social granularities to assess the quality of the artifact produced by each team.

## 3 Methods and Variables

**Coding scheme:** Each Lesson Design group was required to produce a design document with the details of a (possibly multi-hour) lesson that would be taught in their classes. Two authors coded these documents, with an inter-rater reliability of 0.82, according to the coding scheme in Table 1.

**Tf-idf:** For each SIG, we computed the three Term Frequency-Inverse Document Frequencies (Tf-idf), one each for the forum, reviews, and Etherpads. We computed these three tf-idf matrices for every week. The tf-idf value for each term in the matrix denotes two things simultaneously: (1) how important a term is for one document, and (2) how important the term is across the complete set of documents.

**Similarity:** In order to compare the different tf-idf matrices, we computed the cosine similarity between two matrices. The cosine similarity will inform us about the conceptual similarities between the two SIGs, or for the same SIG across forum, reviews, or Etherpads. The similarity value is bounded within the interval

**Table 1.** Coding scheme for design document quality

Code	Description
Learning Objectives (LO)	Level of detail put in the learning objectives mentioned
Activity Design (AD)	Richness in the design of the activities according to the learning objectives
Coherence (CO)	Level of coherence in the various parts of the design document
Innovative use of technology (DT)	Depth of thought put into the innovative use of technology in the design document
Incorporating inquiry-based learning (IB)	The use of inquiry-based learning principles in the design document

(0,1), both values included. A similarity value of zero would depict orthogonal concept spaces, that is, there would be no common themes across those two sets of concepts. On the other hand, a similarity value of one would indicate complete similarity, that is, the two sets of concepts would be the same.

**Betweenness and Withiness:** We computed two types of similarities. The first similarity betweenness is computed between the forums and Etherpad from two different SIGs for every week. The second similarity withinness is computed among the forums, reviews, and Etherpad from the same SIG for every week.

**Uptake of Ideas:** In the present MOOC, the flow of ideas among the participants went in three directions: (1) review to Etherpad; (2) forum to Etherpad; (3) reviews to forum. Every design group received peer feedback on their current state of the design document. This feedback was continuously provided during the course and the peer reviewers were given specific weekly theme-related prompts to suggest improvements to the design documents.

To evaluate the uptake of ideas from the reviews by design groups in the different SIGs we computed the similarity between the reviews they received and the Etherpad (internal group discussion) for the subsequent week. Besides reviews, the design groups also received ideas from the discussions in the forums. We also computed the uptake of ideas from forum using the similarity between forum and Etherpads from the same week.

Finally, to measure the effect of the reviews on the forum discussions, we computed the similarity between the reviews and forums from consecutive weeks.

**Uptake from Videos:** The videos represent a common source of ideas for all SIGs. We computed the similarity between the video transcripts of every week with the forums of every SIG, to evaluate the effect of the information provided by the instructors on the discussions in the different SIGs.

**Case Studies:** The SIGs were designed based on participant interests, however participants chose freely which SIG to join. Since participants differed across multiple dimensions (discipline, age group, etc.), some participants with similar disciplinary focus might have joined different SIGs.

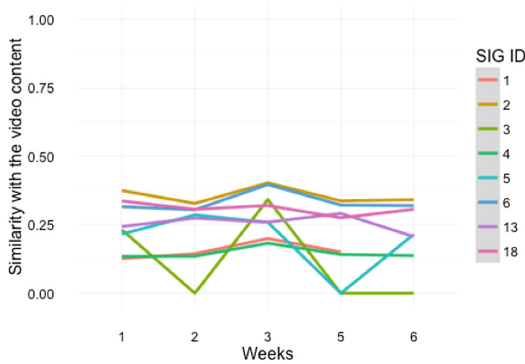
To gauge the effect of the SIG discussion (review, forum, Etherpad) on the individual participants, we extracted all the participants from four specific disciplines (Math, Physics, Chemistry, Biology) from their respective SIGs. We then computed the withinness with their own SIGs and the betweenness with the rest of the SIGs. A higher value of betweenness than withinness will denote that the discussion contributions from individuals are affected by their disciplines; contrary to this, a higher value of withinness will show that the SIG community has a higher effect on the discussions.

### 4 Results and Discussion

Table 4 shows the average betweenness for respectively forums, reviews and Etherpads across all SIGs during the same week, as well as the pair-wise withinness for forums, reviews, and Etherpads (Table 2).

**Table 2.** Average betweenness/withinness for forums, reviews and Etherpads, all SIGs

	Withinness with Reviews	Withinness with Etherpads
Forum (betweenness: $M = 0.50$ ; $sd = 0.12$ )	$M = 0.80$ ; $sd = 0.03$	$M = 0.60$ ; $sd = 0.13$
Reviews (betweenness: $M = 0.48$ ; $sd = 0.12$ )	-	$M = 0.67$ ; $sd = 0.14$
Etherpads (betweenness: $M = 0.22$ ; $sd = 0.07$ )	-	-



**Fig. 4.** Similarity between SIG content and videos per week.

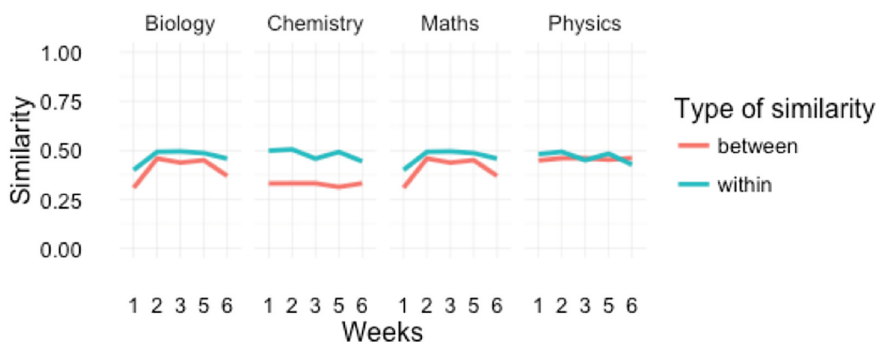
Betweenness for two SIGs is highly explained by their respective similarities with the video content (mean adjusted R-sq 0.76). Figure 4 shows the average weekly similarity between the MOOC-wide videos, and a given SIG (forum, reviews, and Etherpads). From Fig. 4, we can see that most of the SIGs maintain a consistent similarity profile except SIGs 3 (in weeks 2 and 5) and 4 (in week 5). The reason that these three similarity values are zero, is the absence of any activity from the SIGs 3 and 5 during the respective weeks.

#### 4.1 Does SIG Membership Shape the Discourse of Individual Teachers?

We found four disciplines with a large number of teachers, dispersed across multiple SIGs, and tested whether the similarity between teachers from the same discipline (for example physics teachers) were greater than the similarity between a given teacher and his or her SIG (for example Secondary Science). We found that in all four cases, teachers' contributions were significantly more similar to their SIGs, than to teachers with the same disciplinary interests who had joined other SIGs (see Table 3). This shows that the discourse that developed within SIGs informed individual behavior more than what could be explained by looking at individual interests and demographics. Figure 5 shows the development of these relationships for each week of the course.

**Table 3.** Comparing betweenness and withinness for four types of teachers

Similarity $\sim$ Case*Type	df1	df2	F (df1, df2)	p-value
Case (physics, chemistry, biology, maths)	1	62	0.50	0.47
Type (between, within): within discipline < within SIG	1	62	6.14	0.01
Case:Type	1	62	0.16	0.68



**Fig. 5.** Comparing betweenness and withinness for four types of teachers.



## 4.2 Measuring Semantic Diversity of SIGs

We used tf-idf to find the most representative concepts for each SIG (words commonly used in one SIG, and very rarely used in other SIGs). There was a difference between disciplinary-focused SIGs, such as the four listed below, and SIGs focused on a specific age group or audience (museums and informal learning, higher education). The latter SIGs had very few words that were over-represented, suggesting a larger diversity of internal ideas and directions.

Table 4 shows the most representative disciplinary concepts for four discipline-focused SIGs. The number of common terms across SIGs for each week decreases as the course progresses, suggesting that the SIGs become more unified and perhaps more focused on specific applications, and less on the general concepts that unify the course.

**Table 4.** Representative terms from four different SIGs across all weeks

Secondary math	Secondary science	Arts, media and design	Secondary English
Percentage	Enzyme	Atlas	Kinesthetic
Proportion	Hierarchical	Morphology	Invasion
Autograph (math software)	Motion	Art piece	Essence
Geometry	PBL (problem-based learning)	Pastel	Frighten
Circle	Substrate	Watercolor	Marginalization
Symmetrical	Cellular	Pollack	Dramatic
Lag	PhET lab	Storybird	Hannibal
Representation	Respiratory	Van Gogh	Individual
GCF	Ecosystem	Melody	Rome
LCM	Protein	Advocacy	Captivate

## 4.3 Uptake

An important part of the course were the reviews in four weeks of the course, which were disseminated by the participants to their peers scaffolded through the weekly review prompts for each week. In every subsequent week, we found that much of the commonality between the previous weeks' review and SIG discussions and Design groups' Etherpad comments could explain the overall quality of the final design documents. In Table 5 (last two columns), we show the percent of the variance explained of the design document quality ratings by the similarity between reviews a given week and next week's forums and Etherpads respectively. We observe that uptake of reviews in forums is a better predictor of the design document quality than uptake of reviews in Etherpads.

One possible explanation could be that the amount of common knowledge in the forum is much higher than that in the Etherpads; as the whole SIG contributes to the forums, while the Etherpads are specific to one design group.

#### 4.4 Correlation Between SIG Characteristics and Quality of the Design Document

We investigated the correlations between different design document quality metrics (as listed in Table 1), and semantic cohesion. In Table 5, we show the pairwise similarity between Etherpad, reviews, and forums, as well as the individual between similarities between respectively all Etherpads, reviews and forums across SIGs for a given week. We also show the similarity between the forum of a given SIG and the videos of that week (which would indicate idea uptake and focusing on the weekly theme).

The values in Table 5 are the adjusted R-squared of the linear model between the two variables. The dependent variables are the quality ratings and the independent variables are the various similarities. Due to a low number of teams having all the similarity values, we decided to keep the linear models limited to one dependent and one independent variable, thus getting an early estimate of the feature importance for conducting predictions in future. One might argue that we could have used some feature selection mechanisms for reducing the dimensionality of the feature space. Once again, the number of teams ( $n = 8$ ) is not enough to carry out ridge regression. Moreover, it is not less than the number of measures ( $p = 9$ ) so that one could carry out dimensionality reduction suitable for  $n < p$  situations.

We found that Learning Objectives, Design Thinking and Incorporating inquiry-based learning are all very much explained by the video similarity (theme uptake) and the SIG within similarity (cohesion). Activity Design is loosely

**Table 5.** Adjusted R-squared for the five design document quality ratings using the different similarity scores.

	Within similarity			Average between similarity			Similarity of forum w/video	Uptake - Review and the next weeks Etherpads and forums	
	ER	EF	RF	simE	simF	simR	simV	RE	RF
LO	34.1	32	26.3	3.6	4.6	2.7	24.9	7.9	30.9
AD	19.7	15.9	16.1	2.4	4.4	3.9	2.2	4.7	36.5
CO	2.4	4.2	4.6	1.8	3.2	0.1	6.8	2.9	32.6
DT	28.1	23.6	29.8	1.7	19.8	11.6	52.3	2.1	55
IB	23.5	17.8	18.7	4.2	15.1	9.8	59.6	4.7	42.3
Mean quality	21.3	18.9	19.7	1.3	9.8	4.9	32.7	1.9	44

related to within similarity, and for Cohesion, there is no relationship. Activity Design could be seen as more of a measure of individual creativity, and Cohesion is a meta-level indicator.

## 5 Conclusions and Future Work

In this paper, we have presented an analysis of the impact of semantically meaningful cohorts in a unique MOOC. We showed how the ideas discussed in different SIG communities (forums, reviews and Etherpads) were seeded by the weekly videos, which indexed the discussions to the course themes, and were informed by the disciplinary focus of the SIG participants, but were then developed into a coherent discussion that represented something beyond simply a statistical sum of the participants. This can be seen through our analysis of participants with similar disciplinary foci that ended up in different SIGs, and how their expressions of ideas gradually become more similar to the SIG discourse which they are part of, than to the other participants with similar foci in other SIGs.

SIGs and cohorts are an attempt at managing or reducing scale, to avoid overwhelming students, and the large number of students enabled us to form specialized topic-based SIGs in a way that would not have been possible in a small class. However, due to the unequal distribution of interests among students, in what was primarily marketed as a course for STEM K-12 teachers, some of the SIGs were quite specialized around certain disciplines, and others had to group together a number of related disciplines to get a large enough critical mass to support discussions and knowledge work. We have shown that the more specialized SIGs have a higher level of withinness, and are also correlated with a higher quality of the final design documents, perhaps because the forum discussions and reviews were more relevant to the design group efforts.

Student interests have several dimensions, and grouping students in cohorts necessarily prioritizes a subset. We could imagine a physics teacher, working in a high-school, and interested in 3D printing. While she might be grouped with other physics teachers, she would lose out on the comments by the 3D-printing enthusiast in the chemistry SIG. While our analysis has shown the added value of having stable communities whose discourse develops in a coherent manner, future studies could investigate the use of semantic tags (on both participant profiles and content), or text analysis, to promote idea exchange across SIGs.

## References




1. Cobo, G., et al.: Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 248–251. ACM (2012)
2. Dmoshinskaia, N.: Dropout prediction in MOOCs: using sentiment analysis of users' comments to predict engagement. Master's thesis, University of Twente (2016)

3. Dubosson, M., Emad, S.: The forum community, the connectivist element of an xmooc. *Univ. J. Educ. Res.* **3**(10), 680–690 (2015)
4. Fancsali, S.: Variable construction and causal modeling of online education messaging data: initial results. In: *Educational Data Mining 2011*. Citeseer (2010)
5. Håklev, S., Sharma, K., Slotta, J., Dillenbourg, P.: Contextualizing the co-creation of artefacts within the nested social structure of a collaborative MOOC. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017*. LNCS, vol. 10474, pp. 67–81. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_6](https://doi.org/10.1007/978-3-319-66610-5_6)
6. Håklev, S., Slotta, J.D.: A principled approach to the design of collaborative MOOC curricula. In: Delgado Kloos, C., Jermann, P., Pérez-Sanagustín, M., Seaton, D.T., White, S. (eds.) *EMOOCs 2017*. LNCS, vol. 10254, pp. 58–67. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-59044-8\\_7](https://doi.org/10.1007/978-3-319-59044-8_7)
7. Joksimović, S., et al.: What do cMOOC participants talk about in social media?: A topic analysis of discourse in a cMOOC. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, LAK 2015*, pp. 156–165. ACM, New York (2015). <https://doi.org/10.1145/2723576.2723609>
8. Joksimović, S., Manataki, A., Gasevic, D., Dawson, S., Kovanovic, V., De Kereki, I.F.: Translating network position into performance: importance of centrality in different network configurations. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 314–323. ACM (2016)
9. Khan, T.M., Clear, F., Sajadi, S.S.: The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 226–229. ACM (2012)
10. Kulkarni, C., Cambre, J., Kotturi, Y., Bernstein, M.S., Klemmer, S.: Talkabout: making distance matter with small groups in massive classes. In: Plattner, H., Meinel, C., Leifer, L. (eds.) *Design Thinking Research*. UI, pp. 67–92. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-19641-1\\_6](https://doi.org/10.1007/978-3-319-19641-1_6)
11. Lopez, M.I., Luna, J., Romero, C., Ventura, S.: Classification via clustering for predicting final marks based on student participation in forums. In: *International Educational Data Mining Society* (2012)
12. Mor, Y., Ferguson, R., Wasson, B.: Learning design, teacher inquiry into student learning and learning analytics: a call for action. *Br. J. Educ. Technol.* **46**(2), 221–229 (2015)
13. Najafi, H.: Transforming learning in science classrooms: a blended knowledge community approach. Ph.D. thesis, University of Toronto (2012)
14. Peters, V.L., Slotta, J.D.: Analyzing collaborative knowledge construction in secondary school biology. In: *Proceedings of the 9th International Conference of the Learning Sciences*, vol. 1, pp. 548–555. International Society of the Learning Sciences (2010)
15. Peters, V.L., Slotta, J.D.: Scaffolding knowledge communities in the classroom: new opportunities in the web 2.0 era. In: Jacobson, M., Reimann, P. (eds.) *Designs for Learning Environments of the Future*, pp. 205–232. Springer, Boston (2010). [https://doi.org/10.1007/978-0-387-88279-6\\_8](https://doi.org/10.1007/978-0-387-88279-6_8)
16. Slotta, J.: Knowledge Community and Inquiry. Paper presented and published for the Network of Associated Programs in the Learning Sciences (NAPLES). Technical report (2014)
17. Wang, Z., Anderson, T., Chen, L., Barbera, E.: Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework. *Br. J. Educ. Tech.* **48**(2), 683–699 (2017)

18. Wen, M., Maki, K., Wang, X., Dow, S., Herbsleb, J.D., Rose, C.P.: Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. In: EDM, pp. 533–538 (2016)
19. Zhao, N., Najafi, H., Slotta, J.D.: An analysis of teacher-students interactions in three science classes: a pilot study. In: Proceedings of the Ninth International Computer- Supported Collaborative Learning Conference, Hong Kong. International Society of the Learning Sciences (2011)
20. Zheng, Z., Vogelsang, T., Pinkwart, N.: The impact of small learning group composition on student engagement and success in a mooc. In: Proceedings of the 8th International Conference of Educational Data Mining, pp. 500–503 (2015)



# Detecting Learning Strategies Through Process Mining

John Saint<sup>1,2</sup> , Dragan Gašević<sup>1,3</sup> , and Abelardo Pardo<sup>4</sup> 

<sup>1</sup> University of Edinburgh, Edinburgh, UK  
john.saint@ed.ac.uk

<sup>2</sup> Regents University London, London, UK

<sup>3</sup> Monash University, Melbourne, Australia

<sup>4</sup> University of South Australia, Adelaide, Australia

**Abstract.** The recent focus on learning analytics to analyse temporal dimensions of learning holds a strong promise to provide insights into latent constructs such as learning strategy, self-regulated learning, and metacognition. There is, however, a limited amount of research in temporally-focused process mining in educational settings. Building on a growing body of research around event-based data analysis, we explore the use of process mining techniques to identify strategic and tactical learner behaviours. We analyse trace data collected in online activities of a sample of nearly 300 computer engineering undergraduate students enrolled in a course that followed a flipped classroom pedagogy. Using a process mining approach based on first order Markov models in combination with unsupervised machine learning methods, we performed intra- and inter-strategy analysis. We found that certain temporal activity traits relate to performance in the summative assessments attached to the course, mediated by strategy type. Results show that more strategically minded activity, embodying learner self-regulation, generally proves to be more successful than less disciplined reactive behaviours.

**Keywords:** Learning analytics · Process mining · First order Markov models  
Temporal dynamics · Self-regulated learning

## 1 Introduction

Enhancing learning experience is one of the primary goals for many higher education institutions. Approaches such as flipped classrooms offer some promise of advancing student academic performance and satisfaction [1]. However, the emphasis on the self-directed use of technology to complete learning activities increases a need for students to have high skills for self-regulated learning. Poor choices of study tactics and strategies are often reported in the literature, through the collection of student self-reports. Although such approaches can offer some insights to the ways students study, they offer little information that can be used by educators to offer guidance to students in real-time.

The development of the field of learning analytics promises to provide insights into learning strategies by analysis of trace data about students' use of and interaction with online resources provided in learning management systems (LMS). Machine learning

techniques have been used to explore trace data sequences to reveal distinct strategies and approaches to learning e.g., [2–4]. Nonetheless, a section of these studies uses statistical methods and focus more on engagement frequency/categorisation where the dimension of time (critical to this study) is not considered e.g., [5, 4]. Others recognise time as a dimension, but this is restricted to measurement of time on task, and not a reflection of true inter-process temporal dynamics e.g., [6]. Another section of studies provides key insights into learner engagement over time, as opposed to comparative, stochastic inter-strategy analyses e.g., [7].

This paper reports on the findings of a study that was set out to explore the extent to which process mining techniques can provide insights into learning strategies provided by current approaches based on machine learning methods. Specifically, the study used first-order Markov chains to complement the findings of an existing method, based on machine learning, to examine internal dynamics of learning strategies and perform inter-strategy comparison in terms of the temporal sequencing of individual activities can be performed. The results showed that proposed approach provides a genuine insight into inter and intra-tactic dynamics, providing a different dimension to the narrative around learning strategy presently reported in the literature. The study also provides a view of learning malformation as typified by movement through and between study actions.

We use first order Markov models (FOMMs) as an initial exploratory process-mining algorithm with a view to testing their viability as an interpretive tool for learning sciences. FOMMs are based on transition probabilities between sets of processes. It is proposed that this type of stochastic insight combines effectively with the process activity formulation described in the methodology section.

## 2 Background and Related Work

### 2.1 Learning Strategy

The utilisation of effective study strategies is an important factor of effective self-regulated learning (SRL), as is the awareness of the relationship between these strategies and the aspired outcomes [8]. As stated by Boekaerts, self-regulated learners are “...aware of what they know and feel about the domain of study, including which general cognitive and motivation strategies are (less) effective to attain the learning goals...” [9]. Accordingly, they are aware of the attributes of their own knowledge, motivations, beliefs, expectations, and cognitive behaviours, and seek to reapply ongoing task-oriented mediation, in keeping with their defined goals and standards [10]. However, the standards learners use for evaluation of the choices of their learning strategies and products of their learning can be suboptimal. Winne and Noel-Jamieson showed that learners generally overestimate their use of individual study tactics [11]. Bjork et al. [12] suggest that learners mostly use ineffective study strategies – e.g., reading and re-reading text instead of practising memory recall through self-testing. The challenge, therefore, is to determine an effective analytical method of capturing and measuring the choices of study strategies and tactics to enhance the effectiveness of learners’ self-regulation. Study tactics and strategies are closely related concepts. Winne [13] characterises a set of tactics and strategies, as well as an overarching sense of metacognition employed in the learning process. In doing so, he identifies three key aspects

of SRL. A tactic can be viewed as an if-then construct, e.g., *if* I read an article which confirms an aspect of my theory *then* I will add to my corpus. We could extrapolate this to include *else* e.g., *else* I will seek to refine my theory. A strategy is structured arrangement of cognitive tactics. Finally, metacognition is a learner's management of their own cognitive strategies, and the development of an overarching knowledge management strategy, encompassing self-awareness.

## 2.2 Analytics of Learning Tactics and Strategies

The use of trace data to study learning strategies has been galvanised through the foundation of the field of learning analytics. Several authors proposed the use of unsupervised methods for the study of learning strategy. Lust et al. [5] used clustering to identify user-profiles through learner behaviours, identifying profiles through frequency of activity engagement of content management system supported course. In an attempt to add a temporal dimension, Lust et al. [14] augmented their research with an analysis to identify changes in learner strategies between the first and second half of the course. Similarly, Kovanović et al. [6] use a hierarchical cluster analysis to extract learning strategies of learners and to understand the extent to which those strategies were associated with the learners' level of cognitive presence in online discussions. Although the results of these studies are relevant for understanding the connection between learning strategy, academic performance, and cognitive presence, these studies offer little insight into how learners sequence their activities with each of the strategies identified. Thus, learning strategies are looked at as summaries of the quantities of activities rather than temporally sequenced activities based on some strategic choices.

Analysis of temporal links between actions learners take has also been used in the literature on learning strategy. Kinnebrew et al. utilised a computer-based learning environment to measure students' cognitive and meta-cognitive development using sequence mining techniques [15, 16]. Jovanović and her colleagues [3] utilise a combination of an unsupervised machine learning technique with a sequence mining algorithm to explore the extent to which meaningful learning strategies can be extracted from trace data. Their follow-up study showed that learning strategies extracted from trace data are associated with deep and surface approaches to learning [2]. Fincham et al. [7] extract study tactics by using hidden Markov models and then apply a clustering exercise, which partially mirrors Jovanović et al. [3], to extract study strategies. Both Jovanović et al. and Fincham et al. studies found that such the use of learning strategies extracted this way was associated with academic performance. While these studies provide key insights into learner engagement over time, they fall short of providing comparative inter-strategy analyses.

Process mining techniques provide viable tools for comparative inter-strategy analyses, though these methods are typically used on think aloud data. Bannert and her colleagues [17] use process mining techniques to analyse think-aloud data logged from a student-group's navigation through an LMS. The think aloud data were coded for presence of micro-level processes of SRL (e.g., goal-setting) and analysed with the Fuzzy Miner process mining algorithm to compare differences in SRL between high and low performing students. In Sonnenberg and Bannert's follow-up study [18], the same methods are used to measure the impact of metacognitive prompts in similar LM environments.



These studies are significant in that they present a novel way of capturing and measuring SRL on the level of SRL micro-level processes. The studies, however, do not provide insights into learning strategies followed by learners while using an LMS to study.

### 3 Methodology

#### 3.1 Data Collection

The data for this study were collected from an LMS attached to a computing course at a university in [anonymised]. The course was based on a flipped classroom pedagogy and the data used in this study were about students' engagement with the online activities, which served the purpose of preparation for the face-to-face activities. Each time a student engaged with an element of the LMS, a learning event record was generated containing a *student ID* number, a *timestamp*, and the completed *study action*. The study actions were: watching video; reading textual content; response to summative problem-solving exercise along with information about correct and incorrect responses; response to a question from formative quizzes with information about correct and incorrect responses and whether the students asked to see the correct response; dashboard view, and view of lesson objectives. The student cohort consisted of 290 students who collectively generated 184,211 learning events. The course lasted 13 weeks, comprising two main bouts of activity: Weeks 2 to 5 and 7 to 12. In week 6, the students completed a summative mid-term assessment, and in week 13 a final exam. It is crucial to note that successful completion of summative assessment tasks contributed to 10% to the overall module mark. Scores from mid-term and final exam are also used for analysis.

To understand how students managed individual study actions, we added, for each study action, the following four attributes about time management: *preparing* – completing an action on a topic in the designated week; *revisiting* – completing an action on a topic introduced a previous week, having completed the action in the previous week; *catching up* – completing an action on a topic after the week in which that topic was introduced for the first time; and *ahead* – completing an action on a topic ahead of the designated week. This provides an insight into the access timing of the study actions and therefore time management of student tasks.

#### 3.2 Data Analysis

**Extraction of Learning Tactics and Strategies.** The work carried out in [3] is of primary importance to this study. It provides a method for automated extraction of learning tactics and strategies from trace data about students' interaction with online resources. The method was composed of two levels of analytics based on unsupervised machine learning methods – i.e. clustering. Firstly, learning tactics were extracted by analysing study sessions. These sessions were delineated by temporal gaps; a simple example would be a group of study actions beginning and ending in a twenty-minute period. If we observe a gap of more than one-hour between the last action of this period and the start of another action sequence, then we can define it as a session. These sessions were clustered based on similarity of the actions performed by the students. Exploratory

sequence analysis was implemented using TraMinerR R library [19] and followed up with a hierarchical cluster analysis with Levenshtein distance and Ward's method, as proposed in [3]. This generated four strategy types, based on the predominant study action type: *reading course materials*, *formative assessment*, *video viewing with associated formative assessment*, *reading course materials*, and *summative assessment*. Secondly, learning strategies were extracted through an agglomerative hierarchical clustering with Euclidian distance and Ward's method, based on the frequency of the use of the four study tactics by each individual student in the sample. This analysis identified five learning strategies (also referred as strategy groups) which provided insight into how students sequenced individual study actions within each of the strategy groups. These strategy groups, integral to the study in [3], had a significant part to play in the current study. The strategy groups in this study differ slightly from those in the study [3] as we removed single-event sessions from the dataset. This affected sequence clusters and propagated to strategy groups.

**Process Mining.** PM seeks to capture event or process-based data. The starting point of PM is a dataset in the form of an event log. The required elements to run a PM algorithm are:

- **Case:** a process instance. This could represent a human actor, or a more abstract construct, such as a learning cycle. In our study, student ID was the case role.
- **Activity:** a well-defined step in a broader process. In our study, concatenation of strategy types and time management attributes was used, e.g., Formative Assessment & Catch-up, or Summative Assessment & Preparation
- **Timestamp:** ideally one for the beginning and the end of the activity, but more usually just one stamp is available. Timestamps of activities in our trace data were used.

In this sense, trace data supply raw material for examining learning processes. Traditional frequency-based analytic methods do not adequately reflect these learning processes as they flow and change over time. The selection of model discovery algorithm is key. Out of the traditional algorithms: we rejected Heuristic Miner because it is more suited to processes with fewer event types than we have; we rejected Multi-phase miner as it is more suitable for cleanly structured, simple log data (unlike ours); Fuzzy Miner produces interesting overviews of learning processes but does not provide the crucial stochastic metrics we seek to use [20]. We chose FOMMs to explore the novel possibility of combining stochastic analysis and temporal event data [21]. We employ the R package pMineR [21, 22] to train and generate FOMMs based on the learner strategy groups extracted in the procedure as previously explained. The pMineR package provides FOMM visualisations and probability transition matrices which allow analysis and comparison of temporal patterns of process engagement. Examining these patterns provides some insight in the tactical differences between the identified strategy groups in relation to SRL traits.

**Strategy Group Characterisation – Intra-Strategy Group Analysis.** FOMMs were trained and generated for each strategy group. Characterisation is informed by Winne's

construction of learner strategy and study tactics, as articulated by Fincham [7]. We provide an interpretive narrative for each group, and then characterise them accordingly.

**Strategy Group Comparison – Inter-Strategy Group Analysis.** We first identified significantly distinct strategy groups by assessment performance. We undertook pairwise comparisons based on mid-term scores and by final assessment (see Table 2). As ANOVA assumptions were not satisfied, we undertook a Kruskal Wallis test, followed by pairwise Mann Whitney U tests, using False Discovery Rate (FDR) to accommodate alpha inflation. From our pairwise analysis, we elected to compare two pairs of strategy groups. Firstly, we chose only pairs that demonstrate statistical difference in assessment means. From these pairs, we made a valued assessment on the most potentially insightful comparisons, based on high versus low mid-term/final exam performances. To provide comparative insights, we interpreted the comparison diagrams of two pairs of strategy group FOMM models. In each case one strategy group is mapped onto another group (see Fig. 2). The arcs in black represent similar transition probabilities (TPs). Red arcs represent a comparatively lower TP of the mapped model; green arcs represent a higher TP. In cases of disparate TPs, both probabilities are shown. To simplify presentation, a TP threshold of 0.05 is has been set.

## 4 Findings

The findings present the intra-and inter-strategy group analysis performed by using the FOMM. Due to the size of the diagrams representing the final FOMMs, this section includes only excerpts of the main FOMM diagrams. Complete results of the FOMM analysis can be found here:

<https://www.dropbox.com/s/yqtw20uwiwbnmob/FOMM%20Results.pdf?dl=0>.

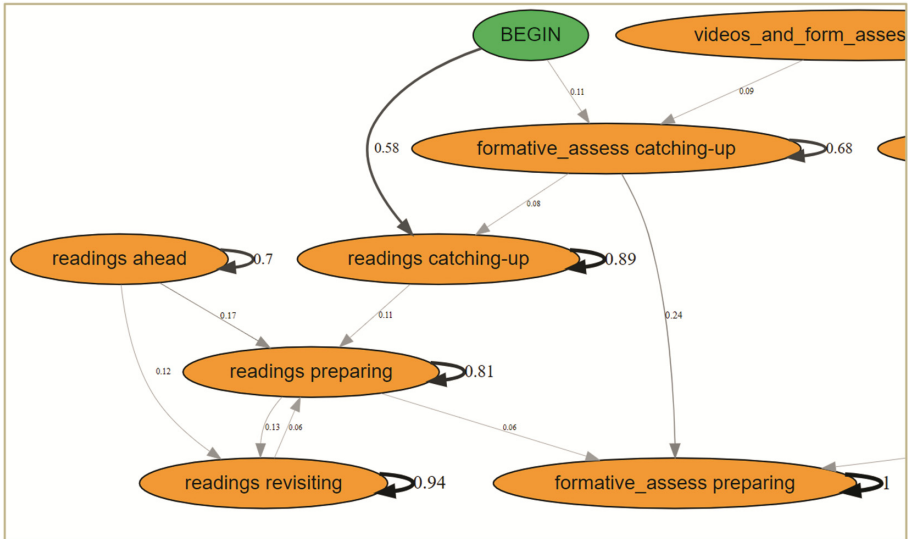
### 4.1 Strategy Group Characterisation: Intra-Strategy Analysis

The strategy extraction method proposed in [3] identified five strategies, also referred to as strategy groups i.e. they represent groupings of the students based on similarities of their learning strategies. By way of context, Table 1 shows the mean and median sample scores for each strategy group, and a measure of group activity i.e. number of events divided by the group sample size.

**Table 1.** Strategy group assessment scores

Strategy group	n	Mean mid-term score	Median mid-term score	Mean final assessment score	Median final assessment score	Events per student
1	19	15.3	15	24.7	24	1634
2	70	14.9	16	22.5	20	1295
3	117	12.9	13	17.4	15	986
4	25	15.5	16	23.7	25	1737
5	59	10.7	11	14.6	14	576

**Strategy Group 1.** This is a relatively well-performing and active group. Figure 1 shows a section of this group’s FOMM, relating to content access. It demonstrates a temporally cohesive approach to the reading tasks. The students, when they are engaged in reading tasks, tend not to get distracted by other activities. There is clear interplay between the four temporal instances of reading activity. Reasonably enough, in some cases reading preparation leads to formative preparation. This is a manifestation of well-formed study patterns. There is a demonstration of movement from video formative assessment and formative assessment in terms of temporal groupings. For example, there is 0.09 chance that students will, on completion of video catch-up session, move to a non-video formative catch-up session. Summative tasks present a neater temporal grouping. Students are likely to stick within this activity group e.g., students are more likely, once they decide on a summative activity, to stick with, or move between time-contextual iterations of the summative task e.g., between summative assessment catch-ups to revisits, or between summative ahead to preparation. In summary, this group show elements of cohesive learning but also a tendency to embrace multiple activity types. In this sense, the students represent an **Active Agile** strategy group.



**Fig. 1.** Partial first order Markov model of strategy group 1 (Active Agile)

**Strategy Group 2.** This group is less active than group 1, and assessment scores suggest an engagement drop-off in the second half of the course. Nonetheless, this group displays a similarly cohesive approach to reading tasks. Formative video tasks are partially associated with certain reading activities; there is a tendency to touch on these video tasks before reading catch-up and preparation. This could represent an attempted strategy to streamline knowledge acquisition through video, before falling back on traditional content access. It demonstrates a regulation of cognitive learning tactics and a broader self-regulatory learning strategy. Formative assessment activities are grouped

temporally, so students do not tend to move out of formative cycles once started. They do not tend to move freely within the summative groupings, aside from a movement between catch-up and preparation. Aside from this, once a summative task is attempted, it is pursued almost without distraction. This group can be typified as **Efficient**.

**Strategy Group 3.** This group shows less engagement with all activities. There is a greater likelihood to attempt the main summative activity without adequate preceding formative preparation. This group's approach points to a minimalist strategy, with inherent gambles on summative success. This group's FOMM diagram highlights a movement to summative and reading revisits after several reading activities. This could indicate a less proactive approach to advanced reading preparation, hinting at a reaction to poor performance in the summative tests. This still indicates regulation of tactics but potentially a less effective learning strategy. This group can be typified as **Summative Gamblers**.

**Strategy Group 4.** This is a strong and active group. It presents a healthy and cohesive approach to preparatory work. In fact, it presents the tightest adherence to activity focus in the sense of the activity self-loops. The students in this group do not tend to move freely from one activity type to another, or even from one activity to another. There is a real sense of disciplined engagement. Interestingly, this group favours video formative assessments more than others, and shows tendencies to engage in focussed video preparation and catch-up tasks. This could indicate a desire to streamline learning using more varied media, in combination with traditional knowledge acquisition tactics. In formulating the best combination, learners are assessing their own comprehension of knowledge, and adjusting to fit. This group is typified as **Active Cohesive**.

**Strategy Group 5.** This is the least active group, and the weakest performer. Apart from the overemphasis on summative assessment without preparation, there is a distinct lack of strategic cohesion. We see a tendency to bounce from activity (type) to activity (type). The exception to this is the formative activity grouping, where there is a semblance of temporal coherence. It is difficult to determine whether this group represents strategic incoherence, or that the collective paucity of engagement data provides inconsistent results. This group exhibits non-ideal navigation through its learning environment. The group is typified as **Extreme Minimalists**.

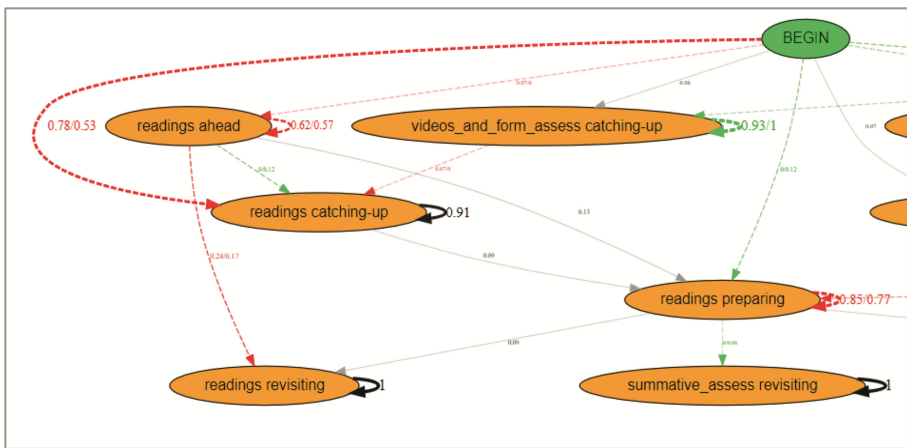
## 4.2 Strategy Group Comparison: Inter-Strategy Analysis

A pairwise comparison of the five strategy groups on mid-term and final examination scores is reported in Table 2. We use this to inform choices of pairs in our comparative analysis.

**Table 2.** Pairwise comparison of assessment scores

Mid-term Scores					Final exam scores				
G1	G2	Z	P	r	G1	G2	Z	p	r
2	5	4.3526	0.00001*	0.3832	2	5	4.3792	0.00001*	0.3856
4	5	3.5534	0.00019*	0.3877	1	5	3.9427	0.00004*	0.4464
1	5	3.3147	0.00046*	0.3753	2	3	3.6004	0.00016*	0.2633
2	3	3.3147	0.00046*	0.2424	1	3	3.4207	0.00031*	0.2933
4	3	2.9387	0.00165*	0.2466	4	5	3.4207	0.00031*	0.3732
1	3	2.4740	0.00668*	0.2121	4	3	2.4227	0.00770*	0.2033
3	5	2.2516	0.01217*	0.1697	3	5	1.8368	0.03312*	0.1385
4	2	0.2475	0.40227	0.0254	1	2	0.4706	0.31897	0.0499
4	1	0.3806	0.64826	0.0574	1	4	0.5112	0.69538	0.0771
1	2	0.5730	0.71669	0.0607	4	2	0.5112	0.69538	0.0524

**Comparative Analysis: Efficient (2) and Summative Gamblers (3).** In this comparison, the *efficient* group are significantly better performers than the *summative gamblers*, based on both midterm and final exam scores. Figure 2 presents a partial example of the comparison diagram for this case.



**Fig. 2.** Partial FOMM comparison diagram: Efficient vs Summative gamblers (Color figure online)

**Reading Activities.** The efficient group demonstrate a greater emphasis on initial reading tasks. The initial TP of 0.78 for reading-catch-up sessions (versus 0.53 for the summative gamblers) points to a greater awareness of the value of preparatory content-based activity. Both groups display a similar self-loop TP of around 0.9 reading catch-ups. The gamblers are more likely to break out of a reading-ahead session to attempt a reading catch-up session (0.12). They are however, less likely to break out of reading revisit sessions (0.24/0.17). This points to a slightly more considered approach to reading

strategy by the efficient group. In the reading preparation task, the efficient group show a higher self-loop TP than the gamblers (0.85/0.77), whereas the gamblers show a more likely propensity to attempt a summative revisit whilst doing this task. This shows that the efficient group are more focussed on the reading task in hand.

**Formative Assessment.** The efficient group demonstrate higher self-loop TPs for formative ahead (0.83/0.63) and catch-ups (0.9/0.83). The gamblers are more likely to break out of these task loops to try formative preparation and revisits. Again, this points to a slightly more considered approach to task management by the efficient group. The gamblers demonstrate a slightly more scattergun approach in this case. Both groups exhibit a strong self-loop focus on formative preparation and revisiting.

**Video Formative Assessment.** Interestingly, the efficient group demonstrate a similar video ahead self-loop. They are however, more likely to break out of this loop to do video preparation (0.22/0.07) and/or reading preparation (0.11/0). The gamblers are more likely to break out to revisit video assessment (0.2/0) and/or attempt a summative assessment ahead of schedule (0.07/0). Again, we can infer that the efficient group are slightly more mindful of preparatory strategies, as befits a self-regulated learner.

**Summative Assessment.** This is, by far, the most popular activity, as it relates to achievable marks on the course. The key point of interest is that gamblers are more likely to attempt this initially, without any other preparation, than efficient members (0.07/0). Regarding catch-up, efficient members are more likely to break out from this loop (0.1/0) to do the main summative preparation activity. This indicates that the efficient group are more likely to move between weekly summative assessments and to tie up loose ends, assessment-wise. This demonstrates a strong sense of self-regulation, as they recognise potential gaps in their understanding that require extra work.

**Comparative Analysis: Active Cohesive (4) and Extreme Minimalist (5).** In this comparison, the active cohesive group are significantly better performers than the extreme minimalists, based on both mid-term and final exam scores.

**Reading Activities.** The cohesive group display a healthy regard for reading activities, as can be seen by the initial activity TPs. This group is nearly as half as likely to embark on an initial reading activity as any other, with a combined TP of 0.48 for preparation and catch-up. The minimalist group's likelihood of starting with a reading activity is 0.28 (specifically catching-up). The weekly-current preparation task is approached differently by the two groups. The minimalist group tends to approach it in isolation, whereas for the cohesive group it provides a valid option from various states: Begin 0.09, video catch-up 0.15, reading catch-up 0.14, reading ahead 0.15. This differs from the normal behaviour of this group but indicates an ongoing focus on this task. In terms of the preparation, the cohesive group maintains a tighter self-loop (0.91), whereas the minimalist group is more likely to move off to other tasks (0.77).

**Formative Assessment.** The cohesive group displays a more considered temporal focus. There is a greater tendency to engage consistently with the formative task in hand,

as highlighted by the higher self-loop TPs around the four formative activities (between 0.9 to 1). We see the minimalist group moving more freely between catch-ups, revisits, and preparation, indicating a less disciplined approach to formative learning. For example, the minimalist group has TP of 0.11 in moving from catch-up to revisiting. It also has a TP of 0.07 in moving from preparation to revisiting. The cohesive group has a TP of 0 in both cases. Temporally, the cohesive group sticks to its formative task groups more closely with less jumping between the week-specific material. This could indicate a different emphasis on controlled, strategy-driven learning.

**Video Formative Assessment.** The cohesive group places more stock in the use of video assessments, particularly preparation and revisits. They are more likely to transition to these activities from other activities, than the minimalist group. Once engaged with these tasks, the cohesive group does not tend to divert, with self-loop TPs of 1 for the two most popular video tasks. The minimalist group approaches these tasks more in isolation. That being the case, they do retain strong self-loops.

**Summative Assessment.** As in the previous comparison, there are differences in the lead-up to this key activity. The minimalist group is much more likely to attempt this as an initial task (0.25/0.09 for prep, 0.23/0 for catch-up), whereas the cohesive group explores content access and preparatory formative activity first. Regarding the minimalists, it is interesting to note that the main summative preparation task could be a destination from several other activities: reading ahead (0.06); reading preparation (0.07), and summative revisits (0.05). For the cohesive group, this task is done more in isolation, apart from as a destination from one task. The cohesive group treat the summative task as a more significant event in and of itself. Both groups, once engaged in the task, retain a tight self-loop.

## 5 Discussion and Conclusion

**Self-Regulation and Summative Tasks.** As previously reported in [3], summative tasks dominate the main activity cycles (as successful completion contributes to the final overall module mark). Using FOMMs, we can gain insights into strategic navigation around the other activity types in the context of these summative main tasks. The two strongest groups, Active Agile and Active Cohesive both demonstrate a healthy regard for pre-summative preparation and engage in more content access and formative assessment before engaging in the summative tasks. Interplay between such states indicates a healthy self-regulatory strategy. In context of the other notable studies that analyse this data [3, 7], this study provides a genuine insight into inter and intra-tactic dynamics, providing a different dimension to the narrative around learning strategy.

**Summative Gambling.** Conversely, the weaker groups exhibit a greater tendency to attempt the summative work without commensurate preparation. There seems to be an underlying attempt to by-pass traditional patterns of self-regulation and gamble on success in the summative tasks. This is a gamble which does not appear to pay off. We also see reactive outcomes in the groups' relationship with catch-up and revisits to past



material. This indicates a more passive, yet performance avoidance, goal-oriented regulation strategy. Whereas previous studies have provided a characterisation of weaker performing groups [3–5, 7, 14], our study provides a view of learning malformation as typified by movement through and between study actions. We therefore have a temporal context.

**Transition Probability Self-loops.** Activity self-loops provide insight into temporal adherence to tasks. It is too simplistic to say that higher self-loop TPs indicate academic discipline. Movement between tasks and task groups can indicate assured self-regulation in learning tactics. Nonetheless, we see that disparate task engagement does seem to indicate a lack of academic focus. This is more apparent in the weaker student groups. Again, through analysing activity engagement patterns, we can pick up on measures of learner focus or lack thereof. This dimension is unseen in previous studies.

**Performance-based Analysis.** There are interpretable differences between higher and lower performing strategy groups. In this sense, we can say that the method can highlight effective versus less-effective learning strategies. Discernible patterns, such as those found in the clustered groups, do appear to exist. This reinforces the need to use effective non-supervised machine learning techniques in studies of this nature. In this sense we are not advancing insight on the fact that we can detect performance differences. Previous studies have linked strategy to performance [3, 7], so in a sense this corroboration provides partial validation of the method.

**Implications for Practice.** This is the first use of a process mining method in combination with unsupervised and sequence mining methods to understand learning strategy. In exploring temporal inter-process dynamics, we have the potential to identify positive and negative instances of learning strategy management. In instances of malformed student learning, interventions and remedial actions are a possibility. We also have the possibility to measure idealised models of student learning against recorded models to inform course design; if we detect weak engagement points in the model, it may indicate weaknesses in course design.

**Limitations and Future Direction.** The study does not provide a set of benchmark metrics for analysis, so its generalisability and replication value cannot be ascertained until more similar studies are undertaken. The option to compare high vs low performers regardless of strategy group, or first half of term vs second half of term, was not explored. This may have provided more crucial strategic insights than the comparison of strategy groups alone. These options will be explored in the next cycle of analysis. FOMMs, by their very definition, provide transition probabilities based on the current event, and therefore lack event “memory”. We are keen to build on this research and explore higher order Markov models, hidden Markov models, and other related techniques, such as conditional random fields.

## References

1. O'Flaherty, J., Phillips, C., Karanicolas, S., Snelling, C., Winning, T.: The use of flipped classrooms in higher education: a scoping review. *Internet High. Educ.* **25**, 85–95 (2015). <https://doi.org/10.1016/j.iheduc.2015.02.002>
2. Gašević, D., Jovanović, J., Pardo, A., Dawson, S.: Detecting learning strategies with analytics: links with self-reported measures and academic performance. *J. Learn. Anal.* **4**, 113–128 (2017). <https://doi.org/10.18608/jla.2017.42.10>
3. Jovanović, J., Gašević, D., Dawson, S., Pardo, A., Mirriahi, N.: Learning analytics to unveil learning strategies in a flipped classroom. *Internet High. Educ.* **33**, 74–85 (2017). <https://doi.org/10.1016/j.iheduc.2017.02.001>
4. Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., Adesope, O.: Analytics of communities of inquiry: effects of learning technology use on cognitive presence in asynchronous online discussions. *Internet High. Educ.* **27**, 74–89 (2015). <https://doi.org/10.1016/j.iheduc.2015.06.002>
5. Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J., Clarebout, G.: Tool-use in a blended undergraduate course: in search of user profiles. *Comput. Educ.* **57**, 2135–2144 (2011). <https://doi.org/10.1016/j.compedu.2011.05.010>
6. Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R.S., Hatala, M.: Does time-on-task estimation matter? Implications for the validity of learning analytics findings. *J. Learn. Anal.* **2**, 81–110 (2015). <https://doi.org/10.18608/jla.2015.23.6>
7. Fincham, O.E., Gasevic, D.V., Jovanovic, J.M., Pardo, A.: From study tactics to learning strategies: an analytical method for extracting interpretable representations. *IEEE Trans. Learn. Technol.* 1–13 (2018). <https://doi.org/10.1109/tlt.2018.2823317>
8. Zimmerman, B.J.: A social cognitive view of self-regulated academic learning. *J. Educ. Psychol.* **81**, 329–339 (1989). <https://doi.org/10.1037//0022-0663.81.3.329>
9. Boekaerts, M.: Self-regulated learning: a new concept embraced by researchers, policy makers, educators, teachers, and students. *Learn. Instr.* **7**, 161–186 (1997). [https://doi.org/10.1016/S0959-4752\(96\)00015-1](https://doi.org/10.1016/S0959-4752(96)00015-1)
10. Butler, D.L., Winne, P.H.: Feedback and self-regulated learning: a theoretical synthesis. *Rev. Educ. Res.* **65**, 245–281 (1995). <https://doi.org/10.3102//00346543065003245>
11. Jamieson-Noel, D., Winne, P.H.: Exploring students' calibration of self reports about study tactics and achievement. *Contemp. Educ. Psychol.* **27**, 551–572 (2002). [https://doi.org/10.1016/S0361-476X\(02\)00006-1](https://doi.org/10.1016/S0361-476X(02)00006-1)
12. Bjork, R.A., Dunlosky, J., Kornell, N.: Self-regulated learning: beliefs, techniques, and illusions. *Annu. Rev. Psychol.* **64**, 417–444 (2013). <https://doi.org/10.1146/annurev-psych-113011-143823>
13. Winne, P.H.: A metacognitive view of individual differences in self-regulated learning. *Learn. Individ. Differ.* **8**, 327–353 (1996). [https://doi.org/10.1016/S1041-6080\(96\)90022-9](https://doi.org/10.1016/S1041-6080(96)90022-9)
14. Lust, G., Elen, J., Clarebout, G.: Regulation of tool-use within a blended course: student differences and performance effects. *Comput. Educ.* **60**, 385–395 (2013). <https://doi.org/10.1016/j.compedu.2012.09.001>
15. Kinnebrew, J.S., Biswas, G.: Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In: *Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012)*, pp. 57–64 (2012)
16. Kinnebrew, J.S., Segedy, J.R., Biswas, G.: Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition Learn.* **9**, 187–215 (2014). <https://doi.org/10.1007/s11409-014-9112-4>

17. Bannert, M., Reimann, P., Sonnenberg, C.: Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition Learn.* **9**, 161–185 (2014). <https://doi.org/10.1007/s11409-013-9107-6>
18. Sonnenberg, C., Bannert, M.: Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *J. Learn. Anal.* **2**, 72–100 (2015)
19. Gabadinho, A., Ritschard, G., Mueller, N.S., Studer, M.: Analyzing and visualizing state sequences in R with TraMineR. *J. Stat. Softw.* **40**, 1–37 (2011). <https://doi.org/10.18637/jss.v040.i04>
20. van der Aalst, W.: *Process Mining: Data Science in Action*. Springer, Heidelberg (2016). <https://doi.org/10.1007/978-3-662-49851-4>
21. Gatta, R., et al.: Generating and comparing knowledge graphs of medical processes using pMineR. In: *Proceedings of the Knowledge Capture Conference 2017, Austin, Texas* (2017)
22. Gatta, R., et al.: pMineR: an innovative R library for performing process mining in medicine. In: ten Teije, A., Popow, C., Holmes, J.H., Sacchi, L. (eds.) *16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna* (2017)



# Low-Investment, Realistic-Return Business Cases for Learning Analytics Dashboards: Leveraging Usage Data and Microinteractions

Tom Broos<sup>(✉)</sup>, Katrien Verbert, Greet Langie, Carolien Van Soom,  
and Tinne De Laet

Catholic University of Leuven, Leuven, Belgium  
{tom.broos,katrien.verbert,greet.langie,  
carolien.vansoom,tinne.delat}@kuleuven.be

**Abstract.** In recent years, Learning Analytics (LA) is finding more and more practical adoption, alongside of continued research interest. However, questions about the impact of LA applications and their underpinning in educational science are still being raised, impeding viability of some LA projects at larger scale. Within this paper we describe two examples using student-facing LA dashboards (LAD) deployed at scale at a relatively low cost. Leveraging data collected by the dashboards themselves, usage data (N = 4070 students) and in-dashboard microinteractions (N = 367 students), we try to put the impact question in perspective. We suggest that when investment is kept limited, a business case with modest but realistic expectations of returns may be feasible.

**Keywords:** Learning analytics · Learning analytics dashboards  
Business case · Realistics expectations · Usage data · Microinteractions

## 1 Introduction

Learning Analytics (LA) “*is about collecting traces that learners leave behind and using those traces to improve learning*” [6]. Typically, studies look for such traces in the virtual learning environment (VLE), or massive open online courses (MOOC) platform. Other approaches involve the use of multimodal data collection, including sensors to capture speech and gestures. However promising, most applications require a sizable investment before generating returns. Especially when aiming for deployment at scale, LA advocates may find it difficult to convince senior management. Our proposition is that at the institutional level, LA projects require a clear business case. From the management perspective, these projects compete for the same pool of resources available to other educational

---

This research is co-funded by the Erasmus+ program of the European Union (562167-EPP-1-2015-1-BE-EPPKA3-PI-FORWARD).

innovation projects. As such, LA advocates need to apply a return-on-investment (ROI) rhetoric. Projects requiring large investments, may require large expected returns and/or low risks. This may be a challenging requirement for LA, still in full development. Several authors have questioned the *impact* (or measurement thereof) of LA interventions (e.g. [4,5]). However, doing so without considering the required investment limits the discussion to the *numerator* of the ROI equation. We found little prior work that explicitly addresses ROI of LA projects. Picciano warned that investment in LA know-how “*will take time and additional resources and may or may not be worth the return on investment*” [7]. Slater underlines that the demonstration of ROI of LA standards may still take years [8].

We argue that with limited investment, a modest but realistic return may be attained. Many small-scale LA case studies incorporate this view implicitly, by focusing on a small group of students within a specific (favorable) setting. This paper moves into the opposite direction by targeting scalability to a large group of students, but at minimal cost. We report on easy to collect traces produced by the dashboard itself: usage data and in-dashboard microinteractions. Such ‘low-cost’ techniques may be used to break the chicken and egg cycle of convincing management without prior positive LA experience within the institution.

In the following sections, two examples using dashboards<sup>1</sup> are addressed. Both examples are part of a hands-on research approach where LADs are deployed at scale, studying usage behavior *in the wild* using qualitative and quantitative methods. The context and implementation of these dashboards using *small data* –results from a pre-existing questionnaire and study results readily available in the administrative systems– has already been described in detail in earlier work [2,3]. The first example connects dashboard usage early in the semester to study results after the semester exams. The second example is used to examine the use of simple microinteractions to get reverse feedback from students.

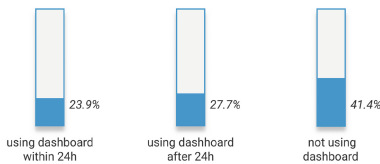
## 2 First Example: Dashboard Usage Data

A first example examines dashboard usage data that can be obtained at minimal cost. KU Leuven operates within an open access educational system: students are free to register to any study program (with the exception of medical sciences) and selective admission is not allowed. Recently, the university introduced a ‘30% rule’: new-coming students that do not succeed in at least 18 out of 60 credit points (study efficiency) are forced to re-orientate. While the rule is applied at the end of the academic year, students and institution could benefit from early detection and remediation. Results of the first-semester exams are currently the first available indicator, but much of the damage may already been done by then. The LA body of work explores many opportunities for early detection, e.g. based on digital learning traces in a VLE. However, from a practical perspective,

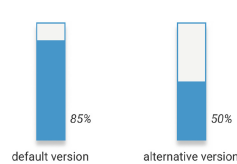
<sup>1</sup> Translated examples of both dashboards are available at <https://learningdashboards.eu/demo/ECTEL2018>.

many of these approaches are difficult and/or expensive to deploy at large scale. For this example, a dashboard about 'learning skills' that was made available to students mid-semester was used to study if dashboard usage data in itself may have the potential of being a cost-efficient probe to detect students at risk of failing.

*Method.* The 'learning skills' dashboard uses the data of a Learning and Study Strategies Inventory (LASSI) test students participated in. Once results were available, students were invited through email to use the dashboard as a feedback instrument about their motivation, concentration, time management, test strategy and failure anxiety. A logging system kept track of students clicking through to the dashboard. Several weeks later, students participated in exams. The results of these exams, more specifically if students managed to reach the 30% study efficiency threshold, was linked to the usage data collected earlier.



**Fig. 1.** Students below the threshold for early, late and non-users of the dashboard.



**Fig. 2.** Students (%) completing the microinteraction for all courses between the two dashboard versions.

*Results.* The analysis was limited to first-year students of 26 study programs who (1) filled out the LASSI questionnaire completely, (2) were invited by e-mail to use the mid-semester learning skills dashboard, and (3) were subsequently still registered for the exams at the end of the semester. Out of N = 4070 included students, 2420 (59.5%) accessed the mid-semester dashboard within 24 h<sup>2</sup>. Another 1355 (33.3%) students did use the dashboard, but not within the first 24 h of its availability. 295 (7.3%) students ignored or missed the e-mail invitation and did not use the dashboard at all. As summarized by Fig. 1, 23.9% of students who accessed the dashboard within 24 h, have a study efficiency below threshold at the end of the semester. This proportion is higher (27.7%) for students only accessing the dashboard after the first 24 h. The difference is significant ( $p = 0.012$ ) at the 5% level, when testing for equality of proportions. Within the group of students not using the dashboard at all, 41.4% of students ended up below the threshold. The difference is significant when testing for equality of the *non-user* group to the within 24 h group ( $p = 5.0e - 10$ ) and the *late-user* group ( $p = 9.7e - 06$ ).

<sup>2</sup> 'Accessing within 24 h' was defined as: accessing the dashboard within 24 h after the first student within the same study program accessed it.

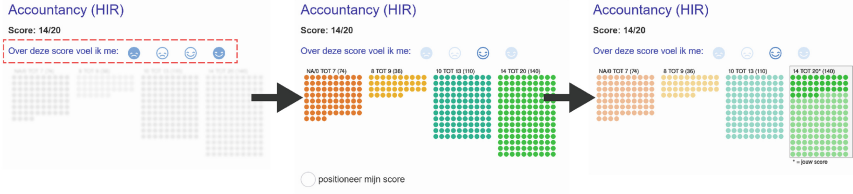
*Discussion.* The aim of this example was not to find strong evidence of the relationship between dashboard usage and study success, nor to explain it and even less to claim any impact in the form of causality. What remains is the observation that dashboard usage data in itself may have potential as an indicator for future study results. If so, it would have the advantage of cost-efficiency, requiring only limited integration with existing systems and not dependent on sensors or other sources of *big data*. The approach could serve as a quick win to demonstrate the potential of further investment in LA. Further research is necessary to refine this approach, e.g. by including more detailed in-dashboard behavior data (time spent, interaction, device type, etc.).

### 3 Second Example: Microinteractions

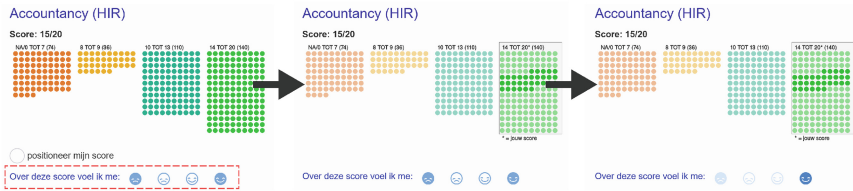
A second example demonstrates another option to collect digital traces from within a LAD at minimal cost. Microinteractions are defined as “*short-time interruptions of primary tasks*” [1]. Facebook’s *thumbs up* and swiping left (dislike) or right (like) in the Tinder dating app are examples. Within the domain of education, microinteractions may have the potential to collect self-reported data in a less obtrusive and more continuous way than existing questionnaire instruments. This principle was applied in a dashboard used to provide first-year students with feedback about their exam results.

*Method.* While the official grade report of KU Leuven only provides the raw scores, the exam results dashboard includes additional context and useful information. Grading in Flanders is typically not done on a curve, and especially first-year students have difficulties in interpreting scores. A prominent feature of the dashboard is the visualizations it provides at the course level about the distribution of results of all exam participants. By default (see Fig. 3), the dashboard requires students to answer a simple question before unlocking this feature. By selecting one of four faces, students share how they feel about their exam result for a given course: very unhappy, unhappy, happy or very happy. Once this microinteraction has taken place, the previously blurred chart becomes fully visible and students are given the option to position themselves more precisely in comparison to peers. To study the willingness of students to share information with the dashboard voluntarily, an alternative version (see Fig. 4) was introduced to a subgroup of students within a single study program (Bachelor of Engineering Sciences). Here the microinteraction is entirely optional: the detailed charts and positioning feature are immediately available to students. Within the selected study program, students were randomly assigned to either the default or alternative version of the dashboard. The latter group was oversampled (3/5).

*Results.* In total  $N = 367$  first-year students of the Bachelor of Engineering Sciences program clicked through, 157 to the default dashboard and 210 to the alternative version. Figure 2 clearly shows the difference in feedback completeness: 85% of students using the default version of the dashboard completed the

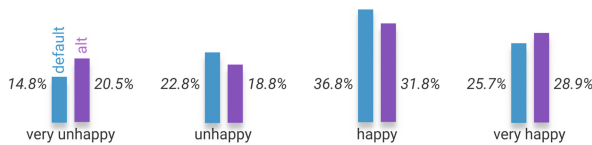


**Fig. 3.** Default version: excerpt of the dashboard. For each course, the number of exam participants within four categories ( $< 7/20$ ;  $8 - 9$ ;  $10 - 13$ ;  $\geq 14$ ) is displayed using dots. Students have to perform the reverse feedback microinteraction first to receive detailed information and to enable a more precise positioning of their scores.



**Fig. 4.** Alternative version: detailed feedback and optional precise positioning of scores is available immediately. The reverse feedback microinteraction is entirely optional.

microinteraction for each of the courses. Using the alternative version, only 50% of students completed all microinteractions. Additionally, it turns out that students using different versions of the dashboard also respond differently. Figure 5 summarizes the selected responses for all courses. Students using the alternative versions seem to be more likely to select more pronounced responses. A possible explanation is that these students were fully able to position their results in relation to peers, altering their interpretation of their own result –as anticipated in the Learning Analytics Process Model from Verbert [9].



**Fig. 5.** Comparison of microinteraction responses between students using the default (see Fig. 3, left-hand columns) and alternative (see Fig. 4, right-hand columns) versions of the dashboard.

*Discussion.* As in the previous example, this study remains on the surface and does not elaborate in detail on the interpretation of the learning traces obtained. Rather it demonstrates the possibility of engaging with students using a bidirectional feedback cycle facilitated by a dashboard instrument that requires only



limited investment. While a 50% response rate may seem limited, it should be taken into account that students did not receive any incentive to provide information in the alternative version. Further research is required to check if the resulting digital traces collected using the microinteractions may serve as valuable components in a cost-efficient LA model; if self-reported emotions using microinteractions can be a valuable feature for detecting students at risk; to see if the response rate can be improved using other incentives than unlocking dashboard features; and to study the different response patterns when feedback is optional.

## 4 Conclusion

This paper provides a plea for extending the question of *impact* of learning analytics (LA) to *return on investment* (ROI) to make the connection to the practice of senior management of educational institutions. Our suggestion is to consider low-cost, realistic-return business cases first to introduce learning analytics within the institution at scale. This approach requires cost-efficient techniques to capture the learning traces necessary to inform LA models and dashboards. Using examples of existing, scalable dashboards based on available, small data, we observed two techniques for using such dashboards to generate additional data: dashboard usage data and in-dashboard microinteractions. While both techniques demonstrate some promising results, further research is required to refine them. If the ROI question would start to take a more prominent place in the LA domain, we expect the introduction of many more cost-efficient techniques to enable low-cost, realistic-return business cases for LA.


## References

1. Ashbrook, D.L.: Enabling mobile microinteractions. Georgia Institute of Technology (2010)
2. Broos, T., Peeters, L., Verbert, K., Van Soom, C., Langie, G., De Laet, T.: Dashboard for actionable feedback on learning skills: scalability and usefulness. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2017. LNCS, vol. 10296, pp. 229–241. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58515-4\\_18](https://doi.org/10.1007/978-3-319-58515-4_18)
3. Broos, T., Verbert, K., Langie, G., Van Soom, C., De Laet, T.: Small data as a conversation starter for learning analytics: exam results dashboard for first-year students in higher education. *J. Res. Innovative Teach. Learn.* **10**(2), 94–106 (2017)
4. Dawson, S., Jovanovic, J., Gašević, D., Pardo, A.: From prediction to impact: evaluation of a learning analytics retention program. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 474–478. ACM (2017)
5. Dyckhoff, A.L., Lukarov, V., Muslim, A., Chatti, M.A., Schroeder, U.: Supporting action research with learning analytics. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 220–229. ACM (2013)
6. Erik, D.: Learning analytics and educational data mining (2012). <https://erikduval.wordpress.com/2012/01/30/learning-analytics-andeducational-data-mining/>
7. Picciano, A.G.: The evolution of big data and learning analytics in American higher education. *J. Asynchronous Learn. Netw.* **16**(3), 9–20 (2012)

8. Slater, N.: *Learning Analytics Explained*. Taylor & Francis, Routledge (2017)
9. Verbert, K., Duval, E., Klerkx, J., Govaerts, S., Santos, J.L.: Learning analytics dashboard applications. *Am. Behav. Sci.* **57**(10), 1500–1509 (2013)



# Identifying Design Principles for Learning Design Tools: The Case of edCrumble

Laia Albó<sup>(✉)</sup>  and Davinia Hernández-Leo 

ICT Department, Universitat Pompeu Fabra, Barcelona, Spain  
{laia.albo,davinia.hernandez-leo}@upf.edu

**Abstract.** Despite the existing variety of learning design tools, there is a gap in their understanding and adoption by the educators in their everyday practices. Sharing is one of the main pillars of learning design but sometimes it is not a sufficient reason to convince teachers to adopt the habit of documenting their practices so they can be shared. This study presents the design principles of edCrumble, an online learning design platform that allow teachers the creation and sharing of blended learning designs with the support of data analytics. The design principles have been learned and extracted from a participatory design process with teachers during the conceptualization and ongoing development of the tool. Several workshops including interviews were carried out as part of a design-based research iteration process. Later analysis has been done to extract and highlight those design principles aiming informing the development of learning design tools towards better learning design adoption.

**Keywords:** Design principles · edCrumble · Learning Design · Authoring tool  
Learning design adoption

## 1 Introduction

Learning Design (LD) tools have been conceived to support teachers in the process of documenting their teaching practices, making their learning design ideas explicit and sharable [1–3]. Despite the existing variety of learning design (LD) tools, there is a gap in their understanding and adoption by the educators in their everyday practices [4, 5]. Sharing is one of the main pillars of LD [6] but sometimes it is not a sufficient reason to convince teachers to adopt the habit of documenting their practices so they can be shared. Thus, one of the near-future LD challenge is reducing this gap and providing LD tools that can facilitate their adoption [5]. Moreover, despite existing proposed representations of pedagogical practice are varied, some are too specific for particular pedagogies and general approaches are not sufficiently accessible for teachers that do not have the required technical skills [7]. More intuitive visual representations of LD are needed [2]. [1] distinguishes two types of LD tools: “tools for visualizing designs” (which can be used to visualize and represent LDs) and “pedagogical planners” (which can guide and support educators in making informed LD decisions).

In this line, we have conceptualized and developed a generic LD tool that aims fitting in both categories bringing together the advantages of both types of tools. ILDE2/

edCrumble can be considered a pedagogical planner which provides an innovative visual representation of the LDs characterized by data analytics with the aim of facilitating the planning, visualization, understanding and reuse of complex LDs (available online at <https://ilde2.upf.edu/edcrumble/>). This study presents the design principles of edCrumble, extracted from a participatory design process with high school teachers during the conceptualization and ongoing development of the tool (Fig. 1).

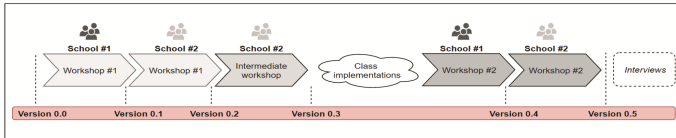


Fig. 1. edCrumble development versions regarding the participatory design workshops outputs.

## 2 Methodology

The development of edCrumble is part of a design-based research project which integrates several iteration cycles [8]. In this paper, we will present the design principles extracted from a complete cycle of this process which has the aim of prototyping and assessing the preliminary versions of the authoring tool. Within this cycle, 24 high school teachers from two different school communities have been involved in several participatory design workshops [9] between October 2017 and February 2018. Those teachers were participating in the context of a Teacher Professional Development program which had the aim of training teachers as designers of TEL and facilitate their inquiry practice with the collection of student data. For this reason, workshops were structured based on the following pattern: (1) **Workshop#1(2 h)**: teachers had to design a LD using edCrumble, with the help of the researchers (participants were asked to come to the workshop with a concrete LD idea); (2) **Class implementations (9 and 4 weeks respectively)**: teachers had to implement their LDs in class and collect students' data; (3) **Workshop#2 (2 h and 1 h respectively)**: joint reflection about the implementation phase and possible redesign (using edCrumble) of their original LDs. In the case of the second school, they had an intermediate 2 h workshop because they needed more time for designing the interventions.

At the end of the workshops phase, we carried out seven semi-structured face-to-face interviews of about 45 min each (three teachers from School#1 and four from School#2 -due time and resources constraints we could not interview all 24). The interviews consisted of a series of open-ended questions that invited participants to share their perspectives regarding (1) how they used to design and document their educational practices before knowing our tool and (2) how was the design process they followed during the workshops using the edCrumble (see the demographics of participants and interviews questions in [10]). The resulting qualitative data were coded, analyzed and triangulated by two researchers familiarized with the data. An open coding was used for identifying the main topics, extracting design principles and highlighting those aiming at informing the development of learning design tools towards better learning design

adoption. Specifically, in this paper we will focus on describing the design principles learned and extracted from the steps' outputs from the first version of the LD tool – conceived from the existing theory of the research field and our previous studies [11, 12] – to the current version (v.0.5) –developed based on the workshops' outputs during this cycle (Fig. 1).

### 3 Design Principles Regarding edCrumble Development Process

#### 3.1 Content and Activity Centered Planning

When we asked teachers “How do you usually design or prepare your courses?” they did not answer from a pedagogical point of view, instead they answered first from the content perspective – i.e. they explained how they structured the content without mentioning any pedagogical details (e.g. how the activities were designed: if they used collaborative learning or any pedagogical model...etc.). On one hand, five out of seven teachers said that they start preparing their courses examining the content that they must deliver and then filtering this content depending on the learning objectives. On the other hand, one participant said that she first starts looking on the objectives and then she plans the content. Last, one said that her preparation consists on a revision of the last year course and the re-adaptation of the content to the current objectives, as she has been teaching the same course for some years. This result is aligned with findings from related research. First, [13] state that the starting point of the design process depends on the nature of the design problem, identifying also three distinct starting points: from the learning outcomes, from a content-area focus and from a direct re-adaptation of previous LDs. Second, there is a need of describing teaching and learning activities as the “content” dimension of education is already captured in books, websites, etc. [14] for the later sharing and reuse of LDs. From our results we have observed that teachers need support to adopt and switch between these two approaches. **Implications for LD adoption:** From the above discussion we argue that it is important to foster the use of activity-centered model for capturing pedagogy beyond the content-based approach. But, at the same time, it is necessary to allow teachers to connect with their content-based approach whereas they adopting the LD aims (e.g. allow them to upload content related with their activities).

#### 3.2 Planning Tool Based on a Timeline

All teachers stated that they design their courses based on time using different tools: paper-based calendars or notes with dates, online calendar applications, LMS which organize the content based on time...etc. The time-based design approach used by teachers is aligned with Laurillard research insights in [6], who points out that the learning sequence is essentially time-based and that a LD does demand a plan. Other research findings also highlight the importance of the time and activity-sequence in course planning [3, 15]. **Implications for LD adoption:** we argue that LD tools which act as pedagogical planners can serve users in connecting their current planning practices with LD as they can foster the LD approach adoption by offering pedagogy support and helping in taking design-informed decisions during the design process.

### 3.3 Facilitate the Design in a Community of Educators

Most of teachers stated that they plan their activities alone, showing a high level of autonomy in deciding what and how to teach—results in line with [13]. The main reason is that usually there is only one teacher per topic and educational level in the school and there is no chance for co-design between teachers of the same educational context. Moreover, from the participatory workshops they highlighted the sharing and reflection phase they had during the second workshop as they really appreciated having found a space to talk with and learn from others' practices—despite they were LDs from other topics. It is known that the sharing is one of the most important aspects of LD field [15], but still there are few learning design tools that offer a social platform for exchange LDs.

**Implications for LD adoption:** we argue that is necessary to have LD tools that facilitate the sharing of the created LDs between educators—creating spaces for sharing LDs and support the seeking of similar topic LDs cross education-communities (open community instead of institutions-based closed communities).

### 3.4 Usability Matters: The Google Apps Effect

When we asked teachers about the weaknesses of the edCrumble, we detected what we name as the “Google Apps effect”: they were continuously referring to Google apps (calendar, drive, etc.) features for suggesting usability improvements to our tool. This result suggests, as other research findings pointed out, that usability is one of the two most important things (together with the usefulness) for users adopting a new technology [1]. Teachers are used to commercial applications, and existing LD applications are far from them in terms of design appeal and usability. **Implications for LD adoption:** Aesthetics and usability are an important factor to consider in the design of LD tools to facilitate their adoption.

### 3.5 Increasing the Utility Perception Solving Teachers' Real Problems

General opinion of teachers regarding edCrumble was positive despite most of them recognized that it will be difficult for them because of lack of time (as they put LD approach at the bottom of their list of day-to-day priorities). **Implications for LD adoption:** We argue that offering LD tools that can solve some of their day-to-day problems can be a way of adopting the LD approach—as it can increase their utility perception of the tools.

## 4 Decisions and Implications for the edCrumble Development

**Content and Activity centered planning:** (1) The LD is based on defining a sequence of activities which are composed by tasks. User can indicate for each task: the cognitive process level associated, the students type of work, the teacher's presence and the evaluation mode; (2) Users can provide the detailed list of learning objectives and relate them with the activities; (3) Users can upload all the content necessary to carry on their courses. **Planning tool based on a timeline:** The main element of the LD tool is a

timeline where users can place their activities sequenced depending on their schedule and type (in-class/out-of-class activities). **Facilitate the design in a community of educators:** edCrumble has been integrated as an authoring tool within the Integrated Learning Design Environment (ILDE2) [16] allowing practitioners to co-edit, share, remix and comment their designs and others' designs within a community of educators. Once teachers have implemented their LDs, they can upload their evaluation, helping others understand their impact and facilitating the adaptation and reusability of their LDs (e.g. describing the challenges found or uploading links to the resulting learning analytics). **Usability matters: the Google apps effect:** edCrumble must be improved in terms of design aesthetics and usability (i.e. allowing users creating grouped activities which follow a certain time pattern as Google Calendar automatically does when you want to create the same event at the same day every week). **Increasing the utility perception solving teachers' real problems:** During the interviews we have detected some teachers' needs arising during the LD process which edCrumble can solve: (1) the need of having a syllabus of the course for sharing it with students and institution (online and printed version) –*edCrumble can generate a LD summary including a printable syllabus with the activities description, the resources' plan and a report with all the analytics generated. Also, it provides an interactive visualization of the LD to be embedded or shared with the colleagues but also with the students to help them organize their courses.* (2) the interest of sharing the plan of the out-of-class activities between the different colleagues of the same educational level to leverage the “homework” of their students in a certain period –*the tool enables users to generate aggregated LD analytics from all the LDs placed in a folder (named as community analytics), supporting teachers' decision making during the LD process not only at their individual level but also allowing the possibility of considering the colleagues' LDs analytics in their community;* (3) the need of decreasing the time needed to document their practices in edCrumble as it is an entry barrier for those teachers that do not plan but only need re-adapting LDs –*further work has to be done to improve the flexibility and connection with existing tools (LMS, calendars...).*

## 5 Discussion and Conclusions

In this paper, we have extracted some design principles from interviews with high school teachers involved in participatory design workshops with the aim of informing the design and development of the edCrumble learning design tool. Of those design principles, we can highlight two rules which we think they can facilitate the adoption of the LD tools by educators in their daily practices: LD tools which seek to connect with teachers' existing practices and LD tools which seek for solving teachers' day-to-day problems [13]. From the first one, the following design principles are derived: *Content and Activity centered planning, Planning tools based on time, Usability matters: the Google apps effect.* And from the second one: *Facilitate the learning design in a community of educators and Increasing the utility perception solving teachers' day-to-day problems.* The final evaluations of ILDE2/edCrumble are part of an ongoing cycle of a design-based research process. Further research is needed to evaluate the edCrumble adoption

by educators and inform the redesign of the existing identified design principles for supporting the development of future learning design tools.

**Acknowledgements.** Authors want to thank all the teachers who participated in the study. This work has been partially funded by RecerCaixa (CoT project) and the Spanish Ministry of Economy and Competitiveness under MDM-2015-0502, TIN2014-53199-C3-3-R, TIN2017-85179-C3-3-R.

## References

1. Conole, G.: Designing for Learning in an Open World, vol. 4. Springer Science & Business Media, New York (2012). <https://doi.org/10.1007/978-1-4419-8517-0>
2. Agostinho, S.: The use of a visual learning design representation to support the design process of teaching in higher education. *Australas. J. Educ. Technol.* **27**(6), 961–978 (2011)
3. Laurillard, D., et al.: A constructionist learning environment for teachers to model learning designs. *J. Comput. Assist. Learn.* **29**(1), 15–30 (2013)
4. Celik, D., Magoulas, G.D.: A review, timeline, and categorization of learning design tools. In: Chiu, D., Marenzi, I., Nanni, U., Spaniol, M., Temperini, M. (eds.) ICWL 2016. LNCS, vol. 10013, pp. 3–13. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-47440-3\\_1](https://doi.org/10.1007/978-3-319-47440-3_1)
5. Cameron, L.: How learning design can illuminate teaching practice. In: The Future of Learning Design Conference (2009)
6. Dalziel, J.: Learning Design: Conceptualizing a Framework for Teaching and Learning Online. Routledge, New York (2015)
7. Pozzi, F., Asensio-Pérez, J.I., Persico, D.: The case for multiple representations in the learning design life cycle. In: Gros, B., Kinshuk, Maina, M. (eds.) The Future of Ubiquitous Learning. Lecture Notes in Educational Technology, pp. 171–196. Springer, Heidelberg (2016). [https://doi.org/10.1007/978-3-662-47724-3\\_10](https://doi.org/10.1007/978-3-662-47724-3_10)
8. Amiel, T., Reeves, T.C.: Design-based research and educational technology: rethinking technology and the research agenda. *Educ. Technol. Soc.* **11**(4), 29–40 (2008)
9. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. CRC Press, Hillsdale (1993)
10. Albó, L., Hernández-Leo, D.: Participants' data and interview questions - Identifying design principles for learning design tools: the case of edCrumble (Version v.1). Zenodo, 2 May 2018. <https://doi.org/10.5281/zenodo.1239740>
11. Albó, L., Hernández-Leo, D.: Blended learning with MOOCs: towards supporting the learning design process. In: OOFHEC 2016, pp. 578–588 (2016)
12. Albó, L., Hernández-leo, D., Oliver, M.: Blended MOOCs: university teachers' perspective. In: HybridEd Workshop, EC-TEL 2015, pp. 11–15 (2015)
13. Bennett, S., Agostinho, S., Lockyer, L.: The process of designing for learning: understanding university teachers design work. *Educ. Technol. Res. Dev.* **65**(1), 1–21 (2016)
14. Dalziel, J., et al.: The Larnaca declaration on learning design. *J. Interact. Media Educ.* **1**(7), 1–24 (2016)
15. Dalziel, J.: Implementing learning design: the learning activity management system (LAMS). In: 20th Annual Conference of the Australasian Society for Computers Learning in Tertiary Education, pp. 7–10, December 2003
16. Hernández-Leo, D., et al.: An integrated environment for learning design. *Front. ICT* **5**, 9 (2018)





# Exploring Causality Within Collaborative Problem Solving Using Eye-Tracking

Kshitij Sharma<sup>1</sup>(✉), Jennifer K. Olsen<sup>2,3</sup>(✉), Vincent Aleven<sup>3</sup>,  
and Nikol Rummel<sup>3,4</sup>

<sup>1</sup> Norwegian University of Science and Technology, Trondheim, Norway  
`kshitij.sharma@ntnu.no`

<sup>2</sup> École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland  
`jennifer.olsen@epfl.ch`

<sup>3</sup> Carnegie Mellon University, Pittsburgh, PA, USA  
`aleven@cs.cmu.edu`

<sup>4</sup> Ruhr University Bochum, Bochum, Germany  
`nikol.rummel@rub.de`

**Abstract.** When students are working collaboratively and communicating verbally in a technology enhanced environment, the system is not aware of what collaboration is happening outside of the technology, making it difficult to adapt the system to better support the collaboration of the students. In this paper, we analyze the causal relationships between collaborative and individual gaze measures and the influence that the students dialogue, prior knowledge, or success has on these relationships to find indicators that can be used within an adaptive system. We found that when students are discussing concrete aspects of the problem, the causal relationship between their eye gaze measures changes compared to other types of dialogue patterns. The results also show a clear difference in causal relations when the pairs with high prior knowledge or success are compared with the pairs with low prior knowledge or success. Collaborative gaze causes the individual gaze for pairs with high prior knowledge and the opposite for the pairs with low prior knowledge.

**Keywords:** Collaboration · Dual eye tracking · ITS  
Granger causality

## 1 Introduction

In technology supported collaborative settings, students not only benefit from the support of the technology, but also from the exchange of ideas and explanations within their group. Currently, many technologies that are developed to support learning focus on the support of the domain material with support for collaboration being an afterthought, if explicitly supported at all, as is the case with individual Intelligent Tutoring Systems (ITSs). Additionally, in the classroom, students are often collaborating face-to-face and communicating verbally

with these verbal interactions occurring outside of the system making it difficult for the system to have a complete picture of the collaboration. For this paper, we are interested in supporting the collaborative interactions that occur between students as they work on a collaborative technology where all of the interactions may not be captured through the system.

Adaptive collaborative learning support (ACLS) can be used to adapt to the collaborative learning environment to provide appropriate support for the students by assessing student interactions, comparing them to a set of productive interactions, and providing interventions that will guide students closer to a productive interaction [49, 53]. Because verbal communication is still difficult to assess in real-time and students may not always be providing input to the learning technology. We propose using eye-tracking to assess student collaboration behaviors by investigating the different causal relationships of different process variables to find indicators that can be tracked and measured in real-time within a collaborative setting.

In this paper, we investigate the causal relationships between students' individual and collaborative gaze patterns (i.e., focus and similarity) for elementary school students working on a collaborative fractions ITS and examine how their dialogue plays a role in this relationship. For this analysis, we used time series data from the students working on the tutor. In the following sections, we will present an overview of the literature, study context, the analysis process, and causal inference results. These results provide insights into how eye-tracking measures can be used within collaborative learning environments to assess the level of collaboration and adapt to the current collaboration state. Specifically, we will address the following research question for this contribution: 1. What is the nature of causality between the collaborative and individual gaze patterns and 2. How do dialogue, prior knowledge and success alter this causality?

## 2 Related Work

Current implementations of ACLS often use either attributes of the student dialogue or interactions with the learning technology to assess the current collaboration state of the group. Many previous ACLS systems have used shallow indicators from dialogue to support student collaborations such as the number of student utterances [17, 41], used sentence openers [5, 34], or tracked particular sequences of dialogue actions (e.g., use of a question mark or dialogue talk moves) [1]. Often this analysis has been done on students who are communicating through chat where the features are easier to extract. Additionally, by including features of the learning environment in the assessment of the collaboration, such as the classification of the dialogue in relation to the actions the students are taking in the learning environment, often the intervention can be more impactful [33, 52, 54]. ACLS systems have also used interactions in the learning technology to gauge the collaboration, such as the request of hints and error patterns [54]. However, these interactions are not as useful in understanding what is happening outside of the system if there are long pauses between interactions when students

may be having discussions. Eye-tracking may be able to be used to make this link between the information provided in the learning technology and the group discussions that occur outside of them.

Eye-tracking may be a promising method to use to assess student collaboration as research has shown that eye gaze is tied to communication [35]. Previous research has shown a link between speech and eye gaze when people are working together on a task. There is a coupling of the collaborators' eye gaze around a reference [40], meaning that the collaborators' gaze may fixate, at approximately the same point in time, at the object referenced in the dialogue, for example just before mentioning it and just after hearing about it. The eye gaze has a closer coupling when each of the collaborators has the same initial information and when collaborators can visually share important objects that they are referencing in speech [26, 40], suggesting that concrete references may have more of an impact on eye gaze compared to abstract references.

Over the past few years, eye-tracking has become a key source of process data in educational research. Research using eye-tracking covers a wide range of educational ecosystems. Eye-tracking has not only been used to understand the learning processes in various contexts [38, 39, 46], but it also has been used to provide students appropriate, real-time, and adaptive feedback on their learning processes [14, 45]. In terms of collaborative learning scenarios, eye-tracking has most often been used with collaborating partners dialogues. Research has shown that there is a time lag between looking at an object and referring to the same object (eye-voice span) [21] and a time lag between a speaker's reference and a listener's gaze on the referred object (voice-eye span) [3]. Additionally, in terms of dual eye gaze, there is a lag in the eye-eye (speakers eye listeners eye) span (i.e., the time difference between the moment a speaker looks at an object and the moment the listener looks at the same object) [40]. Most of the dual eye-tracking studies have shown that the amount of time that the collaborating partners spend while looking at the same objects at the same time (cross-recurrence) is predictive of several collaborative constructs (e.g., collaboration quality [26]; misunderstandings [11]; learning gains [42]). In this paper, we go beyond correlational links to explore where there may be causal links between eye gaze measures and how they change during different forms of dialogue.

In this contribution, we propose a shift from correlation to causality. We borrow methods from finance and environmental studies to understand the causal relation between the different gaze-based variables. The key idea is to use the "cause" to "forecast" the effect and prepare for "adaptation" in ITSs. This has been a traditional practice in finance and environmental studies to use the causality to forecast [10, 12, 22, 28] and to use forecasting for adaptation requirements [7, 8, 32, 55]. We propose to use the causal relationship between the individual and collaborative gaze patterns to be able to forecast the behavior and provide adaptive feedback in a proactive manner.

For understanding the behavioral relation between the individual and collaborative gaze, we will use the Granger causality [20], a method that has been used in a multitude of domains to understand the relationship between

observable variables. For example, neuro-science [16, 19], user-consumption [36], stock-market [24] and economics [27, 50]. We will also explore the nature of this causal relationship using co-variables such as: pairs' dialogue, their prior knowledge and success levels.

In our work, we used a fractions ITS as a platform for our research. ITSs have been shown to be beneficial for student learning [30, 31] and are effective by providing cognitive support for students as they work through problem-solving activities. This cognitive support comes in the form of step-level guidance, namely, an interface that makes all steps visible, error feedback, and on-demand hints, which allow the system to adapt to the students current level of knowledge [51]. The cognitive support provided through the system can provide support for the student learning of the domain but does not provide support for the student collaboration when they are working in groups.

### 3 Methods

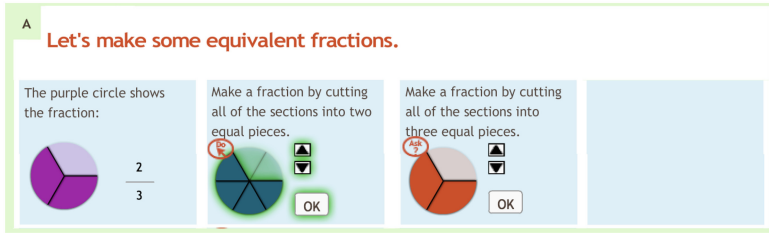
#### 3.1 Experimental Design and Procedure

Our data set involves 14 4th and 14 5th grade dyads from a larger study that investigated the benefits of collaborative versus individual learning [6, 37]. Each teacher paired the students participating in the study based on students who would work well together and had similar, but not equivalent, math abilities. The dyads were engaged in a problem-solving activity using a networked collaborative ITS, which allowed them to synchronously work in a shared problem space where they could see each others actions while sitting at their own computers. The students were able to communicate verbally through a Skype connection. Each dyad worked with the tutor for 45 min in a pull-out study design at their school. The morning before working with the tutor and the morning after working with the tutor, students were given 25 min to complete a pretest or posttest individually on the computer to assess their learning. During the experiment, dual eye tracking data, dialogue data, and tutor log data in addition to the pretest and posttest measures were collected. We collected eye-tracking data using two SMI Red 250 Hz infrared eye-tracking cameras.

#### 3.2 Intelligent Tutoring System

During the study, the dyads engaged with an ITS oriented towards supporting the acquisition of knowledge about fraction equivalence. Within each problem, the tutor provided standard ITS support, such as prompts for steps (i.e., revealing steps sequentially), next-step hints, and step-level feedback (i.e., correct or incorrect feedback) that allows the problem to adapt to the students problem-solving strategy [51]. Each of these different supports were displayed as actions on the screen that could guide the students actions and gaze.

For the collaboration, the ITS support mentioned above was combined with embedded collaboration scripts, which allowed students to take slightly different actions and see different information. The embedded collaboration scripts



**Fig. 1.** Example of a fractions interface showing incremental step reveals, feedback, and hint requests. Students had roles assigned that were displayed through their icon.

included three theoretically proven types of collaboration support: roles, cognitive group awareness, and individual accountability. First, for many steps, the students were assigned roles [29]. In the tutors, on steps with roles, one student was responsible for entering the answer and the other was responsible for asking questions of their partner and providing help with the answer. The tutor indicated the current role for the students through the use of icons on the screen. A second way in collaboration was supported was by providing students with information their partner did not have that they were responsible for sharing for the problem to be completed causing individual accountability [48]. The final feature was cognitive group awareness, where knowledge that each student has in the group is made known to the group [25]. On steps where this feature was implemented, each student was given an opportunity to answer a question individually before the students were shown each others answers and asked to provide a consensus answer.

### 3.3 Variables

For our analysis, we investigated a combination of data streams from eye gaze measures, dialogue, and test scores. For our eye gaze measures, we used focus and similarity because these two variables have been used in the recent research work concerning collaborative eye-tracking [43,46,47] to combine and analyse gaze behaviour at individual and collaborative levels. We used dialogue abstract as it can indicate how grounded the speech of the students is to what is occurring on the problem. Finally, the pretest and posttest scores allowed us to understand the relation of the causality to student knowledge.

**Individual Focus.** This is computed in terms of the entropy of the gaze. To compute the entropy, we divided the screen in 50-by-50 pixels grid. We also divided the whole problem-solving session into 10 seconds time windows. We then computed the proportion of the time spent in each block in the spatial grid for each 10-second time window. This resulted in a series of 2-dimensional proportionality vectors. Finally, we computed the Shannon Entropy for each of the vectors. A low entropy value (the minimum possible value is zero) depicts that the student was looking at only a few elements on the screen, which we called focused gaze. On the other hand, a high value of entropy indicates more

elements being looked at in a given time window, which we called unfocused gaze. Although focus and attention are related concepts, focus, as we defined here, does not contain the idea of processing the stimulus, as is required in the definition of attention. Attentive gaze indicates a certain level of processing of the sensory input. Focused gaze simply indicates a small number of elements looked over a fixed time period.

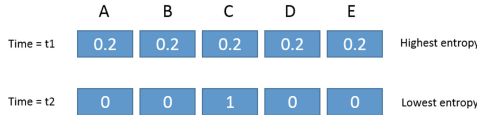


Fig. 2. Entropy computation

**Collaborative Gaze.** In order to compute the similarity between the gaze patterns of the collaborating students, we divided the screen space and the interaction time in the same manner as we did for entropy computation. We computed the similarity between the two proportionality vectors by using the reverse function  $(1/(1+x))$  of the correlation matrix of the two vectors. A similarity value of one will show no similarity between the two gaze patterns during a given time window. On the other hand, a higher value of similarity will show that the two participants spent time looking at the similar set of object on the screen during the same time window. Gaze similarity is an alternative measure of gaze convergence, the only difference between gaze similarity and gaze convergence comes from the mathematical formulation.

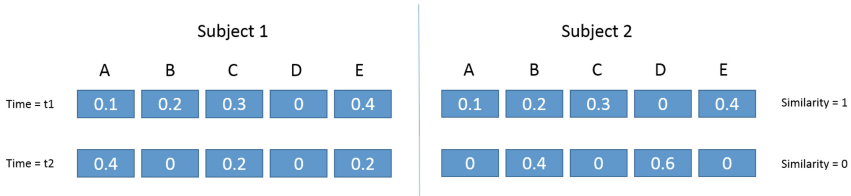


Fig. 3. A typical Similarity computation example

**Dialogue Abstraction.** Each of the student dialogues were transcribed and coded for abstraction levels. Abstraction is how grounded within the concrete aspects of the problem solving and communication the students utterance is. The level of abstraction is fully dependent on what occurs in the dialogue and is not intended to infer all mental processes. Within our transcripts, we coded for abstraction at the utterance level. This allowed us to have a fine-grained coding for each second of the dialogue without losing the context of the words. The abstraction codes consisted of five different levels: acknowledgement, read

out loud, interface, problem solving, and metacognitive (See examples below). The levels of abstraction followed an ordering with acknowledgments being the least abstract and metacognitive being the most abstract. For the coding, all statements that were off-task or were with a researcher were marked as “not applicable” and were discarded from the analysis. An inter-rater reliability analysis was performed to determine consistency among raters ( $Kappa = 0.78$ ).

1. Not applicable (NA): The student engages in off-task behavior, converses with the experimenter, or vocalizations without any context.
2. Acknowledgement (ACK): The student acknowledges their partner, or they request acknowledgment or a repeat of what the partner has said.
3. Read-out-loud (ROL): The student is reading information provided within the problem and presented on the screen.
4. Interface (INT): The student discusses actions that can be taken in the interface or engage in work coordination.
5. Problem solving (PRO): The student is providing an answer to the problem or showing evidence of think aloud as they solve the problem.
6. Metacognitive (META): The student verbally expressing their understanding of their current knowledge/problem solving state.

**Pretest and Posttest Scores.** To measure learning, we administered pretest and posttests to the students. The tests were computer-based and developed to closely align with the target knowledge covered in the tutors. The test comprised of 5 procedural and 6 conceptual test items. Two isomorphic sets of questions were developed, and there were no differences in performance on the test forms across all participants in the original study,  $t(79) = 0.96$ ,  $p = 0.34$ . The presentation of these forms as pretests and posttests was counterbalanced.

### 3.4 Data Analysis

We used Granger causality [20] test to examine the causality between the focus and similarity. The basic definition of Granger causality has two assumptions [20]. First, that cause occurs before effect and that the cause has information about the effect that is more important than the history of the effect. Although Granger causality is defined for linear and stationary time-series contexts, the variations for non-linear [4, 9, 18] and non-stationary [15, 23] contexts exist. The basic principle of Granger causality is to compare two models to test if  $x$  causes  $y$ . The first model predicts the value of  $y$  at time  $t$  using the previous  $n$  values of  $y$ . The second model predicts the value of  $y$  at time  $t$  using the previous  $n$  values of both  $x$  and  $y$ . Mathematically, following is a bivariate linear auto-regressive model for two variables  $x$  and  $y$ :

$$y(t) = \sum_{j=1}^p \alpha_{11j}x(t-j) + \sum_{j=1}^p \alpha_{12j}y(t-j) + \varepsilon_1(t) \quad (1)$$

$$x(t) = \sum_{j=1}^p \alpha_{21j}x(t-j) + \sum_{j=1}^p \alpha_{22j}y(t-j) + \varepsilon_2(t) \quad (2)$$

Where,

$p$  = model order, maximum lag included in the model

$\alpha$  = coefficients matrix, contribution of each lag value to the predicted value

$\varepsilon$  = residual, prediction error

We can conclude that  $x$  *granger-causes*  $y$  if coefficients in  $\alpha_{12}$  are jointly significant from zero. Statistically, this can be tested using F-test with the null hypothesis  $\alpha_{12} = 0$ . Also, the value of  $p$  can be decided based on the AIC [2] or BIC [44] model estimation values.

## 4 Results

In this section, we will provide the different analyses to arrive at a causal relationship between the variables mentioned in the Sect. 3.3. First, we would give an example about how to determine the granger causality between two variables to make the method explained in the Sect. 3.4.

Let us take the case of “focus” (the probability that both the participants have low gaze entropy) and “similarity” (the extent to which the peers looked at a similar set of objects in the a given time window). Table 1, comparison 1 shows the granger causality results for the overall data. The order of the model (Table 1, column 2) denotes how much lag was used to compute the causal relationship ( $p$  in Eqs. 1 and 2). In the case of Table 1, comparison 1, the lags used are 4 time windows (each time window corresponds to 10 s). To check if similarity granger causes focus, we create two models given by Eqs. 1 and 2 and compare them using F-test. The F and p values denote the effect size and significance of the model (Table 1, columns 3 and 4, respectively). We repeat the same process for checking if focus granger causes similarity. As we can see in Table 1 comparison 1, that “similarity granger causes focus” have a higher F (2.51) and lower (and significant) p value (.03) than “similarity granger causes focus” (F = 2.04, p = .09). Thus, we can conclude that “similarity granger causes focus”.

The remainder of this section presents the main results for this contribution. We observe that *similarity Granger causes focus* during the whole interaction (Table 1, Comparison 1). This causality also holds up when the dyads are engaged in a dialogue (Table 1, Comparison 2). Considering the data from the individual dialogue categories, The same causality holds when the peers are talking about interface issues (INTF, Table 1, Comparison 3). However, the causality changes the polarity (that is *focus Granger causes similarity*) while the peers are talking about problem solving (Table 1, Comparison 4); And there is no conclusive causality for ACK and META.



**Table 1.** The Granger causality model, across different data types, for collaborative similarity and probability that both participants have high focus. The direction of causality is denoted with a \*.

Model	Order	F-value	p-value
Overall Data (1)			
Focus <- Similarity	4	2.51	.03*
Similarity <- Focus	4	2.04	.09
Participants Engaged in Dialogue (2)			
Focus <- Similarity	8	2.12	.03*
Similarity <- Focus	8	0.93	.47
Participants Engaged in Dialogue w/INTF Abstraction (3)			
Focus <- Similarity	6	2.83	.009*
Similarity <- Focus	6	1.01	.41
Participants Engaged in Dialogue w/PRO Abstraction (4)			
Focus <- Similarity	5	0.21	.95
Similarity <- Focus	5	2.52	.02*
Dyads with High Average Posttest Scores (5)			
Focus <- Similarity	2	3.91	.02*
Similarity <- Focus	2	1.70	.18
Dyads with Low Average Posttest Scores (6)			
Focus <- Similarity	3	7.04	.00001*
Similarity <- Focus	3	2.04	.11
Dyads with High Average Posttest Scores w/PRO Abstraction (7)			
Focus <- Similarity	2	2.81	.05*
Similarity <- Focus	2	1.01	.31
Dyads with Low Average Posttest Scores w/PRO Abstraction (8)			
Focus <- Similarity	3	0.54	.44
Similarity <- Focus	3	2.74	.05*
Dyads with High Average Pretest Scores (9)			
Focus <- Similarity	3	6.49	.0002*
Similarity <- Focus	3	.04	.98
Dyads with Low Average Pretest Scores (10)			
Focus <- Similarity	3	0.11	.95
Similarity <- Focus	3	4.42	.004*

However, when we divide the data into pairs with high and low average posttest scores, we observe a few different relations. For the pairs with high posttest average *similarity Granger causes focus* (Table 1, Comparison 5) This polarity does not change for “PRO” abstraction (Table 1, Comparison 7). For the pairs with low

posttest average *focus Granger causes similarity* (Table 1, Comparison 6) and the polarity changes for “PRO” abstraction (Table 1, Comparison 8).

This result shows that there is some kind of interaction between the focus, similarity and performance. There is also an interaction between the focus, similarity and dialogue. Finally, we considered the relation between the pre and the post test scores. There is a positive significant correlation between the average pretest and the posttest scores for the pairs ( $r(27) = 0.57, p = .001$ ), indicating that prior knowledge also contributes in the success. Therefore, we divided the dataset into dyads with low and high average pretest scores and found that *similarity granger causes focus* for the pairs with high average pretest scores (Table 1, Comparison 9); whereas, *focus granger causes similarity* for the pairs with low average pretest scores (Table 1, Comparison 10).

## 5 Discussion and Conclusions

Granger causality is useful for forecasting the caused variable. In this paper, we examined the causal relation between individual and collaborative gaze-patterns, and used the dialogue, pretest and posttest scores as co-variates to explain the observed causality in detail. By understanding the causality, we can better use these measures to assess the collaborative state of students and develop interventions to guide the collaborative process.

In our analysis, we found that overall the collaborative similarity is causing the individual focus. This causality switches, that is individual gaze causes collaborative gaze, when the pairs are talking about “how to solve the problem?” One plausible explanation for this is that when two peers are talking about ways to solve problems, they both are individually focused on the problem description areas and hence start looking at the same section of the screen. Moreover, there is no conclusive causality during the episodes when the peers are in “ACK” or “META” abstraction. This may be explained by the fact that there is no need for the stimulus support when acknowledging a partner’s dialogue or a requirement to reflect upon a peer’s own state of understanding.

The key difference between the two causalities “looking at the same place hence focused” and “focused hence looking at the same place” might explain the fact whether collaboration is driving the individual gaze or the other way. In the case of successful pairs the collaboration seems to drive the individual behaviour, while in the case of unsuccessful pairs the relationship seems reversed. The same difference is there for the pairs with high and low prior knowledge. That is “similarity causes focus” for the pairs with high prior knowledge and “focus causes similarity” for the pairs with low prior knowledge. This difference could be a guiding factor about “how to provide adaptive feedback to the students?”

Additionally, the different causal relations for pairs with different levels of prior knowledge and success show that collaborative gaze causing the individual gaze is indicative of a “top-down” approach while individual gaze causing the collaborative gaze points to a “bottom-up” approach. Having coordinated gaze is a result of deeper socio-cognitive mechanisms [26, 40, 42, 43, 46] than just looking at a few elements on the screen (high focused gaze, by definition). In this

way, one can hypothesize that individual focus is similar to gaze reacting to the stimulus (screen or partner's dialogue) that is bottom-up behaviour [13]. On the other hand, the coordinated gaze is similar to cognition-driven gaze (referential gestures or familiarity with the interface or prior knowledge) that is top-down behaviour [13]. Our results show that examining the causality between collaborative and individual gaze patterns can unveil intriguing cognitive mechanisms underlying the collaborative learning with tutoring systems.

By forecasting the focus of the peers, we can take suitable actions for keeping the focus size for students in check. Using our results, when the focus size is large, given the similarity of the students, we can provide appropriate gaze-aware cues to the students, which would increase their similarity. From our results, this increase in similarity should increase the student focus, which can lead to more effective collaboration.

Additionally, we can provide feedback to the students based upon their eye gaze patterns. For example, whenever we detect that the focus is causing similarity, which tells us that they are not talking "PRO" then we can provide prompts to the students to guide their discussion back to the problem. We can test the impact of the prompts if we see that the *similarity is causing focus*, indicating the students' dialogue is discussing the problem.

Another opportunity for the personalized and adaptive feedback arises from the different causal relations based on the prior knowledge of the pairs. We found that for the pairs with high prior knowledge (high average pretest score) similarity causes focus, while for the pairs with low prior knowledge (low average pretest score) this is the focus that causes similarity. For such pairs (low prior knowledge), one can start giving feedback about where the partner is looking at, from the beginning of the session so that the high levels of similarity could be initiated and maintained throughout the collaboration and hence high levels of individual focus.

This work contributes to adaptive learning by revealing causality relations between individual and collaborative eye gaze measures that can be used to assess the collaboration of a group so that interventions can be applied at the correct moments. In future work, we would like to both extend our analysis to account for how features of the tutoring environment impact the findings as well as apply our findings to an adaptive environment to investigate if an adaptive system developed using these indicators is effective. A limitation of our work is that we had a small sample size and this may have impacted the results, which should be addressed in future work. Overall, our results indicate that student dialogue can impact the eye gaze relations as well as student prior knowledge. Understanding these relations allow us to adapt the system to better support student collaboration.

## References

1. Adamson, D., Rosé, C.P.: Coordinating multi-dimensional support in collaborative conversational agents. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 346–351. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-30950-2\\_45](https://doi.org/10.1007/978-3-642-30950-2_45)
2. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. control* **19**(6), 716–723 (1974)
3. Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K.: Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Memory Lang.* **38**(4), 419–439 (1998)
4. Ancona, N., Marinazzo, D., Stramaglia, S.: Radial basis function approach to non-linear granger causality of time series. *Phys. Rev. E* **70**(5), 056221 (2004)
5. Baker, M., Lund, K.: Promoting reflective interactions in a CSCL environment. *J. Comput. Assist. Learn.* **13**(3), 175–193 (1997)
6. Belenky, D., Ringenberg, M., Olsen, J., Alevan, V., Rummel, N.: Using dual eye-tracking to evaluate students' collaboration with an intelligent tutoring system for elementary-level fractions. Grantee Submission (2014)
7. Bertolli, C., Buono, D., Mencagli, G., Vanneschi, M.: Expressing adaptivity and context awareness in the assistant programming model. In: International Conference on Autonomic Computing and Communications Systems, pp. 32–47 (2009)
8. Challinor, A.: Towards the development of adaptation options using climate and crop yield forecasting at seasonal to multi-decadal timescales. *Environ. Sci. Policy* **12**(4), 453–465 (2009)
9. Chen, Y., Rangarajan, G., Feng, J., Ding, M.: Analyzing multiple nonlinear time series with extended granger causality. *Phys. Lett. A* **324**(1), 26–35 (2004)
10. Cheng, C.H., Wei, L.Y., Chen, Y.S.: Fusion anfis models based on multi-stock volatility causality for taiey forecasting. *Neurocomputing* **72**(16–18), 3462–3468 (2009)
11. Cherubini, M., Nüssli, M.A., Dillenbourg, P.: Deixis and gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings. In: Proceedings of the 2008 symposium on Eye tracking research & applications, pp. 173–180. ACM (2008)
12. Clements, M.P., Hendry, D.F.: An overview of economic forecasting. A companion to economic forecasting pp. 1–18 (2002)
13. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: bottom-up versus top-down. *Curr. Biol.* **14**(19), R850–R852 (2004)
14. D'Angelo, S., Begel, A.: Improving communication between pair programmers using shared gaze awareness. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6245–6290. ACM (2017)
15. Ding, M., Bressler, S.L., Yang, W., Liang, H.: Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biol. Cybern.* **83**(1), 35–45 (2000)
16. Ding, M., Chen, Y., Bressler, S.L.: 17 granger causality: basic theory and application to neuroscience. *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, **437** (2006)

17. Dowell, N.M., Cade, W.L., Tausczik, Y., Pennebaker, J., Graesser, A.C.: What works: creating adaptive and intelligent systems for collaborative learning support. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 124–133. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07221-0\\_15](https://doi.org/10.1007/978-3-319-07221-0_15)
18. Freiwald, W.A., et al.: Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *J. Neurosci. Methods* **94**(1), 105–119 (1999)
19. Goebel, R., Roebroeck, A., Kim, D.S., Formisano, E.: Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magn. Reson. Imaging* **21**(10), 1251–1261 (2003)
20. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: J. Econom. Soc.* 424–438 (1969)
21. Griffin, Z.M., Bock, K.: What the eyes say about speaking. *Psychol. Sci.* **11**(4), 274–279 (2000)
22. Hafner, C.M.: Causality and forecasting in temporally aggregated multivariate garch processes. *Econom. J.* **12**(1), 127–146 (2009)
23. Hesse, W., Möller, E., Arnold, M., Schack, B.: The use of time-variant EEG granger causality for inspecting directed interdependencies of neural assemblies. *J. Neurosci. Methods* **124**(1), 27–44 (2003)
24. Hiemstra, C., Jones, J.D.: Testing for linear and nonlinear granger causality in the stock price-volume relation. *J. Financ.* **49**(5), 1639–1664 (1994)
25. Janssen, J., Bodemer, D.: Coordinated computer-supported collaborative learning: awareness and awareness tools. *Educ. Psychol.* **48**(1), 40–55 (2013)
26. Jermann, P., Nüssli, M.A.: Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 1125–1134. ACM (2012)
27. Joerding, W.: Economic growth and defense spending: granger causality. *J. Dev. Econ.* **21**(1), 35–40 (1986)
28. Kavussanos, M.G., Nomikos, N.K.: Price discovery, causality and forecasting in the freight futures market. *Rev. Deriv. Res.* **6**(3), 203–230 (2003)
29. King, A.: Discourse patterns for mediating peer learning (1999)
30. Kulik, J.A., Fletcher, J.: Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* **86**(1), 42–78 (2016)
31. Ma, W., Adesope, O.O., Nesbit, J.C., Liu, Q.: Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* **106**(4), 901 (2014)
32. Manevitz, L., Bitar, A., Givoli, D.: Neural network time series forecasting of finite-element mesh adaptation. *Neurocomputing* **63**, 447–463 (2005)
33. McLaren, B.M., Scheuer, O., Mikšátko, J.: Supporting collaborative learning and e-discussions using artificial intelligence techniques. *Int. J. Artif. Intell. Educ.* **20**(1), 1–46 (2010)
34. McManus, M.M., Aiken, R.M.: Supporting effective collaboration: using a rearview mirror to look forward. *Int. J. Artif. Intell. Educ.* **26**(1), 365–377 (2016)
35. Meyer, A.S., Sleiderink, A.M., Levelt, W.J.: Viewing and naming objects: eye movements during noun phrase production. *Cognition* **66**(2), B25–B33 (1998)
36. Narayan, P.K., Smyth, R.: Electricity consumption, employment and real income in australia evidence from multivariate granger causality tests. *Energy policy* **33**(9), 1109–1116 (2005)

37. Olsen, J.K., Belenky, D.M., Alevan, V., Rummel, N.: Using an intelligent tutoring system to support collaborative as well as individual learning. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 134–143. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07221-0\\_16](https://doi.org/10.1007/978-3-319-07221-0_16)
38. Prieto, L.P., Sharma, K., Dillenbourg, P.: Studying teacher orchestration load in technology-enhanced classrooms. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 268–281. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24258-3\\_20](https://doi.org/10.1007/978-3-319-24258-3_20)
39. Raca, M., Dillenbourg, P.: System for assessing classroom attention. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 265–269. ACM (2013)
40. Richardson, D.C., Dale, R., Kirkham, N.Z.: The art of conversation is coordination. *Psychol. Sci.* **18**(5), 407–413 (2007)
41. Rosatelli, M.C., Self, J.A.: A collaborative case study system for distance learning. *Int. J. Artif. Intell. Educ.* **14**(1), 97–125 (2004)
42. Sangin, M., Molinari, G., Nüssli, M.A., Dillenbourg, P.: Facilitating peer knowledge modeling: effects of a knowledge awareness tool on collaborative learning outcomes and processes. *Comput. Hum. Behav.* **27**(3), 1059–1067 (2011)
43. Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., Pea, R.: Using mobile eye-trackers to unpack the perceptual benefits of a tangible user interface for collaborative learning. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **23**(6), 39 (2016)
44. Schwarz, G., et al.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
45. Sharma, K., Alavi, H.S., Jermann, P., Dillenbourg, P.: A gaze-based learning analytics model: in-video visual feedback to improve learner’s attention in MOOCs. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 417–421. ACM (2016)
46. Sharma, K., Caballero, D., Verma, H., Jermann, P., Dillenbourg, P.: Looking at versus looking through: a dual eye-tracking study in MOOC context. In: International Society of the Learning Sciences, Inc. [ISLS] (2015)
47. Sharma, K., Jermann, P., Nüssli, M.A., Dillenbourg, P.: Understanding collaborative program comprehension: Interlacing gaze and dialogues. In: Proceedings of Computer Supported Collaborative Learning (CSCL 2013), vol. 1, pp. 430–437 (2013)
48. Slavin, R.E.: Research on cooperative learning and achievement: what we know, what we need to know. *Contemp. Educ. Psychol.* **21**(1), 43–69 (1996)
49. Soller, A., Martínez, 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**(4), 261–290 (2005)
50. Thornton, D.L., Batten, D.S.: Lag-length selection and tests of granger causality between money and income. *J. Money Credit Banking* **17**(2), 164–178 (1985)
51. Vanlehn, K.: The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **16**(3), 227–265 (2006)
52. Viswanathan, S.A., VanLehn, K.: High accuracy detection of collaboration from log data and superficial speech features. International Society of the Learning Sciences, Philadelphia (2017)
53. Walker, E., Rummel, N., Koedinger, K.R.: Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *Int. J. Comput.-Support. Collab. Learn.* **6**(2), 279–306 (2011)

54. 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)
55. Wilby, R.: Decadal climate forecasting techniques for adaptation and development planning. Report for DfID (2007)



# Towards an Automated Model of Comprehension (AMoC)

Mihai Dascalu<sup>1,2,3</sup>✉, Ionut Cristian Paraschiv<sup>1,3</sup>, Danielle S. McNamara<sup>4</sup>,  
and Stefan Trausan-Matu<sup>1,2,3</sup>

<sup>1</sup> Department of Computer Science, University Politehnica of Bucharest,  
060042 Bucharest, Romania

{mihai.dascalu, ionut.paraschiv, stefan.trausan}@cs.pub.ro

<sup>2</sup> Academy of Romanian Scientists, Splaiul Independenței 54,  
050094 Bucharest, Romania

<sup>3</sup> Research Technology S.R.L., Sos. Virtutii, nr. 19D, Bucharest, Romania

<sup>4</sup> Institute for the Science of Teaching and Learning, Arizona State University,  
Tempe, AZ 85287-2111, USA  
dsmcnama@asu.edu

**Abstract.** Reading is a complex cognitive process wherein learners acquire new information and consolidate their knowledge. Readers create a mental representation for a given text by processing relevant words that, along with prior inferred concepts, become activated and establish meaningful associations. Our automated model of comprehension (AMoC) uses an automated approach for simulating the ways in which learners read and conceptualize by considering both text-based information consisting of syntactic dependencies, as well as inferred concepts from semantic models. AMoC makes use of cutting edge Natural Language Processing techniques, transcends beyond existing models, and represents a novel alternative for modeling how learners potentially conceptualize read information. This study presents side-by-side comparisons of the results generated by our model versus the ones generated by the Landscape model.

**Keywords:** Comprehension modeling · Semantic models  
Natural Language Processing · Landscape Model

## 1 Introduction

Reading is a complex cognitive process, which has been subject to many studies throughout the years. It is one of the most common means that learners use to acquire new information and consolidate existing knowledge. Moreover, text resources represent one of the primary sources for learning. Readers create mental representations, which includes previous knowledge, enabling them to comprehend the text. However, text materials are not customized depending on the individual reader, and they are usually addressed to specific categories of readers. As such, computational models that simulate the reading process can serve as important tools for creating personalized learning applications that support the educational process by presenting adequate materials to learners.



The Construction-Integration model [1] represents a semi-automated approach to simulating the comprehension process, extracting the information from a text and combining it with the reader's personal experience. The model is based on a cyclical process using sentence units and requires manually setting the words' initial activation scores that appear within the text as well as the connections between the words (or nodes). The CI model's construction process has two phases, each responsible for generating concepts and propositions using a different input set. The first phase, known as text-based construction, represents the initial activation of elements from the linguistic, semantic, and situation levels. During the second phase, the knowledge-based constructions are integrated using vector multiplication along with constraint satisfaction, wherein the various propositional nodes' activation levels and links to other nodes are modified depending on their relations in the network.

The aim of this paper is to introduce a novel state-of-the-art automated model of comprehension that can be used to simulate text reading for different categories of learners by employing different parameters and semantic models. In the next section we present two similar models, namely the CI and Landscape models, which are two of the most frequently employed models of comprehension. In the third section we introduce our automated model based on advanced Natural Language Processing (NLP) techniques, alongside a detailed comparison of the results obtained using the Landscape model. The last section concludes the paper and presents future experiments and improvements for our model.

## 2 Similar Models

### 2.1 The Construction Integration Model

The Construction-Integration (CI) model [1] represents a semi-automated approach that extracts the information from a text and combines it with the reader's personal experience. The CI model describes a framework used for studying memory in the form of a semi-automated computational model inspired from the way humans read and understand texts. The model is based on a cyclical process using sentence units and requires setting manually the words' activation scores that appear inside the text. The CI theory uses a bottom up approach that combines features from a symbolic system and from a connectionist system. On the one hand, the symbolic system consists of a rule-based system used to construct a network representation of the text and the activated words. On the other hand, the connectionist system uses a constraint satisfaction mechanism to generate a stabilized (or coherent) interpretation of the to-be-comprehended text.

The CI model's construction process has two phases, each one responsible for generating concepts and propositions using a different input set. The first phase, also known as text-based construction, combines elements from the linguistic, semantic, and situation levels. Linguistic elements are equivalent to syntactic links and have been neglected in many studies as they only reflect a surface level of comprehension. The semantic level uses rules to generate text propositions, which represent concepts regardless of form (i.e., including images). The situation level is topic specific and relies on domain and general knowledge to make inferences to generate links among concepts in

the text. The second phase uses knowledge-based constructions, where various propositional nodes are added and which can vary in strength, depending on their relations. Each propositional node within the knowledge construction phase has an explicit similarity relatedness with a text-based node.

The CI model builds a square term matrix  $C$  containing  $n + m$  elements, where  $n$  represents the number of words, propositions or concepts that appear in the text, and  $m$  the knowledge propositions selected from the long-term memory net or in response to specific task demands. The model makes use of subsequent multiplications of the manual input activation row vector  $A_1$  with the term matrix until the change in mean activation value is less than some criterion value ( $A_i = A_{i-1} * C$ ). In the end, the model generates the final activation vector  $A$ , which provides the predicted strength for each unit. The CI model also provides a long-term square matrix  $M$  with  $p + q$  elements, where  $M_{ij} = C_{ij} * A_i * A_j$  (See Fig. 1).

$$\begin{aligned}
 A &= [a_i \dots a_j \dots a_n \dots a_{n+m}] \\
 C &= \begin{bmatrix}
 & a_1 & \dots & a_j & \dots & a_n & \dots & a_{n+m} \\
 a_1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_i & \cdot & \cdot & c_{ij} & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_n & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_{n+m} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
 \end{bmatrix} \\
 M &= \begin{bmatrix}
 a_1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_i & \cdot & \cdot & c_{ij} * a_i * a_j & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_p & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 a_{p+q} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
 \end{bmatrix}
 \end{aligned}$$

**Fig. 1.** Activation vectors, coherence matrix and long-term memory matrix corresponding to the CI model.

Albeit an impactful theoretical model, the CI model lacks some aspects of automation. First, the activation scores from each step must to be added manually before the model is able to distribute them in the current cycle. Second, the knowledge expansion is also performed by hand, thus making hard to generalize the approach. These limitations have put on hold the further development of the model because more advanced validations have been challenging to realize without automation capabilities.

## 2.2 The Landscape Model

The *Landscape Model* [2] has been designed to simulate the fluctuation of the concepts' activation scores, similarly to the CI Model. The concepts' activation is set manually through strategic assumptions about the source of activation and the amount of activation [3]. Prior knowledge activation is achieved through two different mechanisms: cohort activation and coherence-based retrieval. The first mechanism serves the function of passively mapping related concepts to the reader's mental representation of the text [3]. Concepts are inter-connected forming cohorts or associative memory traces, and the whole group can be activated at once by simply activating one word within the text. The second mechanism, coherence-based retrieval, uses a coherence parameter ranging from 1 to 5 that represents a word's importance with regards to certain relations from the text (causal, temporal, or spatial connections): more superficial reading processes are represented by smaller parameter values.

A visual representation of results from the Landscape Model are presented in Fig. 3, alongside the target text (see Fig. 2), depicting how the words' activation scores theoretically evolved across subsequent sentences [2].

*A young knight rode through the forest (1).*

*The knight was unfamiliar with the country (2).*

*Suddenly, a dragon appeared (3).*

*The dragon was kidnapping a beautiful princess (4).*

*The knight wanted to free her (5).*

*He wanted to marry her (6).*

*The knight hurried after the dragon (7).*

*They fought for life and death (8).*

*Soon, the knight's armor was completely scorched (9).*

*At last, the knight killed the dragon (10).*

*He freed the princess (11).*

*The princess was very thankful to the knight (12).*

*She married the knight (13).*

**Fig. 2.** The “Knight” story. Sample text used for visualizing the Landscape Model (<http://www.brainandeducationlab.nl/downloads>).

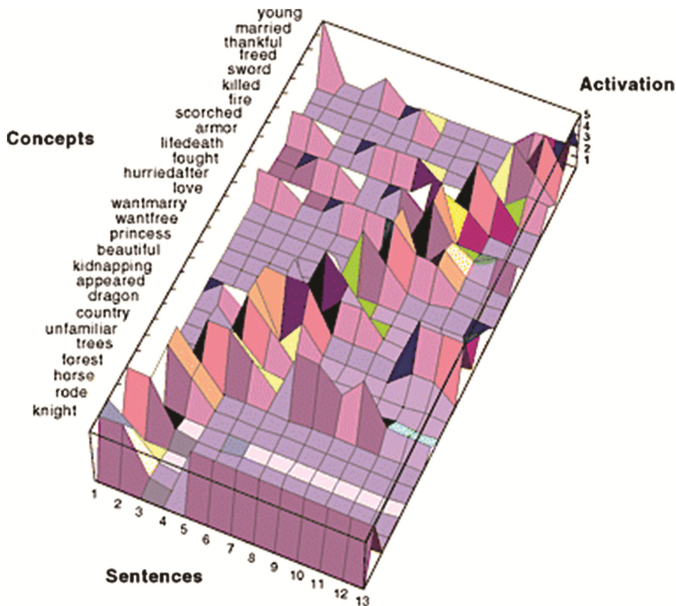


Fig. 3. Visualizing activation scores within the Landscape Model (<http://www.brainandeducationlab.nl/downloads>).

### 3 Current Study

Our automated model of comprehension (AMoC) introduces a fully automated method that analyzes the way in which readers potentially assimilate and conceptualize new text information. AMoC was developed on top of the *ReaderBench* framework [4], containing an extensive set of tools and models to analyze unstructured corpora. *ReaderBench* implements Cohesion Network Analysis which provides an in-depth perspective of discourse by relying on cohesive links identified between different text constituents [5]. Moreover, it contains a wide range of textual complexity indices covering syntactic, semantic and discourse structure levels of text analysis [4].

In its current form, AMoC makes use of lexicalized ontologies to determine synonyms that are used for semantic expansion. The system uses WordNet [6], a frequently used ontology in English, containing more than 150.000 concepts. These inferred words are subsequently compared to the rest of the concepts by using semantic models, representing pre-trained models that associate vectors to textual resources so that their semantic distance can be estimated through their cosine similarity. In our current implementation, we opted to rely on two representative and frequently used semantic models. First, Latent Semantic Analysis [7] creates a term-document matrix which counts words' appearances and applies Singular Value Decomposition followed by a dimensionality reduction. Second, the word2vec model, introduced recently in the literature [8, 9], adds support for words with multiple degrees of similarity along with inflections, and makes use of algebraic operations to determine meaningful similarities and links.

AMoC focuses on viewing a dataset from a micro level, in other words analyzes textual resources individually. The focus is on individual paragraphs, which are analyzed automatically through techniques similar to the way people read texts in general. Humans tend to create mental representations for the words encountered in the text, which in return activate other concepts from their memory. This process results in textual annotations that can be used later on to suggest which are the key points in every paragraph driving the evolution of topics within the text. The textual annotations can be the basis of intelligent reading applications used to help learners to understand textual resources better, even without reading them. AMoC, in its current form, analyzes each paragraph separately, only linking activation scores across sentences.

By reutilizing some basic principles from the CI model along with novel activation scores' computing and concepts' inference, AMoC is a novel approach that fully automates the textual annotations with activation scores. The model analyzes the input text in order to determine which are its most important words. It also infers semantically related words from lexicalized dictionaries to simulate the memory and/or knowledge of the reader. The main idea is that some words from the text are able to activate other terms that are not explicit in the text, but are theoretically available in prior knowledge: e.g., if someone reads a text that contains the word "cat", the concept may activate other concepts such as "feline", "lion", or "tiger" based on the semantic context.

Each sentence is comprised of at least one text-based word, further enriched through its dictionary synonyms, so the graph grows proportionally with the number of sentences and content words. Thus, an activation score is imposed that must be exceeded by all the words within the graph to be considered active. The reason behind this implementation choice is that when reading, humans have a short-term memory consisting of a global context with many inactive and only a few active concepts.

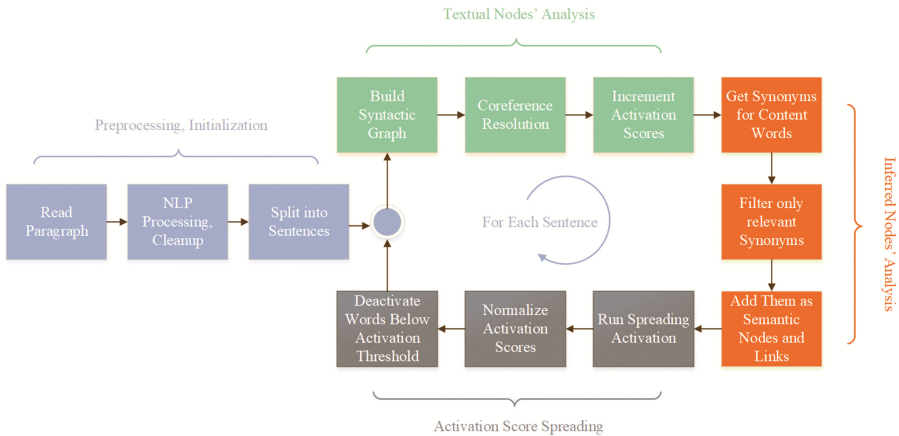


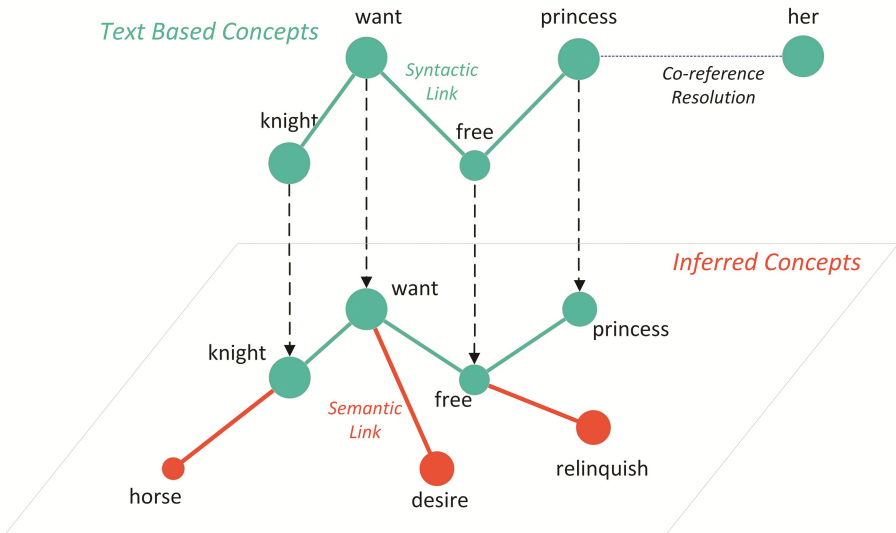
Fig. 4. Automated model of comprehension workflow.

Figure 4 depicts the implemented workflow that uses two types of links: syntactic links reflecting text-based associations between words, and semantic links highlighting semantic relatedness above an imposed threshold in semantic models. In the current

analyses, we opted to use word embeddings from word2vec that provided the highest correlations with the activation scores from the Landscape Model, but other semantic models can be easily employed (e.g., Latent Semantic Analysis). The Landscape and CI models were not implemented within the current research and they represent only inspirational models.

During the preprocessing phase, the document undergoes a complete Natural Language Processing (NLP) pipeline which: cleans the input text, splits it into paragraphs and sentences, removes stop words, applies lemmatization, performs part-of-speech tagging and identifies content words (i.e., nouns, verbs, adjectives and adverbs), identifies syntactic dependencies, and replaces pronouns with corresponding nouns using pronominal resolution [10].

Next, the sentence’s syntactic graph is extracted and merged within the global network graph depicting the memory’s state. The activation scores corresponding to all content words from the sentences are incremented by 1. Afterwards, synonyms are extracted for all the content words using the WordNet ontology [6]. Only the most relevant synonyms (those having the highest semantic correlation with the whole text so far) are retained and merged alongside their corresponding semantic associations within the global network graph. Figure 5 depicts a use case of our model for the fifth sentence in the original text from Fig. 2, in which the semantic and syntactic links are shown, together with the mechanism of co-reference resolution.



**Fig. 5.** Use case for “The knight wanted to free her”.

Subsequently, spreading activation derived from the PageRank algorithm [11] is applied to distribute the activation strengths within the network, and only a limited number of words remain active (or words above a normalized activation score); follow-up sentences are treated in a similar manner by the model. The activation scores for each sentence are saved in order to render the three-dimensional visualizations (terrain

rendering similar to the Landscape Model, 3D bar-charts and an evolutionary grid) in Figs. 6 and 7. As it can be easily observed, the most active concept is “knight”, followed by “princess” and “dragon”, central concepts within the presented story. These visualization techniques aim at depicting the evolution of the words’ activation scores across subsequent sentences.

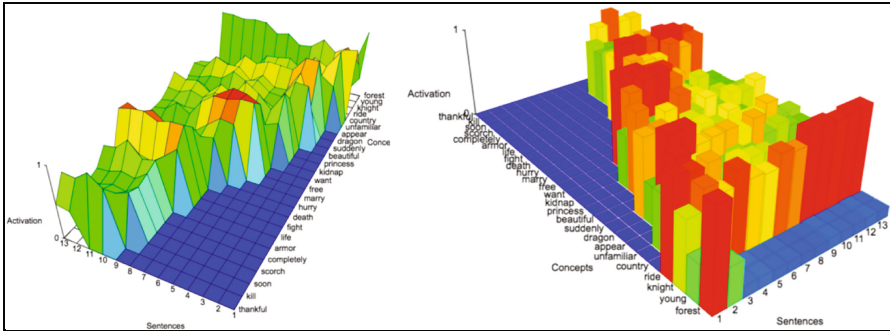


Fig. 6. Visualizing activation scores within AMoC using 3D bar-charts.

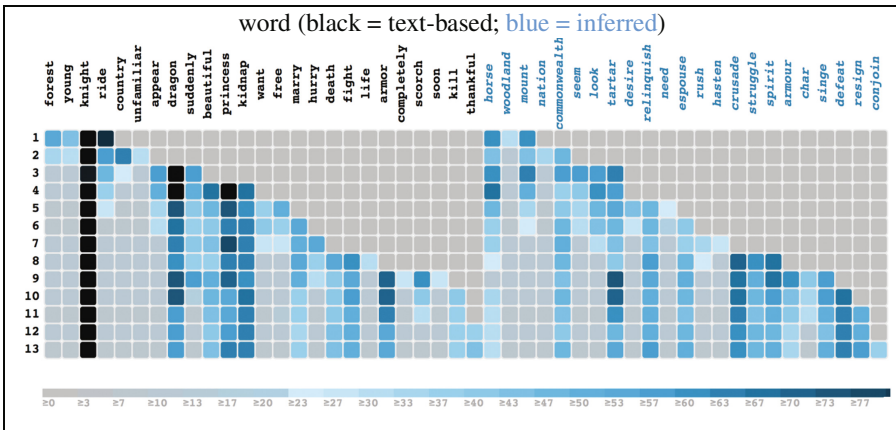


Fig. 7. Visualizing activation scores within AMoC using an evolutionary grid.

Statistical analyses were conducted to assess the extent to which AMoC measures words’ activation scores in relation to predictions reported for the Landscape Model. The Landscape Model’s activation scores were reported in previous experiments [12], and its values were compared with the ones generated from our model. The results of the experiment yielded high correlations (80%) between the activation scores of our model and the ones from the Landscape Model applied on the text from Fig. 2 (see Table 1). As such, the two models derive similar predictions, though the our model is entirely automated.

**Table 1.** Correlations between the activation scores from the Landscape Model and AMoC for the sentences from the “Knight” story (\*\* $p < .001$ ).

Correlation	Sentence													Avg.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	
Pearson	.927**	.872**	.899**	.907**	.825**	.530**	.726**	.809**	.823**	.768**	.550**	.879**	.910**	<b>.802**</b>
Spearman	.898**	.924**	.932**	.810**	.739**	.462**	.547**	.670**	.592**	.554**	.557**	.569**	.737**	<b>.692**</b>

## 4 Discussion and Conclusions

In this paper we introduce AMoC, an automated method for modelling human reading. AMoC can be used to simulate comprehension and offers the means to manipulate variables within the model in order to make model-based predictions. Text learning materials can be personalized by employing AMoC and learners may be helped, for example, by having highlights of central ideas. Understanding and simulating the reading process is a central element towards creating more contextualized learning environments that enhance the assimilation of new information.

AMoC represents a textual analysis tool that utilizes various Natural Language Processing techniques along with CI methods. The model annotates unstructured text with various computational methods that can be used for many purposes. First, the graphs’ nodes are textual units which can be linked with various ontological facts, thereby simulating the activation of the semantic meaning by the reader. Secondly, understanding the most important words across sentences can be utilized for computing the overall textual complexity and to link readers to more detailed explanations.

In addition, understanding the reading process represents a key point in creating more contextualized learning environments that enhance the assimilation of new information and present learning materials tailored to the student’s level. Moreover, this feature can enhance a student’s learning experience in the context of cluttered domains with unstructured information. In other words, it can be a very useful tool in any educational context that relies on reading activities. AMoC was shown to have a high correlation (80%) with the results presented in the Landscape Model, an initial validation on top of which other experiments and validations can be built.

AMoC represents a completely autonomous method for simulating the human reading process. Besides validating the model, there are some parts which can be improved. Verbs tend to be more generic than other words, thus their semantic expansion is usually larger than nouns or adjectives. Furthermore, the static activation threshold seems to not be sufficient for filtering a small number of active concepts, so it should be replaced with a dynamic function. The current version of AMoC analyzes only activation scores within the sentences from a paragraph, thus we need to also account for activation scores between different paragraphs, which are already defined in the literature. Nonetheless, it represents a solid step towards a completely automated model of comprehension.

However, this is clearly only a first, albeit significant step. For example, one limitation is that the model currently focuses solely on a local analysis of texts, addressing only the short-term memory cycle (i.e., AMoC analyzes each paragraph separately, only linking activation scores across sentences). Additionally, various parameters need to be



further tuned and the model needs to be subjected to extensive validations. Our current work is focusing on further extensions of the model and assessing the model's validity by comparing its predictions against prior research findings in the discourse literature.

**Acknowledgment.** The work presented in this paper was funded by the European Funds of Regional Development with the Operation Productivity Program 2014–2020 Priority Axe 1, Action 1.2.1 D-2015, “Innovative Technology Hub based on Semantic Models and High Performance Computing” Contract no. 6/1 09/2016.

## References

1. Kintsch, W., Welsch, D.M.: The Construction-Integration Model: A Framework for Studying Memory for Text, p. 21. Institute of Cognitive Science, Boulder (1991)
2. van den Broek, P., Young, M., Tzeng, Y., Linderholm, T.: The landscape model of reading. In: van Oostendorp, H., Goldman, S.R. (eds.) *The Construction of Mental Representations During Reading*, pp. 71–98. Erlbaum, Mahwah (1999)
3. McNamara, D.S., Magliano, J.: Toward a comprehensive model of comprehension. *Psychol. Learn. Motiv.* **51**, 297–384 (2009)
4. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learner productions and strategies with *ReaderBench*. In: Peña-Ayala, A. (ed.) *Educational Data Mining. SCI*, vol. 524, pp. 345–377. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-02738-8\\_13](https://doi.org/10.1007/978-3-319-02738-8_13)
5. Trausan-Matu, S., Stahl, G., Sarmiento, J.: Polyphonic support for collaborative learning. In: Dimitriadis, Y.A., Ziguers, I., Gómez-Sánchez, E. (eds.) *CRIWG 2006. LNCS*, vol. 4154, pp. 132–139. Springer, Heidelberg (2006). [https://doi.org/10.1007/11853862\\_11](https://doi.org/10.1007/11853862_11)
6. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
7. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* **104**(2), 211–240 (1997)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: *Workshop at ICLR, Scottsdale* (2013)
10. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing toolkit. In: *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. ACL, Baltimore (2014)
11. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. *Stanford InfoLab* (1999)
12. Britton, B.K., Graesser, A.C.: Models of understanding text. In: Britton, B.K., Graesser, A.C. (eds.) *Models of Understanding Text*. Psychology Press, New York (1995)



# Course-Adaptive Content Recommender for Course Authoring

Hung Chau<sup>(✉)</sup>, Jordan Barria-Pineda, and Peter Brusilovsky

School of Computing and Information, University of Pittsburg, Pittsburgh, PA, USA  
{hkc6,jab464,peterb}@pitt.edu

**Abstract.** Developing online courses is a complex and time-consuming process that involves organizing a course into a sequence of topics and allocating the appropriate learning content within each topic. This task is especially difficult in complex domains like programming, due to the incremental nature of programming knowledge, where new topics extensively build upon domain concepts that were introduced in earlier lessons. In this paper, we propose a course-adaptive content-based recommender system that assists course authors and instructors in selecting the most relevant learning material for each course topic. The recommender system adapts to the deep prerequisite structure of the course as envisioned by a specific instructor, while unobtrusively deducing that structure from problem-solving examples that the instructor uses to present course concepts. We assessed the quality of recommendations and examined several aspects of the recommendation process by using three datasets collected from two different courses. While the presented recommender system was built for the domain of introductory programming, our course-adaptive recommendation approach could be used in a variety of other domains.

**Keywords:** Learning content recommendation · Course model

## 1 Introduction

Over the past twenty years, most intelligent tutoring systems (ITSs) have focused their personalization efforts on helping students find an “optimal path” through available learning content to achieve their learning goals. A range of personalization technologies, known as course sequencing, adaptive navigation support, and content recommendation, can account for the learning goals and the current state of student knowledge and recommend the most appropriate content (e.g., a problem, an example, an educational video, etc.). However, in the context of real courses, there is not complete freedom in selecting the appropriate content for students. An instructor usually plans a course as a sequence of topics to be learned. To stay in sync with the instructor and the class, students are expected to work on course topics in the order that is determined by the instructor’s plan. In this context, the personalized selection of learning content should account for both a student’s prospects (i.e., current knowledge levels) and the instructor’s prospects (the preferred order of topics and learning goals).

Unfortunately, the current generation of ITSs rarely support adaptation to a teacher's preferences. In most of these systems, a sequence of topics is predefined and learning content items are statically assigned to these topics. While this approach works well for instructors who are happy to follow the sequence of topics that is defined by the ITS, the instructors who prefer a different topic structure will find such a system unacceptable, since it doesn't support their approach to teaching the course. These considerations are especially important when learning programming, where almost every instructor and every textbook introduces a unique course organization [1,2].

Nowadays, a variety of learning content items could be accessed from different learning content repositories and portals [3,4], while the majority of learning management systems offer authoring tools to structure a course into a set of topics and to add learning content to each topic. However, our work with instructors revealed that limited assistance provided by the current course authoring tool is not sufficient. While defining a sequence of topics is an easy task, selecting the most relevant content for each topic from a large collection of advanced learning content items is a real challenge. The instructors need to carefully review a large number of problems and examples in order to select those that best fit their learning goals for the topic. This is a time-consuming and error-prone process [5,6]. While a number of recommender systems have been developed to assist instructors in finding relevant content in online repositories [7], these systems attempt to adapt to the overall goals and interests of their users and are not able to consider the complex prerequisite-based structures of modern courses.

This paper presents *Content Wizard*, a content recommender system that has been specifically created to assist instructors with the course authoring process by recommending learning activities that are most appropriate to each of the course topics in the context of their preferred model of the course. The system leverages two valuable resources provided by instructors: *the order of course topics* and *problem-solving examples*, which instructors (or textbook authors) present to students to demonstrate course concepts.

This paper is organized as follows: Sect. 2 presents a brief review of related work, while Sect. 3 discusses the place of content recommendation in a course-authoring context and presents the interface of Content Wizard. The internal organization of the system and its recommending approach is described in Sect. 4. Section 5 presents an evaluation of Content Wizard's performance against a more traditional baseline, and Sect. 6 examines the performance of the approach on a deeper level. Finally, Sect. 7 presents a discussion of results and possible avenues for future work.

## 2 Related Work

The problem of authoring support in an ITS context has been extensively explored. Murray [5] defines seven categories of ITS authoring tools and generally classifies them into two broad groups: pedagogy-oriented systems or performance-oriented systems. *Performance-oriented systems* focus on providing a rich educational environment, in which students can gain problem-solving

expertise, procedural skills, concepts, and facts by practicing and receiving feedback and guidance from tutors. Authoring tools in this group include simulation-based learning, domain expert systems, and some special purpose systems. The prominent examples in this category are ASPIRE and cognitive tutor authoring tools (CTATs). ASPIRE [8] allows non-computer scientists to develop new constraint-based tutors with main support for generating a domain model and producing a fully functional system. Cognitive tutor authoring tools (CTATs) [9] allow authors to develop two types of tutors: cognitive tutors and example tracing tutors. The CTAT authoring process requires authors to give a definition of a task domain (such as the fraction addition problem), along with appropriate problems. *Pedagogy-oriented systems* focus on organizing instructional units and tutoring strategies. They support instructors in managing curriculum sequencing and planning, designing teaching strategies and tactics, composing multiple knowledge types (e.g., topics and concepts), and authoring adaptive hypermedia. Two examples in this category are InterBook and SitPed. InterBook [10] provides support for authoring adaptive electronic textbooks. It helps authors to create the book's structure and associate every section to domain concepts. SitPed [11] is a pedagogy-oriented authoring system that supports instructors in creating simple, hierarchical task models, authoring assessment knowledge, and creating tutor feedback and guidance.

Authoring tools in the pedagogy-oriented group frequently focus on supporting authors by defining the domain model as a set of knowledge components (concepts or rules), building a course structure, and associating course units and learning content with domain concepts [10, 12–14]. While our work follows the same approach, we minimize authors' load by deriving the intended course structure from easily available data, rather than requiring the authors to manually provide their intended course structure. The most recent systems in this group also offer learning content recommendation for course authors [7]; yet in most cases, content recommendation is based on instructor interests or on specific goals, and less than a handful of projects [15] focused on using the whole course structure for content recommendation. Our work attempts to advance this research direction by exploring a more powerful yet unobtrusive recommendation approach and by offering a more extensive evaluation than earlier efforts.

### 3 Content Recommendation for Course Authoring

The content recommendation approach presented in this paper was developed for a typical course authoring context. The essence of this scenario is that the author designs the course structure as a sequence of *topics*. To support learning for each topic, a set of items of multiple content types is associated with each topic. This course structuring approach is supported by every major learning management system (where a topic usually corresponds to a course lecture), as well as by most textbooks (where a topic usually corresponds to a chapter).

To facilitate this course-authoring approach and to make it easier for instructors to reuse large volumes of “smart” learning content for computer science

education [16], we developed a drag-and-drop course authoring tool (Fig. 1). The tool allows course authors to define a sequence of topics (shown on the left), select their preferred types of learning activities (the authors could use over 15 types in three domains), and add the desired content to each topic. To add content to a topic, the author selects one topic and one type of learning content (in Fig. 1, the topic *Variables* and the *Problems* content type is selected). Following that, the authoring system shows a list of all activities of the selected type that could be added to the topic through a drag-and-drop interface. The key problem here, as we discovered when working with an early version of the tool, is that this interface doesn't really help the authors to select precisely the right kind of content for each topic. This task is quite difficult: each selected item should cover the learning concepts that the instructor wants to introduce in the selected topic, but at the same time, it should not use concepts that students have not learned yet. The sheer amount of content to consider (e.g., 289 *interactive examples* and 223 *programming problems*) makes this task practically impossible without additional support.

The goal of Content Wizard is to provide the necessary support to make content selection both feasible and efficient. As the right side of Fig. 1 shows, Content Wizard ranks all content items of the selected type by its match to the selected topic in the course context. It also assigns "star" relevance ratings to all content items and puts a warning sign to items that, while superficially a good match, might include concepts that have not yet been introduced at this point in the course. The most important aspects of support offered by Content Wizard are: (1) this support is based on Content Wizard's understanding of the fine-grained course structure and prerequisite relationships on the level of domain concepts; (2) the instructor is not expected to define the fine-grained course structure, as required by some earlier approaches [12,13]; instead, the course's structure is automatically derived from the order of course topics and the set of code examples that instructor shows at the target lecture (or provides in a book chapter). The next section explains this approach in detail.

## 4 Course-Adaptive Content Recommender System

### 4.1 The Course Model

The content recommendation that is provided by Content Wizard is based on a deeper-level course structure modeling. While the instructor may perceive the course as a sequence of topics, the deeper model assumes that the goal of each topic is to introduce a set of fine-grained domain concepts (knowledge components). The course model is defined as a sequence of concept sets that are covered throughout the course (see Fig. 2). The concepts associated with a specific course unit are the concepts that instructor aims to *introduce at that unit*. For example, Unit 2 is the first unit of the course where the concepts *Array Variable* and *Array Data Type* appear, among others. The concepts introduced in earlier units become prerequisite concepts for later units. For example, *Print*, *Int Data Type*, and other Unit 1 concepts are expected to be learned before students start

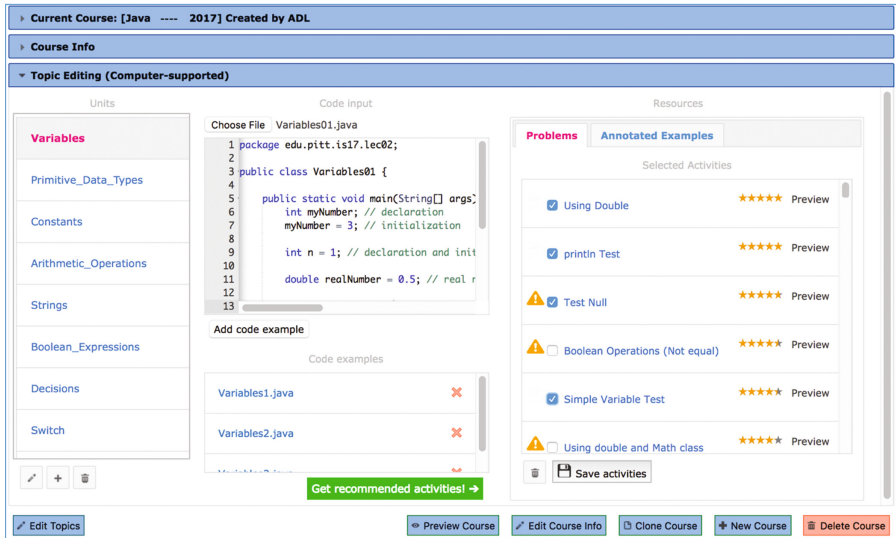


Fig. 1. Course-authoring tool with Content Wizard recommendations

Unit 2, and are considered prerequisites for Unit 2 and all following units. This deeper level concept-based course modeling is popular among ITS and Adaptive Hypermedia authoring systems [10, 12, 13] where it is assumed that course or system authors will create this model manually. The difficulty of manual modeling is a known bottleneck of the fine-grain course structuring approach, which prevents this approach from being more broadly used. However, Content Wizard is able to automatically derive this model by using worked examples that are provided by the course authors.

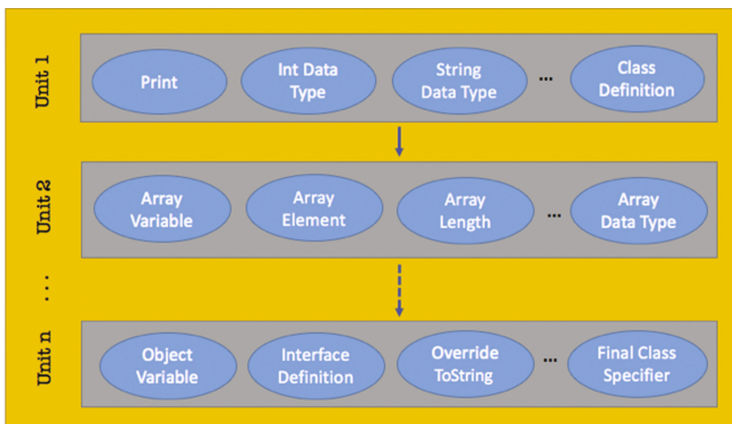


Fig. 2. Knowledge structure of a course

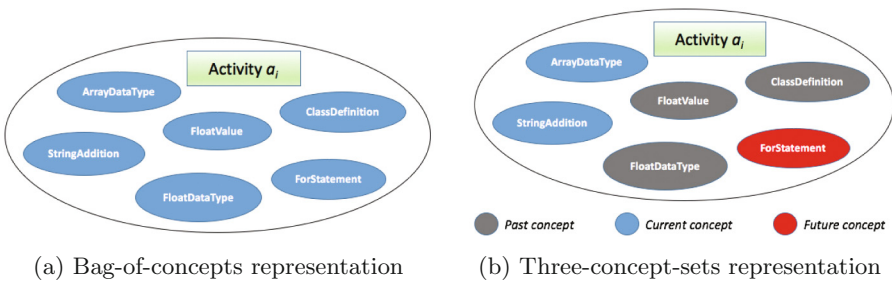
### 4.2 Worked Code Examples

Worked examples, in the form of complete programs or code fragments, are extensively used in teaching programming concepts. In each lecture, an instructor usually presents several worked examples that illustrate newly introduced concepts. Similarly, each programming textbook extensively uses examples and frequently offers access to the code of these examples through a CD included with the textbook or a Web site. The assumption behind the Content Wizard’s automatic course structuring is that a set of examples presented for each unit offers the best way to understand the concepts that the instructors want to introduce in this unit. To build a deep course model that follows the instructor’s preferences, Content Wizard asks the course author to submit the plain code of each example that is used in the unit (see the middle column in Fig. 1).

Using the code examples provided for each unit, Content Wizard automatically creates course knowledge structure, as shown in Fig. 2. First, it extracts all programming concepts that are associated with each code example using a Java concept parser [17], that returns a set of fine-grained concepts from Java ontology. Second, for each unit, it forms a set of *covered concepts* that merges concepts from all of the unit’s examples. Finally, it sequentially processes the units to define the unit’s content as concepts that are first introduced in this unit; i.e., all concepts extracted from Unit 1 examples become Unit 1 concepts; all *new concepts* extracted from Unit 2 examples (i.e., those that have not been introduced in Unit 1) become Unit 2 concepts, and so on.

### 4.3 Content Representation and Analysis

To identify a match between a unit and a learning activity, Content Wizard considers a set of concepts associated with a candidate activity and the course structure. Since all types of of learning activities available in the system (i.e., examples or problems) include code fragments, we use the Java concept parser [17] to represent each activity as a “bag” of Java programming concepts (Fig. 3a). This “bag of concepts” representation could be used by a number of traditional recommendation algorithms. A match to a specific unit, however, depends on



**Fig. 3.** Demonstration of representing learning content as programming concepts.

the position of the target unit within the course. When selecting an activity, instructors usually consider the balance of practicing newly introduced knowledge and reviewing learned knowledge, because students are most engaged when the material to be learned is neither too difficult nor too easy. Wang et al. [2] classify a learning activity in the progression as *reinforcement* (reviewing learned concepts), *recombination* of previously learned concepts, or *introduction* (introducing new concepts). Using the course model presented in Sect. 4.1, Content Wizard classifies each concept that appears in an activity into one of three categories (Fig. 3b):

- *Past concepts (P)*: Concepts that were covered in previous units. These concepts are supposed to be known before starting the current unit.
- *Current concepts (C)*: Concepts that are covered in the current unit (and thus have not been covered in any previous units). We consider these concepts as targets of the current unit, according to the instructor’s vision of the course.
- *Future concepts (F)*: Concepts that have not been covered up to the current unit. We assume that the instructor prefers to cover these concepts in future units (or not to cover them at all). Most likely, these concepts are not yet appropriate for students to learn in the context of the unit.

This representation reflects instructor preferences and enables our recommendation approach, in which recommended activities focus on current concepts, leverage learned concepts, and avoid future concepts.

#### 4.4 The Recommendation Method

Content Wizard adaptively provides two valuable sources of information that can help instructors find the most appropriate content for each unit: a *ranked list* and *warning flags*. At every step of course creation, based on the current course model and the code examples provided for the target unit, the system will update the representations of all candidate learning activities during the recommendation process.

For each learning activity,  $a_i$ , consisting of three concept sets,  $P_i$ ,  $C_i$  and  $F_i$ , the Wizard calculates its ranking score by linearly combining the contribution of the concept covered by the activity, according to the category to which each concept belongs. Equation (1) shows how we compute each content ranking score:

$$score_{a_i} = \alpha|P_i| + \beta|C_i| + \gamma|F_i| \quad (1)$$

$\alpha$ ,  $\beta$ , and  $\gamma$  are the parameters controlling the importance of the three categories. The values for these parameters might be different for different domains and even for individual instructors, depending on how much they focus on current and past concepts and how much they want to avoid future concepts. Indeed, [2] shows the differences among the proportions of reinforcement, recombination, and introduction of concepts of two Japanese textbooks and two online learning tools (Duolingo and Language Zen). Given sufficient volume of data (content selected by instructors for different courses) these parameters could be learned



from data; otherwise, they could be selected by using expert estimation. We explored both of these approaches in our evaluation presented in Sect. 5.

In addition to the ranking, we believe it is important for instructors to be aware of “not ready” learning activities that use concepts that do not appear in the code examples up to the current unit, because they could confuse less prepared students. We identify these activities as follows:

$$warning_{a_i} = \begin{cases} 1, & \text{if } |F_i| > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Activities with *warning* value 1 are annotated using a warning icon (see the 3rd, 4th, and 6th rows in Fig. 1). The instructor can then evaluate whether an activity with potentially premature concepts should be assigned to the unit.

## 5 Evaluation

### 5.1 Datasets

To evaluate our proposed recommendation method, we collected three data sets from two different universities. Each dataset encapsulated instructor preferences in content selection (i.e., “ground truth”).

**Dataset 1:** The data was collected from a Java class taught at the University of Pittsburgh in Fall 2016 (referred as IS17F16). The instructors followed a lecture-based format and created a course structure for IS17F16 that consists of 18 units (each unit includes two types of learning content, *annotated examples* and *parameterized problems*). No content recommendation functionality was used for content selection. As input, we collected code examples presented by the instructors in the course slides. All *annotated examples* in the content pool were used for ranking. As the ground truth, we used *annotated examples* that were selected by the instructors for each unit.

**Dataset 2:** The second dataset uses the same inputs as the first dataset for running the recommendation process. All items in the *problem* pool are ranked for recommendation (as shown in Fig. 1), and the ground truth is the *problems* selected by the instructors for each unit of IS17F16.

**Dataset 3:** This dataset was extracted from the CS1 online programming course<sup>1</sup> taught at University of Helsinki, Finland. We mapped the course structure into ten coherent topics. Each topic has several *code examples* for students to learn new concepts (which we used as input) and several *coding exercises* to practice (which we used as the ground truth). Note that in this dataset, only *coding exercises* that were actually used in the course were ranked in the recommendation process, while in datasets 1 and 2, all items in the content pool were ranked for the course, including those that were not selected by the instructor.

<sup>1</sup> <http://mooc.fi/courses/2013/programming-part-1/material.html>.

## 5.2 Performance Comparison

To assess the performance of our approach in a fair way, we estimate the values of the parameters in Equation (1) based on a preliminary analysis (we can't learn it from the same data that we use to evaluate the approach). We collected code examples of several topics from a Java programming textbook and ran the algorithm using Eq. 1 while adjusting the parameter values in order to get the best recommendation results (by taking the book's contents as ground truth). The estimated values of  $\alpha$ ,  $\beta$ , and  $\gamma$  are set to **0.2**, **1**, and **-1.5**, respectively. As a baseline ranking approach, we use a popular content-based approach that ranks candidate items by cosine similarity between concept vectors that represent units and content items [15]. We refer to this approach as *tf\*idf*, since we use a TF\*IDF approach to assign weights for individual concepts in the concept vector. To measure ranking performance, we use three classical metrics: precision, recall, and F1 score (at top 3, top 5, top 10, and top 15).

**Table 1.** Performance comparison of Content Wizard vs. the baseline

Dataset	Method	Precision@top (%)				Recall@top (%)				F1@top (%)			
		3	5	10	15	3	5	10	15	3	5	10	15
1	<b>Wizard</b>	<b>62.74</b>	<b>50.59</b>	<b>34.7</b>	<b>25.09</b>	<b>51.26</b>	<b>63.36</b>	<b>77.51</b>	<b>80.44</b>	<b>56.42</b>	<b>56.26</b>	<b>47.94</b>	<b>38.26</b>
	tf*idf	21.57	18.82	15.29	15.68	21.57	18.84	29.14	42.33	15.57	18.84	20.06	22.89
2	<b>Wizard</b>	<b>47.05</b>	<b>37.64</b>	<b>32.94</b>	<b>26.66</b>	<b>34.23</b>	<b>41.91</b>	<b>66.32</b>	<b>76.86</b>	<b>39.63</b>	<b>39.66</b>	<b>44.01</b>	<b>39.59</b>
	tf*idf	21.57	17.65	14.11	14.11	17.63	23.05	33.33	43.08	19.40	20.00	19.83	21.27
3	<b>Wizard</b>	<b>96.3</b>	<b>88.89</b>	<b>75.76</b>	<b>64.44</b>	<b>41.45</b>	<b>52.79</b>	<b>73.32</b>	<b>83.90</b>	<b>57.96</b>	<b>66.24</b>	<b>74.42</b>	<b>72.89</b>
	tf*idf	81.48	73.33	66.67	57.04	37.24	44.97	64.34	73.18	51.12	55.75	65.48	64.10

As shown in Table 1, Content Wizard outperforms the *tf\*idf* method for all datasets. In all cases, Content Wizard archives a good recall performance of about 80% when presenting the top 15 results out of a pool of more than 200 learning items (i.e., 80% of all relevant items are included among the top 15 results). The table also shows interesting differences between the F1 performance of both approaches on datasets 1–2 and on dataset 3. First, on the dataset 1 and 2, the performance of the Wizard is 2–3 times better than the baseline while in the dataset 3 the difference is smaller. Second, the precision of both approaches is considerably higher for dataset 3. We believe that both differences stem from the nature of the datasets. The ranking tasks for the dataset 3 was much easier than for datasets 1–2. First, for dataset 3, recommender approaches had to rank only the actual items used in the course (and no “spares”). It was essentially a matter of matching each item to the best unit. The number of units to match was also much smaller (10 vs. 18). Recommendation for datasets 1–2 required ranking or all content items in the repository out of which only a part was used in the course. Some of these items might be a poor match to the course, but some were not used by the instructors when creating the course simply because they wanted to select *some* relevant content for each unit, but not *all* relevant

content. As shown by the data, a simple content-based algorithm might work reasonably well in simple cases, but in a more challenging (and realistic) context, Content Wizard offered a remarkable advantage.

## 6 Deeper Analysis

### 6.1 Finding the Best Values for the Parameters

As presented in Sect. 5.2, the values of  $\alpha$ ,  $\beta$ , and  $\gamma$  used in performance evaluation were selected using a preliminary analysis. The performance of our systems with estimated parameters was better than the baseline, but it might still be improved by learning best parameters from the data. Since the precise balance between past, current, and future concepts may depend on instructor's preferences, the proper way to learn parameters for performance evaluation would be to use another earlier-authored course from the same instructor (with sufficient volumes of instructor-selected content). Since we have only one course for each instructor, the parameters learned from this course could not be used to evaluate recommendation performance on the same course. However, we could still post-assess the effectiveness of the estimated parameters and explore how the quality of the recommendation depends on the value of the parameters.

To achieve this goal, we ran 100 iterations from 0 to 10 with an increment of 0.1 for each of the parameters, for a total of 1,000,000 iterations. We found that within this range, the best values of  $\alpha$ ,  $\beta$ , and  $\gamma$  w.r.t F1@15 performances are respectively **0.167**, **1**, and **-3** for dataset 1; **0.2**, **1**, and **-2.5** for dataset 2; and **0.125**, **1** and **-2.3** for dataset 3 (these values have been normalized by dividing all parameters by the values of  $\beta$  to have  $\beta$  equal 1). Although these best values vary slightly for each dataset, each set offers about the same small performance increase in comparison with the estimated values. For example, the F1@15 performance with the best data-derived values are 39.34, 41.18, and 74.31 for datasets 1, 2 and 3, respectively, as compared to 38.26, 39.59, and 72.89 F1@15 performance with manually estimated data.

Figure 4 shows how Content Wizard's performance changes with changing the parameters' values. Since the results are similar across the three datasets, we report only the results from dataset 2 (see Fig. 4). To generate these figures, we consecutively fixed one of the parameters to its best value and plotted the change of performance for each reasonable combination of the remaining parameters. The results show that when the value of  $\alpha$  increases (leading to the increased occurrence of past concepts in recommended content), the performance of the system tends to decrease. On the other hand, a larger absolute value of  $\gamma$  (leading to stricter penalty for future concepts) usually results in a better performance. However, no single parameter could lead to the best performance; it is the combination of the contribution of all the concepts in the three categories. The results hint that the instructors in the courses used for our analysis do not pay a lot of attention to the *past concepts* and tend to avoid *future concepts* when choosing content for students to learn at the current unit.

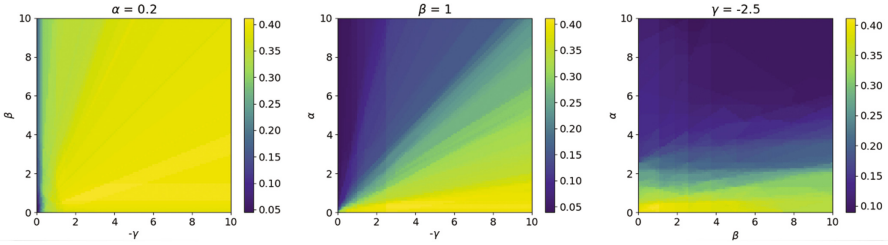


Fig. 4. F1@15 of Content Wizard with different sets of  $\alpha$ ,  $\beta$ , and  $\gamma$  in dataset 2.

### 6.2 How Many Code Examples Do We Need?

One of the most important elements in our recommendation process is the code examples provided by instructors. The examples are vital to understand what an instructor expects students to learn in a given unit. It could be expected that the more examples are provided for each unit, the better items the Wizard recommends. But how many of these code examples are sufficient to achieve good quality results? In order to assess the impact of the number of provided examples, we picked the topics that had at least 10 examples and compared the performance of both approaches using from 1 to 10 examples (randomly selected from all examples provided by the unit) as input. As shown in Fig. 5a, the performance of Content Wizard (measured by F1) consistently improves when the number of examples increases. In contrast, the baseline TF\*IDF approach (Fig. 5b) is not able to learn from an increasing number of examples. It could be also observed that with the first four to five examples the increase of quality reaches a *plateau*, which hints that four to five might be the optimal number of examples to ask instructors to provide.

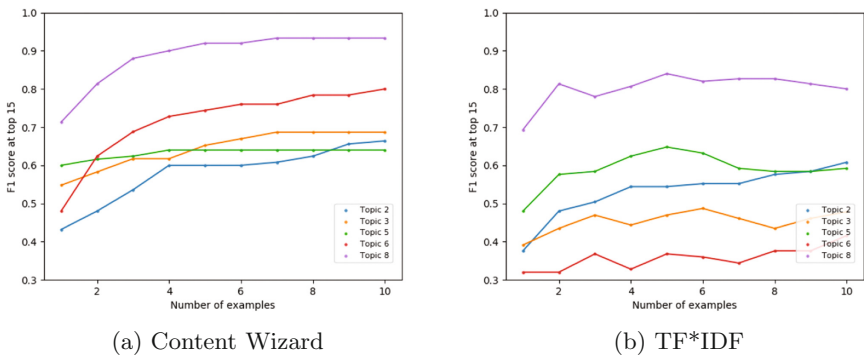


Fig. 5. F1@15 performances with different number of code examples for five topics in dataset 3 that have at least 10 code examples.

### 6.3 Discovered Limitations

A deeper analysis of recommendation performance helped us to reveal some limitations in our approach. First, the current approach doesn't take into account the fact that some concepts within a topic might be more important than others. For example, in the topic *while loop*, the concept *WhileStatement* is more important than concept *BreakExpression*. Exploring the importance of each concept and adding its weight to Eq. (1) may potentially help Content Wizard to achieve a better ranking. However, it is still unclear how the importance of a specific concept can be derived from data rather than by asking the instructor. We tried a natural idea to use TF\*IDF weights as importance weights, but this did not improve the overall performance of Content Wizard. Indeed, TF\*IDF weights more heavily most unique concepts, i.e., *BreakExpression*, while the key topic concept, i.e., *WhileStatement* might dominate topic problems.

Second, we discovered that the code examples provided by the instructors are frequently not *exhaustive* in listing all concepts that the instructor wants to teach in a unit. By observing the automatically deduced course structure that was produced in our study, we noticed that although in most of the cases the concepts from the code examples cover all the concepts from the content selected by the instructors, some concepts such as *PostIncrementExpression* (`+=`) do appear in the selected content, but not in any provided code examples (though these do contain the concept *PostDecrementExpression* (`-=`)).

Finally, the Wizard will fail to recommend items for a unit that doesn't introduce new concepts, but instead uses learned concepts to introduce a specific class of problems; for instance, *finding the maximum value in an array* or the topics *using truth values* and *instructors on code-writing and problem solving* in dataset 3. While a new *type of problems* could be considered as new knowledge [18], it is not recognized by our Java parser.

## 7 Discussion and Future Work

This paper presented a course-adaptive content recommender system called Content Wizard, which assists instructors in authoring adaptive online programming courses. We introduce a novel unobtrusive example-based approach to build a fine-grained course structure that encodes an instructor's vision of course organization. We also presented a novel content recommendation approach that uses the whole course structure to recommend the most relevant content for each course unit. Altogether, these innovations aim to decrease the effort required to build a high-quality course based on reusable learning content (i.e., efficiency and time to author [5, 6]) and facilitate the task of maintaining its coherent sequential structure. We believe that this approach could be most valuable for adaptive educational systems, such as ITSs or adaptive hypermedia, since personalization algorithms require a much larger volume and variety of learning content for each topic (to allow fine-grain personalization), rather than static courses.

We assessed the performance of the proposed approach using three datasets collected from different universities. Comparing our system’s performance with a standard baseline, we demonstrated that Content Wizard provides higher-quality recommendations, especially in more challenging and realistic contexts. The good recall performances suggest that Content Wizard could be efficiently used in a real-life context: with content ranked by Content Wizard, an instructor only needs to review the top 15 items out of several hundreds of items to select the ideal content for each course topic.

While our work indicates the strong potential of the suggested approach, we recognize that the approach itself and our evaluation process have several limitations, which we plan to address in future work. First, this study doesn’t consider that different concepts might have different importance within a topic. Second, it doesn’t account for the fact that the examples provided by instructors might not be exhaustive. While a group of related concepts is usually introduced in the same unit, only some of these concepts are usually illustrated in the code examples. In future work, we plan to extract the relationships between the programming concepts from the ontology introduced in Sect. 4.2, and assume that a group of closely related concepts is added as a whole to a unit once at least one concept is used in the examples. Third, the current Java parser is unable to recognize higher-level domain concepts, such as a specific *type of problem*. We intend to improve the Java parser in order to extract more complex knowledge of programming content.

On the evaluation side, while the data-centered (off-line) performance evaluation approach is a dominant way to evaluate recommender systems, we believe that only an online user study could provide a reliable assessment of the system as a tool to support course authors. In particular, a user study is essential to evaluate “beyond ranking” aspects of the Wizard, such as content warning signs. In future work, we plan to engage course instructors in the evaluation process.

The current implementation and evaluation of our recommendation approach has been performed in the area of learning programming where problem-solving examples in the form of code are both popular and well-structured (allowing concept extraction). The necessity to automatically process instructors’ worked examples limits the applicability of our approach in the suggested form, however, there is a considerable number of popular domains (math, physics, chemistry) where worked examples are well-structured (i.e., formulas, equations) and could be automatically processed for concept extractions. In the future, we plan to explore our approach in some of these domains.

**Acknowledgements.** We would like to thank Arto Hellas from University of Helsinki for providing dataset 3. We would like to thank Yun Huang, Roya Hosseini, and other members of the PAWS lab for their feedback on this paper.

## References

1. Moffatt, D.V., Moffatt, P.B.: Eighteen pascal texts: an objective comparison. *SIGCSE Bull.* **14**(2), 2–10 (1982)
2. 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, New York (2017)
3. Cafolla, R.: Project MERLOT: bringing peer review to web-based educational resources. *J. Technol. Teacher Educ.* **14**(2), 313–323 (2006)
4. Hislop, G., et al.: Sharing your instructional materials via ensemble. *J. Comput. Sci. Coll.* **26**(6), 160–162 (2011)
5. Murray, T.: An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: Murray, T., Blessing, S.B., Ainsworth, S. (eds.) *Authoring Tools for Advanced Technology Learning Environments: Toward Cost-Effective Adaptive, Interactive and Intelligent Educational Software*, pp. 491–544. Springer, Dordrecht (2003). [https://doi.org/10.1007/978-94-017-0819-7\\_17](https://doi.org/10.1007/978-94-017-0819-7_17)
6. Sottolare, R.A.: Challenges to enhancing authoring tools and methods for intelligent tutoring systems. In: Sottolare, R.A., Graesser, A.C., Hu, X., Brawner, K. (eds.) *Design Recommendations for Intelligent Tutoring Systems*, pp. 3–7. U.S. Army Research Laboratory, Orlando, FL (2015)
7. Manouselis, N., Drachler, H., Verbert, K., Duval, E. (eds.): *Recommender Systems for Learning*. Springer, Berlin (2013). <https://doi.org/10.1007/978-1-4614-4361-2>
8. Mitrovic, A., et al.: ASPIRE: an authoring system and deployment environment for constraint-based tutors. *Int. J. Artif. Intell. Educ.* **19**(2), 155–188 (2009)
9. Alevan, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: The cognitive tutor authoring tools (CTAT): preliminary evaluation of efficiency gains. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) *ITS 2006*. LNCS, vol. 4053, pp. 61–70. Springer, Heidelberg (2006). <https://doi.org/10.1007/11774303-7>
10. Brusilovsky, P., Eklund, J., Schwarz, E.: Web-based education for all: a tool for developing adaptive courseware. In: *Proceedings of Seventh International World Wide Web Conference, Brisbane, Australia, 14–18 April 1998*, pp. 291–300 (1998)
11. Chad Lane, H., Core, M.G., Hays, M.J., Auerbach, D., Rosenberg, M.: Situated pedagogical authoring: authoring intelligent tutors from a student’s perspective. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) *AIED 2015*. LNCS (LNAI), vol. 9112, pp. 195–204. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-19773-9\\_20](https://doi.org/10.1007/978-3-319-19773-9_20)
12. Cristea, A., Aroyo, L.: Adaptive authoring of adaptive educational hypermedia. In: De Bra, P., Brusilovsky, P., Conejo, R. (eds.) *AH 2002*. LNCS, vol. 2347, pp. 122–132. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-47952-X\\_14](https://doi.org/10.1007/3-540-47952-X_14)
13. Brusilovsky, P., Sosnovsky, S., Yudelson, M., Chavan, G.: Interactive authoring support for adaptive educational systems. In: *Proceedings of the 2005 Conference on AI in Education*, pp. 96–103. IOS Press, Amsterdam (2005)
14. Cabada, R.Z., Estrada, M.L.B., Garca, C.A.R.: EDUCA: a web 2.0 authoring tool for developing adaptive and intelligent tutoring systems using a Kohonen network. *Expert Syst. Appl.* **38**(8), 9522–9529 (2011)
15. Medio, C.D., Gasparetti, F., Limongelli, C., Sciarrone, F., Temperini, M.: Course-driven teacher modeling for learning objects recommendation in the moodle LMS. In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP 2017)*, pp. 141–145. ACM, New York (2017)

16. Brusilovsky, P., et al.: Increasing adoption of smart learning content for computer science education. In: Working Group Reports of the 2014 Conference on Innovation and Technology in Computer Science Education, Uppsala, Sweden, pp. 31–57. ACM (2014)
17. Hosseini, R., Brusilovsky, P.: JavaParser: a fine-grain concept indexing tool for java problems. In: The First Workshop on AI-supported Education for Computer Science, pp. 60–63. Springer, Heidelberg (2013)
18. 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.) ICFCFA 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)





# Assessing Leadership Competencies Through Social Network Analysis

Faisal Ghaffar<sup>1</sup>(✉), Neil Peirce<sup>2</sup>, and Alec Serlie<sup>3</sup>

<sup>1</sup> IBM Ireland, Dublin, Ireland  
faisalgh@ie.ibm.com

<sup>2</sup> School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland  
peircen@tcd.ie

<sup>3</sup> GITP-Research and Erasmus University, Rotterdam, The Netherlands  
a.serlie@gitp.nl

**Abstract.** Leadership competencies are regularly identified as some of the most in demand workplace competencies. However, the development of these competencies requires appropriate assessments that are often either highly subjective (e.g. manager appraisals) or prohibitively expensive (e.g. roleplays with trained actors). The increasing usage of workplace social networks and increasing prevalence of digital collaboration tools presents a continuous stream of social interactions that can contain evidence of leadership occurring in situ. In this paper we present initial research on the feasibility of Social Network Analysis in the workplace to assess leadership competencies. We examine the assessment in terms of content, construct, and criterion validity. We then present our hypotheses on how the assessment can be conducted including the algorithms necessary to extract relevant features from a social network graph model. Our initial research, to our surprise, shows a weak correlation between an individual's degree centrality and betweenness centrality and the leadership competency that is self-reported. However, experiments indicated a strong positive correlation between network structure based and social collaboration activities based features and the characteristics of the leadership competencies. Our initial machine learning experiments achieved an Area Under the Curve (AUC) score of 0.899 when social network and collaboration activity based features were leveraged to distinguish individuals with self-reported leadership competencies from others. Finally we discuss our findings on the practicality of the approach, and future work on validating and improving the results obtained using parallel conventional assessments for leadership competencies.

**Keywords:** Assessment · Leadership · Social network analysis  
Great Eight · Competencies

## 1 Introduction

The assessment of transversal competencies and skills, are predominantly realized through subjective self-rating questionnaires and interviews [20, 22, 25].

Although certain skills such as problem solving are readily assessed in narrow domains such as mathematics, generalizable assessments remain elusive [25]. Although behavioral interviews and situational judgment tests are also used [20], they are relatively uncommon as their cost and susceptibility to rehearsal, respectively, limit their appeal. The design of soft-skill assessments, in general, is challenging for a number of reasons. The multidimensional nature of these skills, their expression most readily occurring during execution, and the typical absence of a resulting artifact, are obstacles to summative or self-reported assessments [15]. Moreover, these skills typically vary between situations and contexts, where some fuzziness in the definition of the skills is also likely [10].

To realize an effective assessment, clarity is needed on what is being assessed and what instruments are being used. Specifically, there is a need to clarify competency models and social network analysis, respectively.

## 1.1 Competency Models

As work and jobs vary extensively there are numerous competencies and competency models available (e.g. O\*NET, ESCO), there will be little distinction and nuance between them, or as Bartram [2] states “Clearly a balance is needed between highly differentiated models that may not be generalizable and overly broad constructs that fail to capture relevant general dimensions of performance” [2]. To ensure re-usability it is necessary to implement a competency framework that is universal enough to apply to most jobs within a variety of organisations. In order to do so a trade-off is needed between specificity and generalisability.

Kurz and Bartram [19] conducted extensive work on developing a generic competency framework. Bartram [2] distinguished a framework of 112 very specific competencies at the finest level of detail, so called competency components. These competency components are clusters of workplace behaviour. Bartram [2] states that “. . . these components can be thought of as building blocks that can be aggregated together to produce competencies. Sets of competencies, in turn, form competency models. . .” [2]. The research aggregates the 112 components into eight general factors. Based on their work, Kurz, Bartram & Baron [18] propose a general competency model which distinguished eight general factors, also called the Great Eight.

The Great Eight structure provides an articulation of the work performance domain that is consistent with a wide range of models used by practitioners in competency practice and supported empirically by the way in which competency ratings cluster when subjected to factor analysis (e.g. [17,18]). The Great Eight factors [2] are:

- Leading and Deciding
- Supporting and Cooperating
- Interacting and Presenting
- Analysing and Interpreting
- Creating and Conceptualising

- Organising and Executing
- Adapting and Coping
- Enterprising and Performing

This work uses the Great Eight factors as the base competency model that is reusable across diverse jobs and organizations.

## 1.2 Social Network Analysis

Social Network Analysis (SNA) - a recent development in sociology, is a technique of analysing individual's social links arising from social interactions, their formation, evolution and impact. Recently SNA has gained popularity by its ability to solve complex problems in various domains of an individual's life due to the availability of large digital footprints of the social data. Social network data is modeled as a graph where individuals are *nodes* and any "connection of interest" between nodes are *links*. With this graph model of relational data, analysis is possible at various levels; at the node level for the analysis of node's position in the network, at the network level for global structural analysis, and at the dyad level for analysis of links and their properties [4]. Generally, in SNA, special consideration is paid to nodes and network characteristics. The following are basic metrics that can be used for node characteristics:

*Degree Centrality* (DC) is the number of links a node has with others. This indicates the connectedness of a node. In the case of directed graphs, degree centrality has two forms *in-degree centrality* and *out-degree centrality* which could be the indicators of a node being prominent or influential respectively. The basic idea is that many others seeking to connect to a node (in-degree) is a sign of importance of that node and a high out-degree could indicate the node's ability to reach many others and disperse information more quickly, which could be thought as exerting influence.

*Closeness Centrality* (CC) shows a node's reachability to others in the network. It is the measure of a node's *geodesic path* distance to all other nodes in the network. It is normalized (range=(0,1)) and gives an indication of a node's reachability to others. In the context of this paper, closeness centrality is a measure of a node's influence on others and is calculated using the Hubble approach [3] in which all pathways are considered between two actors and weight of each path is assigned using an attenuation factor of 0.5.

*Betweenness Centrality* (BC) shows the node's role as connector in the shortest paths between any two nodes in the network. It determines the relative importance of the node in terms of other's dependency on the node.

*Eigenvector Centrality* (EC) is also a measure of a node's influence. It is an extension to degree centrality but unlike degree centrality it considers inequality in connections of the node's neighbours.

*Structural Holes as Social Capital:* A node is considered to span a structural hole in a social network if it is linked to parts of the network that are otherwise not well connected [7]. It is argued that nodes that span a large number of structural holes have the advantage of access to unique information through friends

who do not know each other [6]. In the workplace, applications that foster collaboration among employees such as blogging applications can be considered as spanning structural holes, in that, they facilitate interactions between employees who would otherwise not be able to find each other. The concept of being in an advantageous position, from the information access point of view, when a node bridges a structural hole in the network is exploited in this paper. We associate the structural hole benefits of having access to unique information with the structural effectiveness of node’s ego network. The ego’s network effectiveness in terms of unique information means that node is less constrained by its direct contacts, have more open network (i.e. node’s alters are not connected with each other). These aspects can be quantified with three measures as proposed by [6].

1. *Constraint (CT)* - It is the extent to which a node is constrained by its connections. The lower the better; a low constraint value of a node indicates less dependency on its connections. It is defined by [7] as:

$$CT_{ij} = p_{ij} - \sum_q p_{iq}m_{jq}, q \neq i, j \tag{1}$$

where  $m_{jq}$  is  $i$ 's interactions with  $q$  divided by  $j$ 's strongest relationship with anyone and  $p_{iq}$  is the proportion of  $i$ 's energy invested in the relationship with  $q$  which is constant as  $\frac{1}{N}$ , where  $N$  is number of nodes in the network.

2. *Effective Size (ES)* - Measure of non-redundancy in the connections of a node. The higher the better; less redundant connections increases an employees chance of accessing more unique information. It is defined by [7] as:

$$ES_i = \sum_j \left[ 1 - \sum_j p_{iq}m_{jq} \right], q \neq i, j \tag{2}$$

3. *Ego Network Efficiency (EFF)* - It is the measure of quantifying the effectiveness of connections on the node. The higher the better; a more efficient network reduces redundant information.

## 2 SNA as an Assessment Method

The use of SNA to assess transversal competencies is promising as it is formative, continuous, based on on-task evidence, and flexible across situations and contexts. However, this approach faces inherent challenges:

- **Sampling:** Captured social network evidence is a proxy for actual social interaction, it is unlikely to capture the full scope of social interactions.
- **Ambiguous Intent:** It may not be clear as to why a social interaction happened, its purpose, and its positive or negative nature may be unreported or ambiguous.
- **Availability:** Evidence may not be available due to circumstance (small teams, availability of digital social tools), organizational policies (e.g. security), and legal restrictions (e.g. data protection). The challenge of incomplete data has been noted by [14].

- **Universality:** Where a social network is built by the necessity of a job role, its usefulness as a predictor of individual traits may be limited. In the case of personality assessment, it has been found that SNA can be a poor predictor for managers as their social network is required and not elective [7]
- **Coverage:** Valid assessments incorporate observations of the skills being assessed. The focus of SNA on network interactions and structure inherently limit it to those skills evidenced through social interactions.

There is a need for reliability, validity, objectivity, and feasibility in assessments [15]. Further requirements include the need for assessments to be "clear and consistent; technically sound, using valid and reliable observations, data and inferences" [10].

The use of social network analysis to assess transversal competencies addresses several of these requirements, as well as the challenges faced by existing approaches. Its objectivity is highly desirable, especially in light of the prevalence of unreliable manager observations [22]. Its reliability over time is likely to be good, although calibration and verification of the algorithms will be needed. However, feasibility is a considerable challenge for this approach. The ability to obtain sufficient social network data for any individual in light of the data actually existing, it being representative of the individual, and legal and ethical requirements have been considered, is a significant obstacle. Moreover, the benefit incurred from the assessment results would need to exceed the cost of gathering and processing this data.

The remaining challenge is that of validity. Although intuitively the approach possesses high validity due to focusing on actual on-task behavior, the assessment needs validity in terms of content, construct, and criterion [15].

**Content Validity.** In terms of content validity, the nature of the social interactions would need to align with the competency or skill under assessment. In the case of assessing collaboration, the SNA evidence would need to represent collaborative activities such as co-authoring or meetings, as opposed to non-collaborative activities such as broadcast marketing or periodic email updates.

In particular, the Great Eight competency factors of Leading & Deciding and Supporting & Cooperating will be the focus of this work.

**Construct Validity.** In terms of construct validity, there is supporting research to justify identifying individual traits through SNA, particularly in terms of personality [1, 7, 11, 13, 16, 24, 26]. However, the direct assessment of workplace transversal competencies through SNA is so far, an under explored area.

For the characteristics of individuals, the research is less prevalent. However, the use of closeness is related to positive performance on learning tasks, a result in contrast to other non-learning task research [8]. This work also showed that communication styles were associated with different ego network compositions. Additionally, the use of network centrality was found to correlate with leadership reputation for known leaders. However, the effect was context dependent and

was evident in networks of subordinates, but not in networks of high-ranking supervisors [21].

Overall the use of SNA to assess specific competencies is evidently limited with little consensus on the metrics in use or the competencies assessable [14]. However, there is growing interest in focussing on attributes of individuals to provide insight in the broader organisational performance. Despite the limited research in this area, there is evidence to support the likelihood that with due consideration of context, SNA for competency assessment has construct validity.

**Criterion Validity.** To evaluate the criterion validity of the approach we will investigate the concurrent and predictive validity of how SNA assessment agrees or predicts the results of other assessments such as game-based assessments [23], 360 degree reviews [5], and self-reported personality tests [12]. Within the DEVELOP research project [9], these parallel assessments are available for the competencies focussed on.

The novelty of the approach lies in validating SNA as an assessment through comparison not only with other objective measures (game-based assessment), but also through indirect assessments (personality tests), and commonly practised 360 degree reviews. These other assessments consider actual, perceived, and predicted behaviour respectively. This approach aims to establish to what extent and accuracy our approach predicts these three criteria, and in doing so, ensure a broad validity of the approach.

## 2.1 SNA Features of Interest

The novel nature of using SNA for individual competency assessment leaves limited insight into the network characteristics resulting from specific competencies. For this reason, this research is largely explorative in nature. However, the following specific hypotheses were investigated.

**Direct Assessment of Competencies Through SNA:** The use of network observations captured digitally, as they happen, creates a chronological dataset, an attribute missing from previous work that used network surveys or untimed friend networks. This availability of timed interactions may correlate with competencies related to Deciding & Initiating Action. In particular, it may correlate with the Great Eight competencies of *Acting on Own Initiative*, *Acting with Confidence*, and *Taking Action*.

**H1:** The initiator of first social interactions, and bursts of social interaction is a predictor of Deciding & Initiating Action competencies. The competency of Providing Direction & Coordinating Action requires a high level of communication and social interaction.

**H2:** The network centrality and relative frequency of interactions is a predictor of Leading & Supervising competencies. The expression of competencies in the work place relating to leading, deciding, supporting, and cooperating, creates

observable social interactions that affect the social network structure. These structural changes may be distinct enough to predict these competencies.

**H3:** Network structure and Social Capital characteristics predict Leading & Deciding and Supporting & Cooperating competencies.

### 3 Social Network Feature Analysis/Extraction

#### 3.1 Data Sources Description

We conducted our experiments in a large multinational organization using the following datasets. An irreversible MD5-hashing technique was used to anonymize and align the records across all the datasets.

1. *Enterprise social network (ESN)* - This is a “friendship” social graph of employees in the enterprise and it has more than 100k<sup>1</sup> nodes and millions of edges between them. The edges are directed where one terminal node of a particular edge indicates the node who initiated the “friend request”.
2. *Collaboration activity streams* - This dataset contains a large number of activities performed by employees on an *Enterprise collaboration platform - IBM Connections*, for a period of over 2.5 years (Jan 2014- June 2016). Collaboration activities include “creating/commenting/linking” a blog, “joining/following” a community, “tagging/following” others, creating or commenting or linking someone’s status update.
3. *Self-reported leaders* - In addition to the social network and collaboration data, we also had a number of employees (approx. 2200) who completed a “Leadership Development Programme” (LDP) between June 2016 and Aug 2016 within the organization. They had their profiles tagged with the “leadership” competency. A tag in IBM Connections is an descriptive label associated with profiles to identify employee skills, interests, or areas of expertise. Employees can tag their own profiles as well as others. In this paper, we extract the profiles who have self-reported or been endorsed by others with the “leadership” tag. It is important to note that LDP is mandatory for employees at senior management or executive roles in the organization but not all the employees who complete LDP have their profiles tagged. Therefore, for our ground truth, we only consider employees who have their profiles tagged with “leadership” skill (self-reported as well as endorsed by others) and are at the management roles in the organization.

#### 3.2 Experiment Settings

In this paper, we approach validation of our proposed hypotheses in two steps: First we extract three different sets of features for each user in our datasets where

---

<sup>1</sup> We are not able to reveal actual numbers here and throughout the paper for commercial reasons.

each set characterise the social behaviour expected from a user having the leadership competency as discussed in the previous section. To identify employees who show leadership characteristics, we use social structural features to statistically compare them with users who don't show leadership characteristics.

In the second step we further demonstrate the construct validity of the SNA technique for leadership competency assessment. This is done by leveraging the extracted features in a machine learning binary classification task. The objective of supervised learning in our case is to see if an employee's social behaviour can be used to characterize an employee from the leadership competency point of view. To quantify the model evaluation, we adopted two measures in this paper, Area Under ROC Curve (AUC) and Precision due to the fact that our dataset was extremely imbalanced in terms of ratio of positive labels with negative (approx. 1:100).

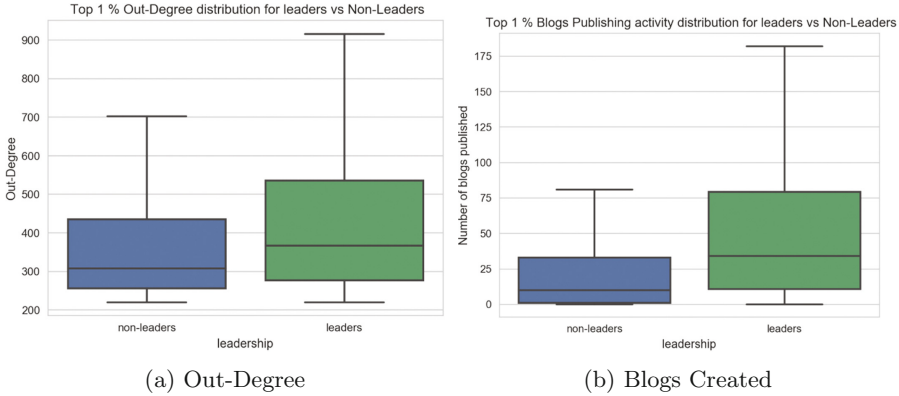
## 4 Results and Discussion

### 4.1 Results of Explorative Analysis

*H1: The initiator of first social interactions, and bursts of social interaction is a predictor of Deciding & Initiating Action competencies.*

To get deeper insight into the characteristics of employees with leadership competencies, in a first step, we looked at how connected these users are. To do so, we compared their network degrees with average degree of the social network. The average degree ( $\langle k \rangle = 331$ ) of 2,573 leaders was 8.5 times higher than the average degree for the rest of the network ( $\langle k \rangle = 39$ ). However, when leaders were compared to the average degree of their neighbours, only 31% of users with the leadership tag had higher degree than their immediately connected nodes. Based on this observation, we derived two rankings: first, we ranked users with respect to their number of out-degree (a high value of out-degree is indicative of a user being a "source" of friendship requests), second, we ranked the number of created blog posts. We compared the top 1% segment of both user rankings. We observed the average number of social activities, which involve *taking action on own initiative* (e.g. creating a blog post, sending a friend request), was higher for leaders than for non-leaders. We observed a combined 60% of top 1% ranked users in out-degree were also the top 1% ranked employees who initiated first the actions of publishing content (e.g. publishing a blog post) on the social platform were also leaders and their averages (out-degree and blog creation) were higher compared to non-leaders (as shown in Fig. 1). This is also confirmed by our comparison analysis of leaders vs non-leaders in Table 1. We can see the top 1% ranked users in features that characterise a user as *taking action on own initiative*, contain the highest overlap with leaders.





**Fig. 1.** Comparison of initiating actions like SNA features for leaders vs non leaders

**Table 1.** Overlap between leaders/non-leaders and the users with high number of action in collaboration streams and high out-degree

leaders vs non-leaders	Out Degree			#blog posts created			#comments on blogs			# blogs liked			total visits on created blog		
	Top 1%	Top 5%	rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest
leaders (%)	17.17	8.71	0.07	11.28	6.21	0.36	6.99	3.63	0.75	1.96	1.56	0.65	12.43	6.33	0.37
non-leaders (%)	82.83	91.29	99.93	88.72	93.79	99.64	93.01	96.37	99.25	98.04	98.44	99.35	87.57	93.67	99.63

*H2: The network centrality and relative frequency of interactions is a predictor of Leading & Supervising competencies*

For H2, we mined collaboration activity streams to identify and analyse user actions that can be considered as evidence of their competency to supervise others. To this end, we extracted two features for all users in our dataset; (i) “number of employees mentored”, (ii) “a yes/no flag based on user’s history if he/she has mentored others or not”. In addition, to evidence of mentoring and sharing experience, we analyzed employee’s centrality in the ESN. We applied in-degree centrality, closeness centrality, betweenness centrality, local clustering co-efficient and eigenvector centrality to employee’s social network and ranked them for each centrality measure to be classified with respect to the categories top 1%, top 5%, and “rest”. Table 2 shows the percentage overlap for leaders and non-leaders in their respective categories. Results of this analysis reveal that 7.51% of top 1% ranked employees who mentored others are leaders and this percentage gets smaller for the other two categories. Our analysis shows that leaders not only exhibit the behaviour of supervising others but are also central in their respective social networks. In particular, the results for closeness centrality reveal that leaders are generally close to all others in the network: 19.24% of the top 1% leaders users belong to the category of the top 1% users with respect to closeness centrality. However, unlike the global measures of centralities, results indicate that leaders have very low overlap in Local Clustering Co-efficient (LCC) and the overlap slightly increase with low ranked users (top 5% and “rest” category). This might be understandable when the context of LCC is considered. LCC is

**Table 2.** Overlap between leaders/non-leaders with users of high centrality and higher evidence of supervising others

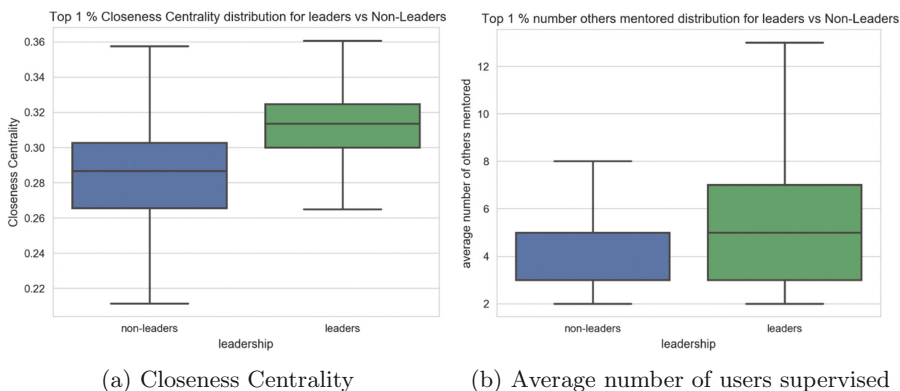
leaders vs non-leaders	# employees mentored			Closeness Centrality (CC)			Betweenness Centrality (BC)			Eigenvector Centrality (EC)			Clustering Coefficient (CLCO)		
	Top 1%	Top 5%	rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest
leaders (%)	7.51	2.26	0.84	19.24	8.06	0.05	3.37	2.47	0.33	7.95	7.5	0.05	0.04	0.06	0.07
non-leaders (%)	92.49	97.74	99.16	80.76	91.94	99.95	99.63	97.53	99.67	92.05	92.5	99.95	99.96	99.94	99.93

the degree to which nodes tend to bind together in triangles locally. It can also be thought as a measure “openness” or “closeness” of a node’s social network. Lower values of LCC means a node’s network is more open and the node has more opportunities to connect with others in the network. This network characteristic might be desired for a leader’s social network as network openness allows leader’s to connects with diverse individuals who themselves are not connected, and ultimately giving access to unique information.

Altogether, our analyses of the ESN for this H2 hypothesis shows that many users with leadership characteristics are among the best-connected users in the ESN. This holds for all centrality measures taken into account.

*H3: Network structure and Social Capital characteristics predict Leading & Deciding and Supporting & Cooperating competencies*

Social capital measures are indicative of a user’s ability to utilise resources in the network. In this paper, we leverage measures proposed by Burt [6] to quantify the social capital in the form of user’s local constraint (CT), effective network size (ES) and efficiency (non-redundant connections) in a user’s ego network. Generally, an individual with high social capital is expected to have high ES and efficiency but lower CT value in his/her ego network [4, 6]. Our analysis reveal that there is high positive correlation between social capital measures and individuals with leadership characteristics. Our experiments demonstrate that the top 1% ranked users in the CT category have no leaders and overlap



**Fig. 2.** Comparison of supervising and network centrality features associated with H2 for leaders vs non leaders

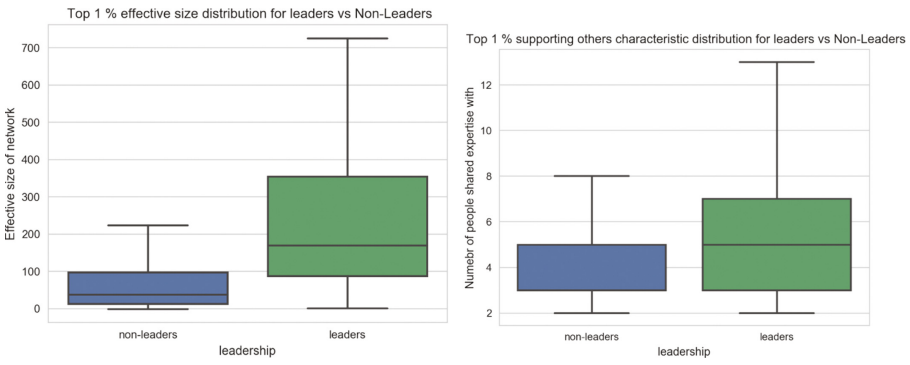
**Table 3.** Overlap between leaders/non-leaders with users of social capital measures and high evidence of supporting others

leaders vs non-leaders	#employees shared expertise with			Ego Network Constraint (CT)			Efficiency (EFF)			Effective Size (ES)		
	Top 1%	Top 5%	rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest	Top 1%	Top 5%	Rest
leaders (%)	10.17	4.35	0.69	0.00	0.07	6.25	10.51	6.00	0.07	18.02	8.86	0.05
non-leaders (%)	89.83	95.65	99.31	100	99.93	93.75	89.49	94.0	99.93	81.98	91.14	99.95

increases for 5% and “rest” category, whereas percentage of leaders in top 5% is high for ES and efficiency measures.

The definition of CT is *the extent to which a node is constrained by its connections to reach others in the network*. In light of this, it is expected for leaders to have a lower CT value and a more open and less redundant social network, as per our discussion for H2. This is the kind of the behaviour proposed by our H3 hypothesis and is shown by individuals with leadership characteristics in our results. Since in our proposed H2 and H3, there is overlap in characterising leading & supporting characteristics of the leaders from their network structure perspective, these results in a way confirms both hypotheses.

Summing up our analysis on network structure and social capital to characterize leadership competency, the top 1% ranked users in the social network have high overlap in exhibiting supportive behaviour towards others (i.e. they have high number of others with whom they have shared their expertise). Additionally, they have high structural social capital as shown by high overlap for measures such as effective size and efficiency along with low overlap in local constraint. For *supporting and cooperating*, an additional feature, number of employees with which leaders and non-leader have shared their expertise with was also used for this hypothesis (Table 3).

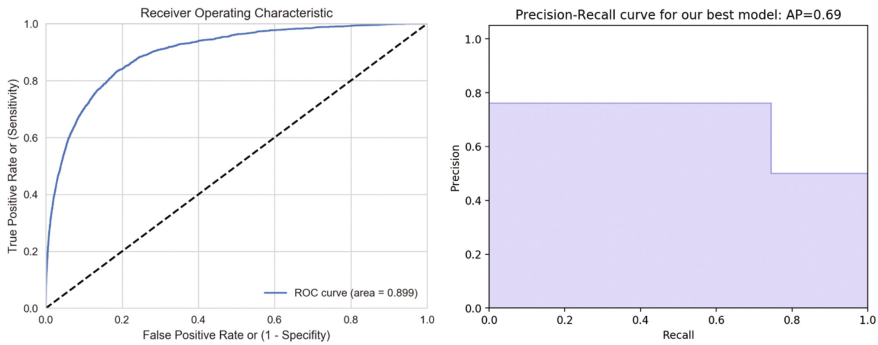


(a) Effective Size (Social Capital) (b) Average number of users supervised

**Fig. 3.** Comparison of supporting others and one of social capital features associated with H3 for leaders vs non-leaders

### 4.2 Results from Classification Analysis

To further demonstrate the ability of our extracted features as potential indicators of a user’s leadership competency, we trained a set of linear and non-linear supervised machine learning models to categorize users with “leadership” tags as “leaders” compared to users with no “leadership” tag as “non-leaders” in our dataset. We divided our dataset into training and test sets with a test size of 30%. To evaluate performance of the algorithms, a  $k$ -fold cross validation approach was adopted, which randomly divides all links into  $k$  subsets and models are trained with  $k - 1$  subsets while tested with the one remaining set. This process is repeated  $k$  times, with each of subset used exactly once as the testing set. In this paper, we used 10-fold cross validation and the trained models are Logistic Regression (LR), Linear Discriminant Analysis (LDA), Nearest Neighbours (KNN), Decision Tree (CART), eXtrem Gradient Boosting (xgb), Naïve Bayes (NB), and Multi-layer Perceptrons (NN). Input to our classification models are a set of network centrality and social capital based features inferred from employees’ social activity graphs. The extracted features are considered as the indicators of behaviours as described in our hypotheses and the associated Great Eight competency cluster. The prediction target was a variable with a binary value of 1 or 0 representing whether an employee had the “leadership” tag or not associated with their social profile. The xgb model performed best across both performance measures compared to all trained models. Therefore, we selected xgb as the final model for our dataset in this particular case and tuned it further to get the best score in terms of AUC and precision. The best score achieved with xgb classification model is AUC of 0.899 and average precision of 0.69 under the parameters of *learning rate* equal to 0.1, *max depth* equal to 5, and the *number of estimators* equal to 1000.



(a) Area Under ROC curve and score (b) Precision-Recall Curve with Average-Precision (AP) score

**Fig. 4.** AUROC curve with AUC score and precision-call curve of best performing model (xgb)

An analysis of features (network and social capital measures) input to our classifier revealed interesting insights. One of the social capital measures constraint (CT) and two of the network centrality measures Closeness Centrality (CC) and Eigenvector Centrality (EC), were the top three most important features in classifying instances of data (i.e. employees) into “leadership” versus “non-leadership” classes. The correlation between features showed a very low dependency on each other except simple network degree based features Degree Centrality (DC) and Effective Size (ES). Interestingly, by training the model with only the top three features the impact on performance in terms of AUC is negligible (a degradation of 0.007) whereas the performance in terms of average precision improves from 0.69 to 0.71.

From a leadership competency perspective, the top three features identified in our classification analysis were also among the most important measures identified for leaders in our explorative analysis. For example, our exploration phase showed the top 1% of leaders had zero constraint (CT), high CC and EC which means leaders generally have high reachability to others, are connected with more significant others and their network is generally less constrained (i.e. less dependent on their connections).

## 5 Conclusions and Future Work

The use of SNA for competency assessment is both promising and under developed. In the absence of research attributing causal relationships in this area, the DEVELOP research project will be explorative in identifying such relations. We have demonstrated the usefulness of SNA for competency assessment through a set of features extracted from a real-world enterprise social and collaboration data. Our experiments have shown that users with leadership traits exhibit a set of behaviours that can be captured from their network structure and social capital characteristics to assess their leadership competency. However, there are certain limitations associated with direct assessment of leadership competencies through SNA in terms of its validity and we were only able to show construct validity of SNA for competency assessment.

There are two key limitations to the current study, the accuracy of the leadership tagging, and the identification of non-trained leaders. The accuracy of the self-reported tagging may underestimate leadership due to those who accidentally, or otherwise, failed to add the tag to their profiles. Additionally those who participated in the leadership training but did not complete it, may have significant competence, yet lack the binary leadership tag. The presence of leadership competencies in those not taking the training is a larger challenge that we aim to address in future work. Specifically, leaders will be identified through game-based assessment, 360° reviews, and personality assessments through actual, perceived, and predicted leadership behaviours. This approach will aim to consider the spectrum of leadership competency, as opposed to just highly trained leaders.

The application of this approach to non-workplace settings such as formal education, is an open area of research. Although the approach taken should be

replicable in other contexts, there are likely to be significant barriers to obtaining appropriate data. Notably, there are ethical, and data protection challenges with accessing data of minors. Moreover, there would be challenges in establishing content validity as the relevance of captured social interactions may not be clear, and criterion validity may be limited in the absence of parallel assessments.

Other validity measures of SNA as highlighted at the start of this paper are not demonstrated here and we envisage to further investigate them in our future work. As part of future work the outlined hypotheses will be further tested as part of the DEVELOP project evaluations by gathering leadership competency assessment data through other direct or indirect methods such as personality assessment, 360 degree feedback, GMA assessment and serious-game simulation. This research will further contribute to the state-of-the-art in social network analysis for transversal competency assessment.

**Acknowledgement.** This work has received funding from the European Union's Horizon 2020 research and innovation programme through the DEVELOP project, under grant agreement No 688127.

## References

1. Amichai-Hamburger, Y., Vinitzky, G.: Social network use and personality. *Comput. Hum. Behav.* **26**(6), 1289–1295 (2010)
2. Bartram, D.: The Great Eight competencies: a criterion-centric approach to validation. *J. Appl. Psychol.* **90**(6), 1185–1203 (2005)
3. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. *Soc. Netw.* **28**(4), 466–484 (2006)
4. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. *Science* **323**(5916), 892–895 (2009)
5. Brett, J.F., Atwater, L.E.: 360 degrees feedback: accuracy, reactions, and perceptions of usefulness. *J. Appl. Psychol.* **86**(5), 930–942 (2001)
6. Burt, R.S.: The network structure of social capital. *Res. Organ. Behav.* **22**, 345–423 (2000)
7. Burt, R.S., Jannotta, J.E., Mahoney, J.T.: Personality correlates of structural holes. *Soc. Netw.* **20**(1), 63–87 (1998)
8. Cho, H., Gay, G., Davidson, B., Ingraffea, A.: Social networks, communication styles, and learning performance in a CSCL community. *Comput. Educ.* **49**(2), 309–329 (2007)
9. DEVELOP Consortium: DEVELOP project. <http://www.develop-project.eu>
10. Gibb, S.: Soft skills assessment: theory development and the research agenda. *Int. J. Lifelong Educ.* **33**(4), 455–471 (2014)
11. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA 2011*, p. 253 (2011)
12. Goldberg, L.R.: A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models (1999)
13. Gosling, S.D., Augustine, A.A., Vazire, S., Holtzman, N., Gaddis, S.: Manifestations of personality in online social networks: self-reported Facebook-related behaviors and observable profile information. *Cyberpsychol. Behav. Soc. Netw.* **14**(9), 483–488 (2011)

14. Hoppe, B., Reinelt, C.: Social network analysis and the evaluation of leadership networks. *Leadersh. Q.* **21**(4), 600–619 (2010)
15. Kechagias, K.: Teaching and Assessing Soft Skills. No. September, 1st Second Chance School of Thessaloniki, Thessaloniki, Greece (2011)
16. Klumper, D.H., Rosen, P.A., Mossholder, K.W.: Social networking websites, personality ratings, and the organizational context: more than meets the eye? *J. Appl. Soc. Psychol.* **42**(5), 1143–1172 (2012)
17. Kurz, R.: Automated prediction of managerial competencies from personality and ability variables. In: BPS Test User Conference, pp. 96–101. British Psychological Society, Leicester, UK (1999)
18. Kurz, R., Bartram, D., Baron, H.: Assessing potential and performance at work: the Great Eight competencies. In: British Psychological Society Occupational Conference, pp. 91–95. British Psychological Society, Leicester, UK (2004)
19. Kurz, R., Bartram, D.: Competency and individual performance: modelling the world of work. In: *Organizational Effectiveness: The Role of Psychology*, pp. 227–255 (2002)
20. Kyllonen, P.C.: Soft skills for the workplace. *Change: Mag. High. Learn.* **45**(6), 16–23 (2013)
21. Mehra, A., Dixon, A.L., Brass, D.J., Robertson, B.: The social network ties of group leaders: implications for group performance and leader reputation. *Organ. Sci.* **17**(1), 64–79 (2006)
22. Muzio, E., Fisher, D.J., Thomas, E.R., Peters, V.: Soft skills quantification for project manager competencies. *Proj. Manag. J.* **38**(2), 30–38 (2007)
23. van Nimwegen, C., van Oostendorp, H., Serlie, A., Modderman, J.: Assessing the personality trait compliance in a game context. In: Spink, A., Grieco, F., Krips, O., Loijens, L., Noldus, L., Zimmerman, P. (eds.) *Measuring Behavior 2012*, Utrecht, The Netherlands, pp. 231–234 (2012)
24. Staiano, J., Pianesi, F., Lepri, B., Sebe, N., Aharony, N., Pentland, A.: Friends don't lie - inferring personality traits from social network structure. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, p. 321 (2012)
25. Stasz, C.: Assessing skills for work: two perspectives. *Oxford Econ. Pap.* **53**(3), 385–405 (2001)
26. van de Ven, N., Bogaert, A., Serlie, A., Brandt, M.J., Denissen, J.J.: Personality perception based on LinkedIn profiles. *J. Manag. Psychol.* **32**(6), 418–429 (2017)



# Concept Focus: Semantic Meta-Data for Describing MOOC Content

Sepideh Mesbah<sup>(✉)</sup>, Guanliang Chen, Manuel Valle Torre, Alessandro Bozzon,  
Christoph Lofi, and Geert-Jan Houben

Delft University Of Technology, Delft, Netherlands  
{s.mesbah,guanliang.chen,m.valletorre,a.bozzon,  
c.lofi,g.j.p.m.houben}@tudelft.nl

**Abstract.** MOOCs promised to herald a new age of open education. However, efficient access to MOOC content is still hard, thus unnecessarily complicating many use cases like efficient re-use of material, or tailored access for life-long learning scenarios. One of the reasons for this lack of accessibility is the shortage of meaningful semantic meta-data describing MOOC content and the resulting learning experience. In this paper, we explore *Concept Focus*, a new type of meta-data for describing a perceptual facet of modern video-based MOOCs, capturing how focused a learning resource is topic-wise, which is often an indicator of clarity and understandability. We provide the theoretical foundations of *Concept Focus* and outline a methodical workflow of how to automatically compute it for MOOC lectures. Furthermore, we show that the learners' consumption behavior is correlated with a MOOC lecture's *Concept Focus*, thus underlining that this type of meta-data is indeed relevant for user-centric querying, personalizing or even designing the MOOC experience. For showing this, we performed an extensive study with real-life MOOCs and 12,849 learners over the duration of three months.

## 1 Introduction

Reusing and sharing teaching material is considered a central societal challenge by several policy makers. Despite continuously advancing open education policies [25], the vision of easy and personalizable access to open educational resources has still not been realized. To a large extent, this can be attributed to the lack of semantic capabilities of current courseware platforms: with access to only shallow *system-centric* meta-data (e.g. video length, authors names, publication date), these platforms are mostly degraded to be simplistic repositories for storing and serving learning resources. As a result, such platforms are often lacking in usability [29], and rarely take advantage of emerging technologies as for example intelligent digital assistants or conversational interfaces [27]. In this paper, we advocate for the availability of semantic meta-data for educational resources. In contrast to *system-centric* meta-data, *semantic* meta-data – e.g. didactic intent, perceived difficulty, required expertise, or educational quality – describes the



expected learning experience that a MOOC student might have with a given learning resource. This type of meta-data is generally hard to obtain as it either relies on subjective user-feedback, or needs to be indirectly approximated from the actual learning content. While some standards implicitly, introduce such meta-data types (e.g. LOM [6] – Learning Object Meta-data – covers “semantic density” or “difficulty”), it is usually not specified how such meta-data is defined, nor how it can be obtained from learning resources.

The main goal of this paper is to introduce the notion of **Concept Focus**, a measure of semantic relatedness of all concepts expressed in a learning resource. We set up a large-scale study on 3 MOOCs that engaged more than 12K learners over the duration of three months. We show that *Concept Focus*, while describing an intrinsic property of the learning resource, is also closely related to learner behavior patterns that are usually associated with difficulty or obstacles in the learning process. This can allow future work to use *Concept Focus* as a lever for learning personalization, e.g. steering certain types of learners towards content with high or low focus based on their personalities and learning goals. In summary, our original contributions include:

- The theoretical foundations for *Concept Focus*, a novel meta-data type capturing a relevant aspect of the learning experience of a MOOC video.
- The design space for methods that automatically obtain *Concept Focus* scores of a given MOOC video in a unsupervised fashion.
- The analysis of 3 real-life MOOC courses featuring 67 videos and 12,849 enrolled learners. We show that *Concept Focus* is a characterizing property of video scripts, describing their topical depth or width. We also report the presence of a significant correlation between *Concept Focus* scores and behavioral patterns indicating learning difficulties, e.g. video watching behaviour, quiz scores, and number of forum questions.

## 2 Concept Focus: Foundation and Implementation

Educational resources have been described by a multitude of different meta-data types, e.g. the IEEE LOM standard [7] includes a variety of different meta-data types, which can roughly be categorized into 9 groups. Most of these groups describe a learning object from a *system-centric* point of view: for example, general meta-data (e.g., id, title, language), technical aspects (e.g. length or size of videos), life-cycle (e.g., name of authors, version numbers), copyright, and usage restrictions. Only few types of meta-data actually cover the content itself: for instance, LOM group “classification” describes topic and keywords. Only one group of meta-data in LOM (“educational”) is dedicated to *learners* and their actual *learning experience*, with information about interactivity, difficulty or semantic density. This is analogous to other educational meta-data standards, as for example Ariadne [9]. Additionally, also bottom up approaches employing folksonomy techniques have emerged [4], with *educational* meta-data related to topical depth and didactic purpose being of central importance there.

This *educational* meta-data has been shown to be very beneficial for personalization and querying (especially data on difficulty, interactivity and density [22]), and its effectiveness even increases when combined with content-related meta-data [1]. Despite this fact, educational meta-data is rarely used in real-life MOOC systems. This can be attributed to the fact that it is expensive to obtain, and usually either expert judgments or crowd-sourcing needs to be employed to this end [22]. Furthermore, in [8], it has been shown that for effective personalization, more semantically deeper types (like learning styles or content properties) are beneficial, as they would allow for more meaningful similarity measurements between learning resources [8] for recommendations and explorative queries. Also Concept Focus could be used to that end, allowing to distinguish broader lectures from topically narrower ones.

## 2.1 Intuition

We define *Concept Focus* as a measure of semantic relatedness of all concepts expressed in a learning resource (e.g. a recorded lecture, or a script). Intuitively, *Concept Focus* characterizes how strongly a learning resource focuses on a specific topic: Concept Focus is *high* when the concepts of a resource share topical affinity – e.g., a lecture on natural language processing, which discusses a technique like “word embeddings” is implemented, mentioning only related NLP techniques and mathematical concepts.

We will test in our evaluation the hypothesis that learning material covering different topics, possibly loosely related, lead to learning difficulties. Even in cases where low *Concept Focus* does not always lead to confusion and learning problems (as it might also characterize material giving summaries or overviews), we argue that it is in either case a valuable meta-data field to be considered by an educational personalizing information system, as we will show, it drives behaviours similar to the ones of meta-data that are harder to obtain, as for example clarity or difficulty. *Concept Focus* can be computed automatically by relying on a combination of NLP and information extraction techniques, thus overcoming the aforementioned limitation of prohibitively high costs of crowd-sourcing or expert feedback. In short, *Concept Focus* can be realized as follows:

1. Extract all concepts (i.e., filtered named entities) from the textual representation of a given learning object.
2. Measure the *Semantic Relatedness* of a given concept in the learning resource, w.r.t. all other concepts in the same resource.
3. Calculate the *Concept Focus* of a resource, as a function of the semantic relatedness of all the concepts therein contained. Intuitively, if all concepts are semantically closely related, the *Concept Focus* focus of the resource is high; or low, otherwise.

## 2.2 Concept Extraction

In the following, we discuss how to extract concepts from videos, or more precisely the textual scripts of lecture videos. Arguably, the most important

educational material in MOOCs are the videos, as they are the principal mean for content delivery.

They are therefore our main object of analysis. Due to their interactivity, videos have the additional benefit of enabling in-video interaction analysis (i.e. users click actions such as pauses, replaying, etc.) to observe and assess the learning status of the students (e.g. difficulty in understanding the content) [15]. We exploit this fact in our evaluation.

Formally, a concept  $c$  can be defined as a k-gram that represents ideas and entities expressed in the video transcript text (e.g. “machine learning”, “stock price index”) [24]. Automatic concept extraction from text has received much attention in the past decade [3, 18–20, 26], and thus there exist a number of publicly available concept extractor tools, relying on techniques such as term-frequency analysis [26], co-occurrence graph [20], etc. Extracting concepts from MOOCs content is, however, a challenging task due to the low-frequency problem [23]: MOOCs videos are relatively short documents and due to the small number of words, statistical techniques (e.g. co-occurrence) are not applicable. To cater for such limitations, we employ an ensemble approach, running a battery of concept extractor tools on a video’s script, and extracting all the concepts contained in it. We adopt:

- **TF-IDF**<sup>1</sup>: A well-know Information Retrieval technique, used to rank candidate concepts based on their tf-idf (term frequency - inverse document frequency) in the corpus.
- **TextRank** [20]: A technique that extracts concepts by ranking them according to their co-occurrence graph.
- **TopicRank** [3]: An extension of Textrank. A graph-based concept extraction approach which relies on a topical representation of the text.
- **KPMiner** [10]: A simple technique, which employs a set of heuristic rules (e.g. length of the concept, position in the sentence) to extract concepts from the text.
- **Rake** [26]: Rapid Automatic Keyword Extraction is able to identify concepts by relying on the term frequency, term degree, and ratio of degree to frequency.
- **TextRazor**<sup>2</sup>: A text analysis API that returns detected entities, possibly decorated with links to the DBpedia or Freebase knowledge bases.

As a next step, we merge all the concepts individually extracted from each tool, filtering stopwords (e.g. something, anything, etc.) and concepts coming from “common” English language (e.g., “events”, “data”) that could be found in Wordnet. We retain only concepts that have been detected by the majority of the extractor tools (i.e. 4 out of 6) to filter out irrelevant concepts (e.g. “six months”, “new stories”). Intuitively, a concept will be considered as a correct concept if it has been harvested by different combinations of concept extraction tools [5]. By merging all concepts extracted from a given video scripts  $v$ , we obtain a final list of Candidate Concepts  $concepts(v) = \{c_1, \dots, c_N\}$ .

<sup>1</sup> <http://www.hlt.utdallas.edu/~saidul/code.html>.

<sup>2</sup> <https://www.textrazor.com/>.

### 2.3 Concept Focus

Concept Focus relies on measuring and aggregating the semantic relatedness of concepts contained in a lecture transcript: the higher the semantic relatedness between all concepts, the higher the focus of the lecture. While there can be many implementations for capturing semantic relatedness, previous studies [17] have shown that word embeddings [21] perform this task particularly well by e.g. measuring the cosine similarity of the word embedding vectors. We exploit Wikipedia to learn the word embedding representation of each concept. We first extract English articles from the latest publicly available Wikipedia dump<sup>3</sup>. Next, we built an embedding lexicon based on *fastText* [2]. *FastText* embeds each term (uni-gram and bi-gram) of a large document corpus into low-dimensional vector space (100 dimensions in our case) and overcomes the problem of out-of-vocabulary words by representing each word as a bag of character n-grams.

We adopt a typical measure of semantic relatedness  $SR(c_1, c_2)$ , that is computed between two specific concepts  $c_1$  and  $c_2$  by measuring the cosine similarity of their word embedding vectors [17].

In addition, we now also introduce the semantic relatedness  $SR(c, v)$  between a concept  $c$  and all other concepts contained in a video transcript  $v$ . We also value the relatedness to the title of a video. For instance in a video  $v$  about “Propensity score matching”<sup>4</sup>, concepts such as *propensity score*, *p-value* and *paired t-test* will get a higher semantic relatedness measure with respect to  $v$ , while a concept like *heart catheterization* is less related within  $v$ .

We define  $SR(c, v)$  for a concept  $c$  and a MOOC video transcript  $v$  as follows:

$$SR(c, v) = \frac{\sum_{c \in \text{concepts}(v)} SR(c, cv) * SR(c, \text{titleOf}(v))}{|\text{concepts}(v)|} \quad (1)$$

$SR$  is a value in  $[0, 1]$ , where 1 represents the maximum relatedness that a concept can have in a video.

Consequently, the Concept Focus of a given lecture video  $v$  can be defined as the average concept relatedness of each concept in  $v$  within the context of  $v$ , i.e.:

$$CF(v) = \frac{\sum_{c \in \text{concepts}(v)} SR(c, v)}{|\text{concepts}(v)|} \quad (2)$$

$CF$  is also in  $[0, 1]$ , where 1 is the highest *Concept Focus* value.

## 3 Evaluation

This section reports the results of an extensive study on real-life MOOCs, to showcase and discuss our new *Concept Focus* meta-data. We organize the study around the following research questions:

<sup>3</sup> <https://dumps.wikimedia.org/enwiki/20180201/>.

<sup>4</sup> <https://www.coursera.org/learn/crash-course-in-causality/lecture/VtFdu/propensity-score-matching-in-r>.

- **RQ1:** To what extent do properties of video scripts affect a course's *Concept Focus*? We investigate properties of learning material like video length, number of concepts, and position of the course in the MOOC.
- **RQ2:** To what extent does *Concept Focus* affect students' learning behaviour? We investigate the learners video watching behavior, quiz performance, and discussion behavior in relation to the *Concept Focus* of their consumed learning material.

### 3.1 Dataset Description

We analyze the log traces of learners collected from three MOOCs in edX<sup>5</sup>: itemize **DA** Data Analysis: Visualization and Dashboard Design, **IWC** Introduction to Water and Climate, **IWT** Introduction to Water Treatment.

We selected these 3 MOOCs for the following reasons: (1) they feature comparable amount of videos, and engaged students; (2) they cover a variety of topics; and (3) the scripts of their videos, and the interaction data for the engaged students are available. Table 1 summarizes the main properties of the selected MOOCs. We consider only *engaged* learners, i.e. learners that watched at least one video for more than 15 s. Interaction data is collected through click log traces. We analyzed in total 9,899,369 log trace records of 12,849 learners. Statistics of the MOOC and learners are summarized in Table 1.

**Table 1.** Overview of the three MOOC datasets analyzed. Legend: REG – Registered; Eng – Engaged; CR – Completion Rate

ID	Name	Start	End	Videos	# Learners		
					REG	ENG	CR
DA	<i>Data Analysis</i>	03/2016	06/2016	22	32,682	5,711	3.74%
IWC	<i>Introduction Water and Climate</i>	09/2014	11/2014	27	9,267	4,947	2.60%
IWT	<i>Introduction Water Treatment</i>	01/2016	03/2016	18	13,198	2,191	3.07%

**Properties of MOOCs.** To answer **RQ1**, we study the relation between the following features of videos in a MOOC, and their *Concept Focus*:

- **VD** – *Video Duration*: the length of a video, expressed in seconds.
- **VL** – *Average Video Length*: the average number of words in the video scripts of the given MOOC.
- **ANC** – *Average Number of Concepts*: the average number of concepts extracted from the video scripts of the given MOOC.
- **SC** – *Session of the Course*: the date the lecture was given (i.e. first session, second session, etc.)

**Learners Behaviour.** To address **RQ2**, we study the relationship between the measured behaviour of learners, and the *Concept Focus* score of videos. From

<sup>5</sup> <https://www.edx.org/>.

the log traces, we extracted the following 7 features. Each feature is calculated by aggregating all learner activities, including activities in the video player and in the course’s forum, and their proficiency with the subject as assessed by the MOOC’s grading system.

- **WT** – *Watching Time* of video material: the amount of time a learner has spent watching a video’s material in the MOOC.
- **NWT** – *Normalized Watching Time* of video material: the total amount of time a learner has spent watching video material in the MOOC divided by the duration of the video.
- **FS** – *# Forward Seek*: the total number of times a learner seeks forward while watching a video.
- **BS** – *# Backward Seek*: the total number of times a learner seeks backward while watching a video.
- **SU** – *# Speed Up*: the total number of times a learner increases the play speed while watching a video.
- **SD** – *# Speed Down*: the total number of times a learner decrease the play speed while watching a video.
- **FG** – *Final Grade*: the percentage of quiz questions the learner. answered correctly after having interaction with a video.
- **NFP** – *# New Forum Posts*: the number of new forum posts (i.e., questions) created by the learner after having interaction with a video. Here we consider posts created within 15 min from the last interaction with a video.

### 3.2 RQ1: Video Properties Vs. Concept Focus

Table 2 summarizes the properties of the video scripts part of our analysis, including the number of unique concepts extracted from the MOOCs, the average, median and standard deviation number of concepts extracted from their videos, as well as the length of the videos in terms of the number of words. Here we consider extracted concepts that were also present in Wikipedia, and for which a vector representation exists. Notably, 98% of the candidate concepts extracted from the concept extraction phase have a vector representation in our corpus. Figure 1 shows samples of extracted concepts organized in word clouds, where the size of the concept is proportional to their Semantic Relatedness (*SR*) score.

DA videos, compared to IWC and IWT, feature on average 60% less concepts, and half the number of words per video. The standard deviation is proportionally higher, thus showing more variability within the course. Figure 2 shows the distribution of the *Concept Focus* for all the videos of the three MOOCs. The average *Concept Focus* for the courses are respectively 0.29 for DA, 0.26 for IWT, and 0.19 for IWC. An example of IWC video with low focus score ( $CF = 0.16$ ) is the lecture “Urban Engineering”<sup>6</sup>, which includes a rather diverse concepts such as “cloaca maxima”, “city wall”, or “permeable pavements”. The lecture

<sup>6</sup> <https://www.youtube.com/watch?v=nhMcB-bwSF0>.

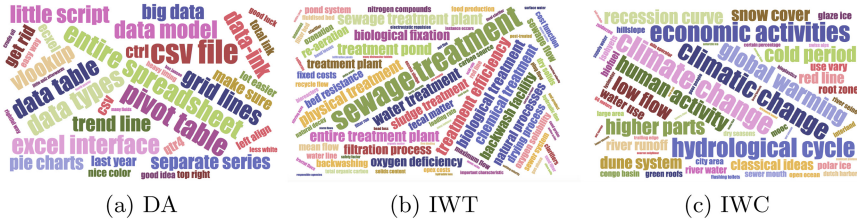


Fig. 1. Extracted concepts from video scripts of IWC, IWT and DA.

belongs to introductory course on Water Climate, a subject that is bound to embrace several topics. The “Solver” lecture in the DA course<sup>7</sup> is an example of very focused video ( $CF = 0.36$ ), including concepts such as “data table”, “excel sheet”, or “spreadsheet”. This is also expected, as the lecture is exclusively about an Excel plug-in program called “Solver”.

Figure 3 shows the relation between the length of the video (in terms of words) and the *Concept Focus* for each MOOC. Intuitively, one would argue that the longer the text of the video script, the higher the number of concepts contained in it, thus the lower *Concept Focus*. Indeed, this is not necessarily the case. We can find a moderate significant positive correlation only for videos in the IWC course (Fig. 3c:  $\rho = -0.59$ ,  $p - value : 0.0069$ ). However, as shown in Fig. 4, videos with higher number of concepts do have lower concepts focus, but only for the DA course a moderate significant negative correlation could be found (Fig. 4a:  $\rho = -0.60$ ,  $p - value : 0.01$ ). These results show that *Concept Focus* is a lecture-specific property that is not biased by the length of a video or by the sheer number of concepts contained in it. Arguably, this is a desirable properties for a content-centric meta-data.

Table 2. Descriptive statistics for concepts C and number of words W of the analyzed MOOCs video scripts. Legend: UC, Unique Concepts;  $\mu$ , average; m, median;  $\sigma$ , std.

MOOC ID	UC	$\mu C$	mC	$\sigma C$	$\mu W$	mW	$\sigma W$
DA	298	17	16	6	680	624	262
IWT	687	49	46	12	1268	1303	365
IWC	1095	43	43	7	1481	1398	366

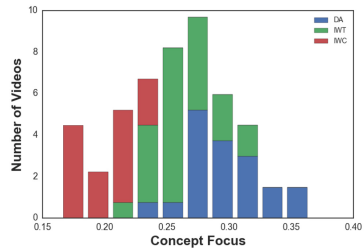
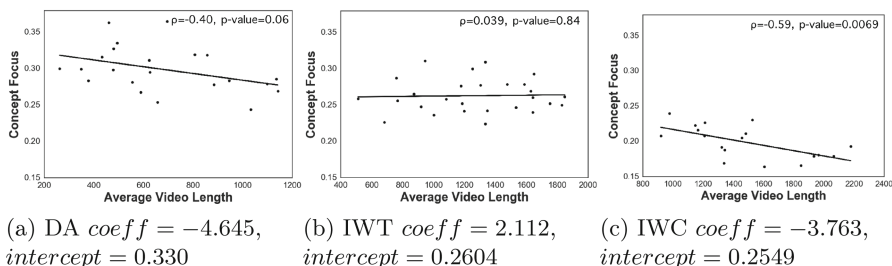


Fig. 2. Distribution of *Concept Focus* for the videos of IWC, IWT and DA in the shape of a stacked histogram

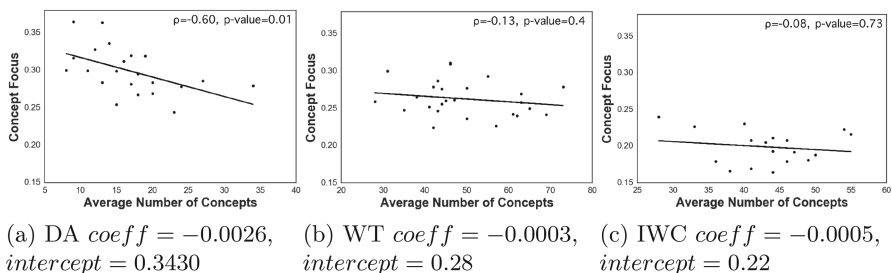
Finally, we study if the position of a video in a MOOC can be related to *Concept Focus*. Courses might feature different progression and organization of

<sup>7</sup> <https://www.youtube.com/watch?v=DgYmpmwBybQ>.

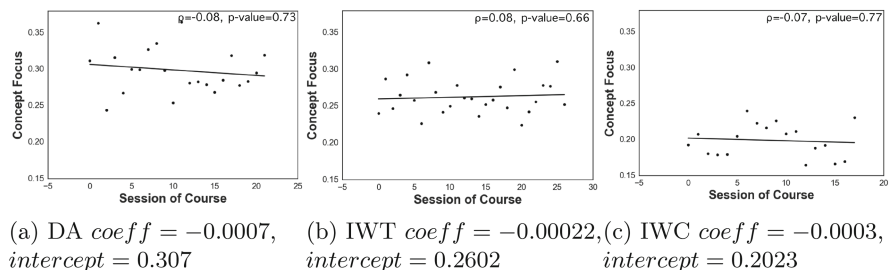
subject, with introductory lecture in the beginning (low *Concept Focus*) and specialized lectures later on (high *Concept Focus*). As shown in Fig. 5, the three courses feature very different teaching profiles. Despite the lack of statistically significant relation with *Concept Focus*, we can see how DA, for instance, starts with two very focused videos while, over time, lectures show consistent variations of *Concept Focus* scores. In IWT and IWC, on the other hand, the first lecture has low *Concept Focus*, and there is less variations in score across lectures, roughly remaining the same.



**Fig. 3.** *Concept Focus* and the number of words in the video transcripts



**Fig. 4.** *Concept Focus* and the average number of concepts in video transcripts



**Fig. 5.** *Concept Focus* and the position of the related video in the MOOC.



### 3.3 RQ2: Learning Behaviour Vs. Concept Focus

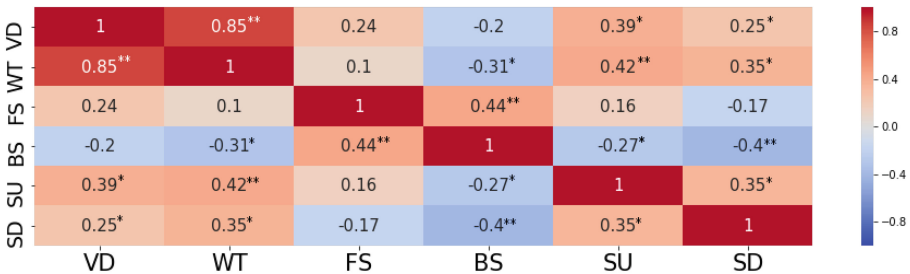
We first study how the length of a video is related to the behaviour of learners, Fig. 6 summarizes the Spearman correlation between all measures as a heatmap. The *Video Duration* VD is obviously highly correlated with the learners *Watching Time* WT. The longer learners spends time watching videos, the higher the amount of video interactions such as FS (*# Forward Seek*), SU (*# Speed Up*) and SD (*# Speed Down*). We believe that the high WT is not associated with learning difficulty, as we observe a negative correlation between WT and BS, and positive correlation with SD which are indicators of higher level of difficulty [16].

**Table 3.** Spearman correlation  $\rho$  between *Concept Focus* and learners behavioural features for all the videos in the dataset. \* $p$  - value  $< 0.05$ , \*\* $p$  - value  $< 0.001$

	$\rho$
NWT - # Normalized Watching Time	0.44**
FS - # Forward Seek	0.31*
BS - # Backward Seek	0.50**
SU - # Speed Up	-0.36**
SD - # Speed Down	-0.55**
FG - Final Grade	0.19
NFP - # New Forum Posts	-0.25*

Table 3 reports the measured Spearman correlation between the *learners behaviour* metrics and *Concept Focus* of the corresponding videos. *Concept Focus* is significantly correlated with NWT, BS, SU, SD, and NFP. We observe a moderate positive correlation between the amount of time learners spent watching video lectures and the number of times they seek backward - i.e., in the videos with higher *Concept Focus*, learners watch the video for a longer time and are more likely to re-watch parts

of them. This observation aligns with the previous study [28] where the authors showed that difficulty correlates negatively with dwelling time (i.e. time students spend watching a video). We interpret this result as a sign of students disengaging with videos having lower focus i.e. that cover a wider range of concepts. A similar result can be found in [12] where it has been shown that many



**Fig. 6.** Correlation heatmap of video interaction. Legend: VD - Video Duration; WT - Watching Time; FS - Forward Seeks; BS - Backward Seeks; SU - Speed Ups; SD - Speed Downs. \* $p$  - value  $< 0.05$ , \*\* $p$  - value  $< 0.001$

students stop engaging with a courses (e.g. watching the videos) when they haven't enough knowledge to understand the context.

We also observe a weak negative correlation with the number of new forum post - i.e., after watching videos with lower Concept Focus, learners are more likely to post in the forum. This can be an indicator of having difficulty understanding the concepts in video scripts with low focus. The number of times the learner speed up and down the video have also a significant moderate negative correlation with the Concept Focus - i.e., in the videos with higher Concept Focus, learners continue watching the video without changing the speed of the video, possibly a sign of well-designed content progression. Finally we do not observe any statistically significant correlations between the final grade of the students and the Concept Focus.

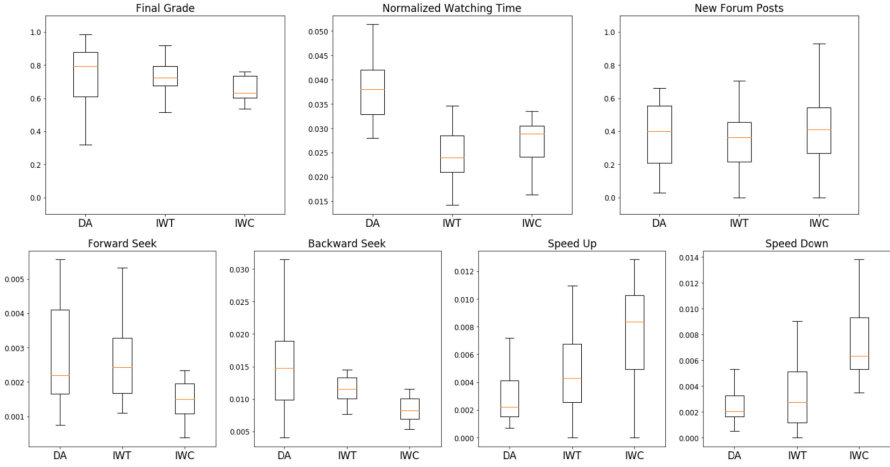
The box plots in Fig. 7 depict the break down of the distribution of final grade, normalized watching time, # of new forum post, # of forward seek, # backward seek, # speed up and # speed down of three courses. In order to check if the samples are drawn from different population groups we performed the Kruskal-Wallis H Test (KWHT). In DA, where average *Concept Focus* is higher (0.29) than IWT (0.26) and IWC (0.19), the learners achieve a slightly higher grade (KWHT *statistic* = 5.99, *pvalue* = 0.049); a statistically significant higher normalized watching time (KWHT *statistic* = 26.73, *p-value* =  $1.56e - 06$ ), forward seek (KWHT *statistic* = 10.49, *pvalue* = 0.005) and back ward seek (KWHT *statistic* = 17.31, *pvalue* = 0.0001); and slightly lower number of speed up (KWHT *statistic* = 9.94, *pvalue* = 0.006) and speed down (KWHT *statistic* =  $1.35e - 05$ , *pvalue* =  $22.42e - 05$ ). The difference in the distribution of number of new forum posts is not statistically significant (KWHT *statistic* = 5.16, *pvalue* = 0.07).

Altogether, these results show that *Concept Focus* is indeed a measure that relates to user-centric properties of videos, giving insights into potential engagement of learners, types of content, or potential learning problems.

## 4 Related Work

A growing body of literature has examined different attributes (e.g. video length [11], interface characteristics [13], video textual complexity [28], displaying the instructor's face to video instruction [14]) of MOOC videos and their effect on learners' dwelling time [15,16,28] or dropout [11].

Recently, several studies focused on the in-video interactions analysis (e.g. measuring the number of pauses, skipping, re-watching) to measure the level of the perceived video difficulty [15,16] and to model students learning behaviour [28]. The existing research capitalize on the relationship between the user and the content to measure the perceived video difficulty. We still have a limited understanding about the intrinsic properties of the text (i.e. without the interpretation of the users) that make a MOOC video clear for the students. Our work is inspired by [28], where the researchers focused on the textual analysis (e.g. word and sentence length, frequency of words, etc.) of the video scripts



**Fig. 7.** Distribution of Final Grade (FG), Normalized Watching Time (NWT), # New Forum Posts (NFP), # Forward Seek (FS), # Backward Seek (BS), # Speed Up (SU) and # Speed Down (SD) for the three courses.

and showed the effect of video complexity on the users video interaction (i.e. dwelling time and rate of the learners). However, the properties of the concepts (i.e. k-grams that represent ideas and entities expressed in the text such as: machine learning, stock price index, etc.) used in the text and the semantic relation between them are not well understood to characterize the lecture clarity and understandability. Thus, in this paper we focus on analyzing the content of MOOC videos to obtain their concept focus topic-wise, which is often an indicator of clarity and understandability of a lecture.

## 5 Conclusion

In this paper, we introduced *Concept Focus*, a novel type of meta-data capturing an aspect of a user’s learning experience when interacting with learning content in an online MOOC platform. *Concept Focus* describes how focused a learning resource is w.r.t. a restricted set of topics. It can be used to semantically characterize a learning resource (as for example an in-depth explanations vs. a general overview), but might also be an indicator for potential learning challenges. In contrast to other meta-data types, we show that *Concept Focus* can be computed fully automatically by relying on a combination of natural language processing and information extraction techniques, thus avoiding the common detriment of having to rely on costly crowd-sourcing or experts. We believe *Concept Focus* can play a role as part of the feature set of more elaborate methods for automatically deriving meta-data on teaching methods or learning styles.

We conducted an extensive study covering three real-life MOOCs with 67 videos on the edX MOOC platform. We show that *Concept Focus* is a property that does not depend on video length, it is lecture-specific, and it characterizes the organization of a MOOC. By analyzing the activity logs of 12,849 learners, we investigated their video watching behavior, quiz performance, and discussion behavior in relation to the concept focus of their consumed learning material. Furthermore, we investigated properties of learning material like video length or number of contained concepts. The analysis indicates a correlation between low *Concept Focus*, and behaviors which are associated with learning difficulties.

While these results are supported by general intuition and previous findings, our study is limited to three MOOCs. Additional studies are therefore necessary to better understand the relationship between this novel meta-data, and behavioural properties of learners.

## References

1. Abdelali, S., et al.: Education data mining: Mining MOOCs videos using metadata based approach. In: Information Science and Technology (CiSt), pp. 531–534. IEEE (2016)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017)
3. Bougouin, A., Boudin, F., Daille, B.: Topicrank: Graph-based topic ranking for keyphrase extraction. In: International Joint Conference on Natural Language Processing (IJCNLP), pp. 543–551 (2013)
4. Catarino, M.E., Baptista, A.A.: Relating folksonomies with dublin core. In: Dublin Core Conference, pp. 14–22 (2008)
5. Chen, L., Ortona, S., Orsi, G., Benedikt, M.: Aggregating semantic annotators. *Proc. VLDB Endow.* **6**(13), 1486–1497 (2013)
6. Learning Technology Standards Committee. IEEE Standard for learning object metadata. IEEE Standard, 1484(1), 2007-04 (2002)
7. Consortium, I.G.L.: Learning resource meta-data specification (2002). <https://www.imsglobal.org/metadata/index.html>. Accessed 26 Feb 2018
8. Dorça, F.A., Carvalho, V.C., Mendes, M.M., Araújo, R.D., Ferreira, H.N., Cattelan, R.G.: An approach for automatic and dynamic analysis of learning objects repositories through ontologies and data mining techniques for supporting personalized recommendation of content in adaptive and intelligent educational systems. In: Advanced Learning Technologies (ICALT), pp. 514–516. IEEE (2017)
9. Duval, E., Vervae, E., Verhoeven, B., Hendriks, K., Cardinaels, K., Oliivié, H., Forte, E., Haenni, F., Warkentyne, K., Forte, M.W., et al.: Managing digital educational resources with the ariadne metadata system. *J. Internet Cat.* **3**(2–3), 145–171 (2000)
10. El-Beltagy, S.R., Rafea, A.: KP-miner: A keyphrase extraction system for english and arabic documents. *Inf. Syst.* **34**(1), 132–144 (2009)
11. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: An empirical study of mooc videos. In: ACM Conference on Learning @ Scale Conference, L@S 2014. pp. 41–50. ACM, New York, NY, USA (2014)

12. Khalil, H., Ebner, M.: MOOCS completion rates and possible methods to improve retention—a literature review. In: EdMedia: World Conference on Educational Media and Technology, pp. 1305–1313. Association for the Advancement of Computing in Education (AACE) (2014)
13. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: Understanding in-video dropouts and interaction peaks in online lecture videos. In: Conference on Learning@ Scale Conference, pp. 31–40. ACM (2014)
14. Kizilcec, R.F., Papadopoulos, K., Sritanyaratana, L.: Showing face in video instruction: effects on information retention, visual attention, and affect. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 2095–2102. ACM (2014)
15. Li, N., Kidzinski, L., Jermann, P., Dillenbourg, P.: How do in-video interactions reflect perceived video difficulty? In: European MOOCs Stakeholder Summit, No. EPFL-CONF-207968, pp. 112–121. PAU Education (2015)
16. Li, N., Kidziński, L., Jermann, P., Dillenbourg, P.: MOOC video interaction patterns: what do they tell us? In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 197–210. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24258-3\\_15](https://doi.org/10.1007/978-3-319-24258-3_15)
17. Lofi, C.: Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. Database Soc. Japan **14**(3), 1–9 (2016)
18. Mesbah, S., Fragkeskos, K., Lofi, C., Bozzon, A., Houben, G.-J.: Semantic annotation of data processing pipelines in scientific publications. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017, Part I. LNCS, vol. 10249, pp. 321–336. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58068-5\\_20](https://doi.org/10.1007/978-3-319-58068-5_20)
19. Mesbah, S., Lofi, C., Valle Torre, M., Bozzon, A., Houben, G.J.: TSE-NER: an iterative approach for long-tail entity extraction in scientific publications. In: International Semantic Web Conference. Springer (2018)
20. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Conference on Empirical Methods in Natural Language Processing (2004)
21. 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)
22. Miranda, S., Ritrovato, P.: Automatic extraction of metadata from learning objects. In: Intelligent Networking and Collaborative Systems (INCoS), pp. 704–709. IEEE (2014)
23. Pan, L., Wang, X., Li, C., Li, J., Tang, J.: Course concept extraction in MOOCS via embedding-based graph propagation. In: International Joint Conference on Natural Language Processing, vol. 1, pp. 875–884 (2017)
24. Parameswaran, A., Garcia-Molina, H., Rajaraman, A.: Towards the web of concepts: extracting concepts from large datasets. VLDB Endow. **3**(1–2), 566–577 (2010)
25. Parliament, E.: Opening up education: Innovative teaching and learning for all through new technologies and open educational resources. Communication from the commission to the European Parliament (2013)
26. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Berry, M.W., Kogan, J. (eds.) Text Mining: Applications and Theory, pp. 1–20. Wiley, Hoboken (2010)
27. Sarikaya, R.: The technology behind personal digital assistants: an overview of the system architecture and key components. IEEE Signal Process. Mag. **34**(1), 67–81 (2017)

28. Van der Sluis, F., Ginn, J., Van der Zee, T.: Explaining student behavior at scale: the influence of video complexity on student dwelling time. In: ACM Conference on Learning @ Scale, L@S 2016, pp. 51–60. ACM, New York, NY, USA (2016)
29. Tsironis, A., Katsanos, C., Xenos, M.: Comparative usability evaluation of three popular MOOC platforms. In: 2016 IEEE Global Engineering Education Conference (EDUCON), pp. 608–612. IEEE (2016)



# Help Me Understand This Conversation: Methods of Identifying Implicit Links Between CSCL Contributions

Mihai Masala<sup>1,3</sup>(✉), Stefan Ruseti<sup>1,3</sup>, Gabriel Gutu-Robu<sup>1</sup>, Traian Rebedea<sup>1,3</sup>,  
Mihai Dascalu<sup>1,2</sup>, and Stefan Trausan-Matu<sup>1,2</sup>

<sup>1</sup> University Politehnica of Bucharest,  
313 Splaiul Independentei, 060042 Bucharest, Romania  
mihai.masala@gmail.com, {stefan.ruseti,gabriel.gutu,traian.rebedea,  
mihai.dascalu,stefan.trausan}@cs.pub.ro

<sup>2</sup> Academy of Romanian Scientists,  
54 Splaiul Independentei, 050094 Bucharest, Romania

<sup>3</sup> Autonomous Systems, 22 Tudor Vladimirescu, 050883 Bucharest, Romania

**Abstract.** Multi-participant chat conversations are one of the most frequently employed Computer Supported Collaborative Learning tools due to their ease of use. Moreover, chats enhance knowledge sharing, sustain creativity and aid in collaborative problem solving. Nevertheless, the manual analysis of multi-participant chats is a difficult task due to the mixture of different topics and the inter-twinning of multiple discussion threads during the same conversation. Several tools that employ Natural Language Processing techniques have been developed to automatically identify links between contributions in order to facilitate the tracking of topics and of discussion threads, as well as to highlight key contributions in terms of follow-up impact. This paper proposes a novel method for detecting implicit links based on features computed using string kernels and word embeddings, combined with neural networks. This method significantly outperforms previous results on the same dataset. Due to its smaller size, our model represents an alternative to more complex deep neural networks, especially when limited training data is available as is the case of CSCL chats in a specific domain.

**Keywords:** Computer Supported Collaborative Learning  
Implicit links identification · Natural Language Processing  
String kernels · Neural networks

## 1 Introduction

In an ubiquitous digital and online connected society, a significant part of communication between individuals has shifted to online messaging, either in social networking platforms or standalone chat applications. These technologies are no longer used just for entertainment or staying in touch with kin, as they are

employed for more and more complex activities. In education, online chats have been used for distant and lifelong learning, serious games, but also as a supplement to traditional learning activities. One of the most up-to-date use cases resides within Massive Open Online Course (MOOC) platforms in which chat and online discussion forums allow participants to communicate with the tutors and among themselves. From a broader perspective, chats have been often used for Computer Support Collaborative Learning (CSCL) tasks [1] as they support frequent changes of context and interest, thus potentially generating multiple discussion threads within the same conversation fostering collaboration and creativity [2]. However, exactly these multiple discussion threads make chat conversations difficult to understand and follow, especially as the number of participants increases. To solve this problem, environments were designed to support multi-participant collaborative chats by allowing users to manually annotate a set of referred contributions [3]. We call these annotations *explicit links* between chat utterances as they are added by participants when issuing an utterance.

Explicit links are useful for adding some structure to CSCL chats, but complex conversations with several parallel discussion threads are still hard to follow. Despite this explicit annotation facility, in practice participants do not annotate every utterance as this process is tedious and it interrupts the conversation flow. Thus, a mechanism for discovering unannotated links between utterances is useful to facilitate the understanding of CSCL chats. These are called *implicit links* and they are important to improve the readability of multi-participant chats. The labeling of implicit links can be done with the help of Natural Language Processing (NLP) techniques which support the automated analysis of texts [4].

Our main objectives have been two-fold. First, we highlight that detection of implicit links is a similar, albeit more complex task to sentence selection from question answering. Second, we present a supervised approach using string kernels and neural networks previously used for answer selection [5] that improves the performance for detecting implicit links when compared to previous studies employing different semantic models and semantic distances in the WordNet ontology [6,7]. In sentence selection for question answering, the most suitable answer is considered the sentence most (semantically) similar to the question. Similarly, given the current utterance and a list of previous contributions within a specific (time or distance) frame, an implicit link can be considered as the most similar utterance to the current one. In a first simplification, we eliminate the context of the conversation and only compute the similarity between the current utterance and each candidate, followed by the selection of the one with the highest score. Another important difference between question answering and implicit link detection is that the datasets for the former are an order of magnitude larger than the ones for the latter. This means that simpler supervised models can achieve better results than more complex, deep learning solutions.

The paper continues with a review of the linguistic techniques and features used for identifying implicit links. The following section contains a presentation of the proposed supervised method, with additional details about the corpus of conversations and the neural network model. Afterwards, results are presented



together with a comparison to previous studies in order to highlight the performance increase of the proposed supervised method. The last section concludes the paper and includes a discussion on the advantages of our approach.

## 2 Related Work

### 2.1 Implicit Links Detection

The process of manually annotating explicit links has two main limitations: (a) it is time consuming and it breaks down the conversation flow; (b) it may be subjective to the particularities of the user. Mechanisms for automated annotation of links have been designed to replace the manual labour performed by chat participants. As the links discovered by such algorithms and techniques are not explicitly added by users, the process is called implicit links detection [8] or chat disentanglement [9]. Multiple methods can be employed for solving this task.

**Semantic Models and Ontologies.** Previous experiments used semantic distances based on the WordNet ontology, together with several semantic models [6,7] to determine the optimal window (in terms of distance or time) to identify implicit references. The utterance belonging to a considered window that had the highest semantic similarity score with the referred utterance was chosen as its implicit reference. A corpus consisting of 55 chat conversations manually annotated with explicit links was used to evaluate the performance of this approach [6].

The experiments used three semantic distances computed using the WordNet [10] lexical ontology. In addition, three semantic models widely used in NLP tasks were also employed in the experiment. Latent Semantic Analysis (LSA) [11] builds a matrix of term-document occurrences which is decomposed using Singular Value Decomposition and then the dimensionality is reduced to a latent semantic space; the semantic relatedness between words is computed using cosine similarity in this space. Latent Dirichlet Allocation (LDA) [12] stores each word or text as a probability distribution over latent topics; the Jensen-Shannon dissimilarity is used to compute the relatedness between two units of texts (e.g. utterances). Word2vec [13] is based on neural word embeddings which are computed starting from word n-gram co-occurrences; the similarity between words is computed within the embedded space by means of cosine similarity.

**Neural Networks.** Neural networks have greatly contributed to recent advancements in various NLP tasks as they are able to automatically model complex combinations of simple inputs such as word embeddings. One relevant experiment for our study considered both meta information (e.g. time and distance between utterances, same/different author for the utterances, mentioning the author's name in the other utterance) and the content of the utterances. State-of-the-art results were obtained on chat disentanglement tasks by using Recurrent Neural Networks (RNNs) [14]. Our method is aimed at using a slightly

simpler neural network, with fewer parameters, that receives as input meta information, semantic features (e.g. word embeddings) and lexical features computed using string kernels.

## 2.2 Lexical and Semantic Models for Text Similarity

In this section we present two recent methods for computing semantic similarity between text documents, namely string kernels and neural models over word embeddings. Both are seen in the NLP community as powerful alternatives to ontologies and semantic models like LSA and LDA frequently used in the educational community for processing different types of texts, including CSCL chats.

**Word Embeddings.** Several alternatives for computing word embeddings on very large datasets have been proposed in recent years. Word embeddings are a method for representing words in a lower dimensional space based on their context of appearance in the corpus. While word2vec [15] is a generative neural model, GloVe embeddings [16] are computed using a count-based approach. However, both models are working on the word level as opposed to fastText [17] which is considered an extension of word2vec working on character n-grams.

**String Kernels.** String kernels [18] are kernel functions that work at the character level. Instead of projecting the documents into a high-dimensional space and performing computations in that space, string kernels employ a kernel function that simulates the dot-product of two elements in that high-dimensional space; the more similar the documents are, the higher the value of the kernel function. String kernels assume that any good measure of similarity between two documents is strongly related to the number of shared sub-strings of a given size in those documents. String kernels are obtained by varying the sizes of the n-grams (usually between 2 and 10 characters) and the function used for computing the n-gram overlap. The most common string kernels (i.e. intersection, presence and spectrum [19]) are based on the number of co-occurrences of shared n-grams. Spectrum kernel (see Eq. 1) is computed as the dot-product of shared n-grams frequencies. Instead of multiplying the frequencies, intersection kernel (see Eq. 2) uses the minimum of these frequencies. In contrast, the presence kernel (see Eq. 3) uses presence bits to encode if a n-gram is present or not in a string. For a fair comparison of strings of different sizes, normalized versions of these kernels are used in follow-up experiments.

$$k_p(s, t) = \sum_{v \in \Sigma_p} num_v(s) \cdot num_v(t) \quad (1)$$

$$k_p^\cap(s, t) = \sum_{v \in \Sigma_p} \min\{num_v(s), num_v(t)\} \quad (2)$$

$$k_p^{0/1}(s, t) = \sum_{v \in \Sigma_p} in_v(s) \cdot in_v(t) \quad (3)$$

where:

- $\sum_p$  = all  $p$ -grams of a given size  $p$
- $num_v(s)$  = number of occurrences of string ( $n$ -gram)  $v$  in document  $s$
- $in_v(s) = 1$  if string ( $n$ -gram)  $v$  occurs in document  $s$ , 0 otherwise

String kernels have also been used as a feature extraction method and combined with different classifiers to solve various problems such as native language identification [20], digit recognition and protein fold predictions [21]. Recently, Beck et al. [22] used Gaussian Process regression on string kernels to optimize the weights related to each  $n$ -gram size and the decay parameters for gaps and matches. Their model outperforms linear baselines for sentiment analysis, but lags behind a non-linear baseline, giving evidence that extending string kernels with non-linearities can provide better results. In a similar manner, Masala et al. [5] have used a neural network to assign weights to different  $n$ -gram sizes and also to non-linearly combine different kernels using a neural network. Their results show that a shallow neural network using string kernels and word embeddings can achieve very good results in question answering with a much smaller model than state-of-the-art deep models. We propose to use a similar approach for implicit links detection.

**Neural Models for Text Similarity.** Neural models for computing similarity between sentences have been widely used in question answering in recent years [23–25]. For the specific task of answer selection, a question and a pool of candidate answers are given and the model must discriminate the most likely answer from all other candidates. In general, neural networks computing the similarity between two sentences (or documents) generate inner representations for both text and then apply a similarity function on these representations. Usually, the representation is computed using a Bidirectional Long Short-Term Memory (Bi-LSTM) network [26] or a convolution neural network (CNN) [27].

Adding attention mechanisms to neural models proved to be a very efficient method in question answering, outperforming previous models. The intuition behind the attention mechanism is that, by looking at the question, different weights can be assigned to different parts of the candidate answer, thus allowing the model to focus on the relevant parts of the candidate. dos Santos et al. [24] combine the question and candidate representations obtained from the Bi-LSTM or CNN into a single, fixed-length matrix. Using this matrix, attention weights are extracted and used to modify both the question and the answer representations.

Instead of computing the attention weights by only looking at the inner representations of the question and the answer, Bachrach et al. [23] also use a global view of the question and of the answer, obtained using a multilayer perceptron (MLP) on a bag-of-words representation. In addition, Wang et al. [28] propose a general method for word-level sentence matching. After computing the attention weights, comparison functions (e.g. element-wise subtraction and multiplication, a simple MLP) are used for combining the representation of the

answer with the attention-weighted representation of the question, at word level. For the final classification, a CNN is used on top of this new representation.

## 3 Method

### 3.1 Corpus of CSCL Chat Conversations

The corpus used for this experiment consists of 55 chat conversations among undergraduate Computer Science students [6]. Students had to discuss about web technologies supporting collaborative work and how these can be efficiently used by a software company. While each participant had to be the supporter of a different technology, in the end they had to reach an agreement on the solution that best suited the company. To this aim, the discussions were similar to the problem-solving tasks usually encountered in other CSCL platforms - e.g., Stahl's Virtual Math Teams project [29]. Stahl demonstrated that problems which are difficult to be solved independently can be answered more effectively by groups of students involved in collaborative learning.

Two methods were considered for the matching process between the automatically detected implicit links and the manually annotated explicit links. The first one is the *perfect match* in which the two referenced utterances (explicit and predicted link) are identical. The second one is the *in-turn matching* - i.e., the implicit link belongs to a uninterrupted block of subsequent utterances written by the same participant, as the explicit link.

The conversations were performed using *ConcertChat* [3], which enables participants to explicitly refer one or more previous turns, when uttering their own contribution. These explicit annotations were used for computing the accuracy of the proposed method using both exact and in-turn matching. The corpus contains about 4500 explicit links and 17600 utterances, meaning that 29% of contributions have a corresponding explicit link. Table 1 shows fragments extracted from chat conversations depicting an exact match and an in-turn match, where the emphasized text marks the utterance which denotes the implicit link. The explicit link added by the participants within the conversation is presented in the Ref ID (reference ID) column.

Gutu et al. [6] have previously shown that a distance of 5 utterances covers 82% of explicit links in the dataset, a distance of 10 covers 95%, while a distance of 20 covered more than 98%. As for time, a 1 min timeframe covers 61%, whereas 93% of explicit links are covered by a 3 min timeframe, and more than 97% by 5 min window. For this reason, windows of 5 and 10 utterances, and 1, 2, and 3 min were used for the current experiments.

### 3.2 Network Model and Design

One of our key insights is that there is a strong resemblance between the way implicit links relate to their respective utterances and how an answer connects to a question. Therefore we propose a neural model inspired from the answer

**Table 1.** Fragments extracted from conversations showing exact and in-turn matching. (Implicit link is highlighted in bold)

Utt. ID	Ref. ID	Speaker	Content
<i>Exact matching</i>			
87		Tibi	<b>you can't rely on anyone ..there should be authorised people writing on this site</b>
88		Octavian	but if want to find organized information about that product wiki is the way to go
89	87	Oana	the people writing on the web site are authorized
<i>In-turn matching</i>			
193		Alin	and they embed only what you need
		...	(several utterances of the same participant, Alin) ...
196		Alin	<b>this is just an example of how hidden markov model can be used</b>
199	193	Razvan	you talked about the prior, does this mean that the method ignores the sequences that are after the word it's tagging and only takes into account the ones before it?

selection task. The goal of our model, inspired from the work of Masala et al. [5] is to find a combination of string kernels that can better capture the notion of implicit links between utterances. The previous most similar utterance to the current one is selected as the implicit link.

We combine three string kernels (spectrum, presence and intersection) with five n-gram ranges: 1–2, 3–4, 5–6, 7–8 and 9–10. We thus compute for each pair of sentences a feature vector  $v \in \mathcal{R}^{15}$ . A simple feed-forward multilayer perceptron (MLP) with one hidden layer is trained over these features. The MLP computes a similarity score for each utterance that is a candidate for an implicit link. The utterance that has the highest similarity score is selected as the discovered implicit link. For all experiments the hidden layer size is set to 8, using a batch size of 100 and Adam [30] optimizer for training. The objective function is the hinge loss defined in Eq. 4, similar to the one proposed by Hu and Lu [31] for finding similarities between two sentences, with the margin  $M$  set to 0.1.

$$e(u_r, u^+, u^-) = \max(0, M + \text{sim}(u_r, u^-) - \text{sim}(u_r, u^+)) \quad (4)$$

where:

- $u_r$  is the current utterance, for which the link is computed
- $u^+$  is the correct (explicitly) linked utterance
- $u^-$  is an incorrect utterance from the current window
- $\text{sim}(u_r, u)$  is the similarity score computed by the MLP between the representations of two utterances
- $M$  is the desired margin between positive and negative examples

In addition, we experiment with augmenting the features obtained using string kernels with semantic and conversation-specific features. Given two utterances, we compute the cosine similarity between the average vector computed

in the embedding space (using word2vec, FastText, and GloVe) for all words in each utterance. The information retrieved from the chat structure consists of differences expressed as counts of in-between utterances and time between the two considered utterances. Finally, for each candidate utterance (for a link) we compute two conversation specific features: if its author is the same as for the current utterance and whether the utterance contains a question.

## 4 Results

We evaluated the proposed methods on the previously described dataset. Our supervised neural model is compared with an unsupervised method based on string kernels and with state-of-the-art methods for implicit links detection and answer selection. For the unsupervised methods, the n-gram range (3–7) was selected to optimize the performance on a small evaluation set. For all supervised methods, a 10-fold cross-validation procedure was employed. Note that all results are reported on the test set. The word2vec [15] embeddings were pretrained on the Google News Dataset. The GloVe embeddings [16] were trained on a Wikipedia 2014 dump and Gigaword 5<sup>1</sup>. The FastText embeddings [17] were also trained on Wikipedia. For computing string kernels we employed an open-source library<sup>2</sup>.

As baselines, we have used both supervised and unsupervised methods employed for detecting implicit links and answer selection, namely:

- Path Length [6]: The best results for detecting implicit links on the same dataset were achieved using WordNet Path Length as similarity distance. Path Length computes the length of the shortest path between two concepts in the WordNet ontology.
- AP-BiLSTM [24]: The current utterance and the candidate utterance are both passed through a Bidirectional LSTM network. The outputs of both Bi-LSTMs, containing the hidden states at each time step, are afterwards combined into a single matrix. From this matrix, attention vectors are extracted via column-wise and row-wise max pooling, and new representations for the utterances are computed. For the classification step, cosine similarity is used on the new representation of the utterances. AP-BiLSTM is one of the top performing deep learning models for answer selection.

The accuracy obtained by the baseline methods are presented in Table 2. While the AP-BiLSTM model is capable of capturing complex semantic relations for the answer selection task [24], its accuracy is low for our problem, offering performance just on par with the unsupervised path length semantic distance for the exact match (and even worse for in-turn match). The poor performance can be explained by the small size of the training dataset relative to the high number of parameters required by the model.

<sup>1</sup> <https://catalog.ldc.upenn.edu/LDC2011T07>.

<sup>2</sup> <http://string-kernels.herokuapp.com/>.

**Table 2.** Proposed baselines for implicit links detection (Exact matching accuracy - top row & In-turn matching accuracy - bottom row).

Window (utterances)	5			10		
	1	2	3	1	2	3
Path Length [6]	32.44%	32.44%	-	31.88%	31.88%	-
	41.49%	41.49%	-	40.78%	40.78%	-
AP-BiLSTM [24]	32.95%	32.39%	33.97%	33.86%	28.89%	24.49%
	34.53%	35.89%	37.58%	35.10%	31.82%	28.32%
Intersection kernel	31.40%	33.87%	33.58%	31.71%	32.24%	29.47%
	34.59%	39.58%	40.01%	34.78%	37.66%	35.24%
Presence kernel	31.84%	33.97%	33.58%	31.80%	32.33%	29.67%
	34.94%	39.81%	40.01%	34.89%	37.71%	35.41%
Spectrum kernel	31.21%	33.45%	33.17%	31.39%	31.56%	28.75%
	34.34%	39.12%	39.49%	34.46%	36.72%	34.26%

Similarly, string kernels as an unsupervised method provide mixed results when compared to path length: improvements are small and only for a larger time window (e.g. 2-min time window) for exact match. Furthermore there is no significant difference between any of the three string kernels functions.

Table 3 introduces the results obtained using the proposed neural model, with and without additional chat features, but without any semantic information. The results highlight the fact that chat and conversation specific features, especially the time and in-between turns distances between utterances are very important for detecting implicit links. A similar conclusion was established in previous studies as the Path Length method from Table 2 also uses a weighting for the path length semantic score, given the distance between the two utterances [6]. Nevertheless, conversation specific features (same author and whether the candidate utterance contains a question) are also relevant features improving the results both on their own and additional to window-based ones.

Compared to previous results using path length, the proposed neural model achieves a substantial improvement from 32.44% (window/time frame: 5 utterances/1 min) to 47.85% (window/time frame: 10 utterances/3 min) accuracy for exact match. Two important results should be highlighted. First, the neural network model achieves improvement for all combinations of frames considered. Second, this model is able to improve the results even when the number of candidates is higher (e.g. larger window/time frames); this was not the case with any of the baselines presented in Table 2.

The results of the experiments using semantic information are presented in Table 4. For all models involving semantic information, experiments were conducted using several word embeddings: word2vec, FastText, and Glove (embedding sizes 100 and 300). While semantic information increased the performance of our model, the gain is not significant especially compared to the performance gain obtained by adding chat specific features. This shows that

**Table 3.** Proposed methods without semantic information (Exact matching accuracy & In-turn matching accuracy).

Window (utterances)	5			10		
	1	2	3	1	2	3
NN using sk	35.21%	35.55%	35.77%	35.55%	34.08%	30.24%
	36.90%	39.39%	39.95%	37.02%	37.47%	33.74%
NN using sk + window + time [32]	33.40%	40.40%	41.87%	33.74%	41.30%	42.66%
	35.21%	44.01%	45.25%	35.44%	45.25%	47.29%
NN using sk + question + author	37.02%	39.84%	39.50%	37.35%	38.26%	35.10%
	39.05%	44.46%	44.46%	39.16%	42.32%	39.16%
NN using sk + window + time + question + author	37.92%	45.48%	47.06%	38.14%	46.27%	<b>47.85%</b>
	39.39%	49.66%	51.80%	39.50%	50.79%	<b>52.93%</b>

Note: sk - string kernels; window - # of in-between utterances; time - elapsed time between contributions; question - whether the utterance contains a question; author - if the utterance shares the same author as the utterance containing the link.

**Table 4.** Proposed methods enhanced with semantic information (Exact matching accuracy & In-turn matching accuracy).

Window (utterances)	5			10		
	1	2	3	1	2	3
NN using sk + sem [32]	36.45%	36.90%	36.00%	36.68%	35.10%	31.26%
	38.14%	40.47%	40.29%	38.14%	38.26%	34.76%
NN using sk + sem + window + time [32]	34.98%	41.64%	44.24%	35.32%	42.21%	44.48%
	36.68%	45.03%	48.53%	36.90%	45.93%	49.32%
NN using sk + sem + question + author	38.14%	41.19%	40.85%	38.48%	39.95%	36.34%
	39.84%	45.03%	45.03%	40.06%	43.56%	39.72%
NN using sk + sem + window + time + question + author	37.02%	46.38%	48.08%	37.24%	47.29%	<b>49.09%</b>
	38.60%	50.00%	52.25%	38.71%	51.46%	<b>53.83%</b>

framing the implicit link detection problem as a purely answer selection task will yield a inherently limited model. The largest gains can be observed for a longer frame (e.g. window/time frame: 10 utterances/3 min, improvement from 47.85% to 49.09%) which means that semantic information becomes relevant for capturing more distant implicit links.

Turning to a qualitative interpretation of the results obtained by the neural model, Table 5 provides examples in which our model is correctly predicting the implicit link. In the top of the table, we present two examples of utterances that represent direct answers to previously asked questions. The proposed model can also detect when an author continues his idea in a new utterance (see lower part of Table 5).



**Table 5.** Example of correct implicit link prediction (**explicit link**/*predicted link*).

Utt. ID	Ref. ID	Speaker	Content
74	74	<i>Cristi</i>	<i>and if it is not a free chat, then it's not that easy</i>
75		Oana	well, there are tons of free chats on the web, I don't believe that this is a real problem
76		Luis	On a chat you can have multiple thread discussions creating confusion
77		Alex	but, I don't think that the chats purpose is to store the information..
78	78	<i>Cristi</i>	<i>right, again but haw do you advertise to use the same chat?</i>
79		Oana	it's the same advertisement as with the forums, or the blogs: on-line advertisement
176		<i>Oana</i>	<i>can't they (wikis) be made private: i mean have groups of users who have permission to post?</i>
177	176	Florin	But if one evil man wants to make joke, anothers 10 will repair the damage
178		Mihaela	in all my experience i did not encounter such i thing...
179		Oana	for example, in an university: only teachers can add content

**Table 6.** Example of wrong implicit link prediction (**explicit link**/*predicted link*).

Utt. ID	Ref. ID	Speaker	Content
91	91	<b>Ciprian</b>	<b>we shall our know-how to make the best use of our separate products by combining them in one integrated tool</b>
92		Cristi	I think that we can use all of them because it's clear that they have different qualities for example ...
93		<i>Ionut</i>	<i>Yeah,let's think of a joint venture for our companies to create a product tocombine them all, an all-in-one learning application for large groups.</i>
94		Ciprian	so, wiki is the perfect tool to create semantic networks as tools for knowledge representation
95		Rudi	I propouse to give chatting options for the registrated users
96		Cristi	Let all think how his product will integrate better and what value it adds to the joint product.
87	87	<b>Radu</b>	<b>Blogs are also based on PHP and MySQL, and are really easy to use and put in practice</b>
88		<i>Raluca</i>	<i>do you have another ideas about the implementation?</i>
89		Radu	See WordPress or Blogger...

However, as the best accuracy is 49.09%, in about half of the cases our model is unable to detect the correct link. This is due to, but not limited, to more complex utterance interaction that can mislead even human annotators (see upper part of Table 6). In other cases, utterances simply do not provide enough information (see the last utterance in the bottom of Table 6). These limitations may be overcome by extracting more complex features from each utterance. Nevertheless some limitations are also due to the way the problem was formulated (as an answer selection task).

## 5 Discussions and Conclusions

Chat conversation have been used in CSCL tasks especially for solving difficult problems in larger groups of students. These conversations foster multiple parallel discussions threads and competing discussion topics that make the conversations hard to follow. Automated NLP techniques come to help by interpreting chats and detecting links between utterances. This process aims at supporting or even replacing the time-consuming work of explicit annotation. For example, it would be great to have a tool that suggests an implicit link for each utterance in a conversation (either chat, but maybe even a discussion forum within a MOOC). As the accuracy is slightly below 50%, the participants would still need to correct the automatic suggestion in half of the situations. On the bright side, it means that half of the time the predicted link is correct and the conversation flow will not be interrupted to manually pick an explicit link.

This paper proposes answer selection techniques for implicit links detection in chats. We explored a supervised neural model using string kernels, as well as additional domain-specific and semantic features. While string kernels alone performed similarly to semantic similarity methods used in previous studies, the neural network learned how to combine efficiently lexical, semantic and chat related features, and significantly increased the accuracy for the detection of implicit links. The method was also compared with state-of-the-art deep-learning models for question answering and achieved better results, proving to be a viable solution for smaller datasets. To our knowledge, this is the first approach of its kind.

Performance was not improved by a large margin by adding semantic information. More experiments need to be conducted with other semantic similarity measures as features, considering that each model might capture different facets of the relations between sentences. Another improvement can be achieved by also considering the context of the conversation, and not only a pair of utterances. This highlights a limitation of our current assumption which oversimplifies the problem, albeit that implicit links can be modelled as a sentence selection task, ignoring the context of the conversation in which utterances occur. Models that use the whole conversation for link detection might be more suitable in this case, but require a larger dataset for training.

The described approach has multiple practical implications. First, it introduces the possibility to split the conversation and easily follow multiple conversation threads, a functionality of great benefits for modelling online conversations in education and beyond. Second, summarizing relevant contributions for each participant by taking into account inter-dependencies between contributions enables the generation of an overview of their involvement. This process also creates a strong basis for assessing the degree of collaboration between participants. Third, implicit links also model cohesive links among contributions; thus, the avoidance of a high inter-twining of multiple concurrent discussion threads and keeping a cohesive discourse makes the conversation easier to follow.

**Acknowledgements.** This research was partially supported by the FP7 2008-212578 LTFLL, EC H2020-644187 *Realising an Applied Gaming Eco-system* (RAGE), and POC-2015 P39-287 IAVPLN projects.

## References



1. Stahl, G.: *Group Cognition: Computer Support for Building Collaborative Knowledge*. MIT Press, Cambridge, MA (2006)
2. Trausan-Matu, S.: Computer support for creativity in small groups using chats. *Annal. Acad. Rom. Sci. Ser. Sci. Technol. Inf.* **3**(2), 8190 (2010)
3. Holmer, T., Kienle, A., Wessner, M.: Explicit referencing in learning chats: needs and acceptance. In: Nejdil, W., Tochtermann, K. (eds.) *EC-TEL 2006*. LNCS, vol. 4227, pp. 170–184. Springer, Heidelberg (2006). [https://doi.org/10.1007/11876663\\_15](https://doi.org/10.1007/11876663_15)
4. Manning, C.D., Schütze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA (1999)
5. Masala, M., Ruseti, S., Rebedea, T.: Sentence selection with neural networks using string kernels. *Procedia Comput. Sci.* **112**, 1774–1782 (2017)
6. Gutu, G., Dascalu, M., Rebedea, T., Trausan-Matu, S.: Time and semantic similarity what is the best alternative to capture implicit links in CSCL conversations? In: *12th International Conference on Computer-Supported Collaborative Learning (CSCL 2017)*, pp. 223–230. ISLS (2017)
7. Gutu, G., Dascalu, M., Ruseti, S., Rebedea, T., Trausan-Matu, S.: Unlocking the power of word2vec for identifying implicit links. In: *17th IEEE International Conference on Advanced Learning Technologies (ICALT 2017)*, p. 199200. IEEE (2017)
8. Trausan-Matu, S., Rebedea, T.: A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants. In: Gelbukh, A. (ed.) *CICLing 2010*. LNCS, vol. 6008, pp. 354–363. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-12116-6\\_29](https://doi.org/10.1007/978-3-642-12116-6_29)
9. Elsnier, M., Charniak, E.: Disentangling chat. *Comput. Linguist.* **36**(3), 389–409 (2010)
10. Miller, G.A.: Wordnet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
11. Landauer, T.K., Dumais, S.T.: A solution to plato’s problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* **104**(2), 211240 (1997)

12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**(4-5), 9931022 (2003)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: Workshop at ICLR (2013)
14. Mehri, S., Carenini, G.: Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). vol. 1, pp. 615–623 (2017)
15. 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)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). <http://www.aclweb.org/anthology/D14-1162>
17. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint [arXiv:1607.04606](https://arxiv.org/abs/1607.04606) (2016)
18. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. *J. Mach. Learn. Res.* **2**(Feb), 419–444 (2002)
19. Ionescu, R.T., Popescu, M., Cahill, A.: Can characters reveal your native language? a language-independent approach to native language identification. In: EMNLP, pp. 1363–1373 (2014)
20. Ionescu, R.T., Popescu, M., Cahill, A.: String kernels for native language identification: insights from behind the curtains. *Comput. Linguist.* **42**, 491–525 (2016)
21. Gönen, M., Alpaydm, E.: Multiple kernel learning algorithms. *J. Mach. Learn. Res.* **12**(Jul), 2211–2268 (2011)
22. Beck, D., Cohn, T.: Learning kernels over strings using Gaussian processes. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 67–73 (2017)
23. Bachrach, Y., Zukov-Gregoric, A., Coope, S., Tovell, E., Maksak, B., McMurtie, C.: An attention mechanism for answer selection using a combined global and local view. arXiv preprint [arXiv:1707.01378](https://arxiv.org/abs/1707.01378) (2017)
24. dos Santos, C.N., Tan, M., Xiang, B., Zhou, B.: Attentive pooling networks. *CoRR*, abs/1602.03609 (2016)
25. Tan, M., dos Santos, C.N., Xiang, B., Zhou, B.: Improved representation learning for question answer matching. In: ACL, vol. 1 (2016)
26. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of 2005 IEEE International Joint Conference on Neural Networks, IJCNN 2005, vol. 4, pp. 2047–2052. IEEE (2005)
27. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746–1751, August 2014
28. Wang, S., Jiang, J.: A compare-aggregate model for matching text sequences. arXiv preprint [arXiv:1611.01747](https://arxiv.org/abs/1611.01747) (2016)
29. Stahl, G.: *Studying Virtual Math Teams*. Springer, New York, NY (2009). <https://doi.org/10.1007/978-1-4419-0228-3>
30. Kingma, D.P., Adam, B.J.: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)

31. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: *Advances in neural information processing systems*, pp. 2042–2050 (2014)
32. Masala, M., Ruseti, S., Gutu-Robu, G., Rebedea, T., Dascalu, M., Trausan-Matu, S.: Identifying implicit links in CSCL chats using string kernels and neural networks. 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, Part II. LNCS (LNAI)*, vol. 10948, pp. 204–208. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_37](https://doi.org/10.1007/978-3-319-93846-2_37)



# The Effect of Personality and Course Attributes on Academic Performance in MOOCs

Mahdi Rahmani Hanzaki<sup>1</sup>  and Carrie Demmans Epp<sup>1,2</sup> 

<sup>1</sup> Department of Computing Science, University of Alberta, Edmonton, Canada  
{rahmanih, cdemmansepp}@ualberta.ca

<sup>2</sup> EdTeKLA Research Group, Edmonton, Canada

**Abstract.** Predicting academic performance has been a topic of research for years, with different factors having been used to predict student grades. One of those factors is personality, with little work having focused on the effect of personality on academic performance in Massive Open Online Courses (MOOCs). Contributing to our lack of understanding of how personality is linked to academic performance in MOOCs, studies that predict academic performance by combining personality with attributes of online course design to have yet to be reported. In this paper, we try to tackle this problem by using personality and level of collaboration (a course attribute) to predict academic performance. We chose level of collaboration as one of the course attributes in our research because social factors, such as the amount of student interaction, can impact learner attrition in MOOCs. We apply machine learning algorithms to two different feature sets. The first feature set only uses personality as a predictor and the second feature set uses personality and level of collaboration in a course as predictors of academic performance. A comparison of these predictive models revealed that adding level of collaboration can increase their performance significantly. These results provide further evidence of the importance of validating classroom-based research in online settings. Moreover, the results of this work can be useful in several ways. For example, we may be able to give better recommendations to users based on their personality and the attributes of courses. We may also be able to adapt course attributes to match the personality characteristics of each student.

**Keywords:** Personality · Course attributes · MOOC · Educational data mining  
Learning analytics · Grade prediction

## 1 Introduction

Attempts to show that a student's personality can be used to predict his or her academic performance have shown a relationship between these two factors [7, 9]. However, personality is only one potential predictor of how students behave and perform in online courses. Other factors that can affect learners' performance are the attributes of an online course's design [19], such as its length [18] or the facilitation method employed to support student discussion [20]. For instance, students with certain personality traits might feel more comfortable in courses with certain attributes and therefore do better in those courses. As explained in [17], different personalities prefer different ways of

learning. For example, introverts like to study in quiet environments while extroverts prefer situations that are more dynamic, like classroom discussion.

A relationship between academic success and students' personality traits was found in traditional, face-to-face learning environments [7, 9]. Since the learning environment influences student behaviors, this means that we do not know the extent to which this relationship holds in online learning environments such as Massive Open Online Courses (MOOC). Even though this newer form of online learning environment shares many characteristics with more established online learning settings, it differs in one fundamental way: the course size, diversity in the student body, and design may be biased towards the learning preferences of some students. Furthermore, there is a clear bias in the demographic backgrounds (e.g., sex and country of origin) of those who withdraw from MOOCs [21], and we do not yet have a strong understanding of how learner personality may be tied to student retention or performance in this setting. Consequently, there is a need to study the effect of learner personality and course attributes on the academic performance of MOOC learners (i.e., personality traits together with the style of the course may be a good predictor of academic success). We address these gaps by employing machine learning algorithms to predict academic performance using personality traits and level of collaboration in a course. We chose level of collaboration in a course as one of the course attributes in our research because previous work (e.g., [5]) has shown that social factors, such as the amount of student interaction, can impact learner attrition in MOOCs. To provide a basis for the current study, we will first discuss a theory of personality to support the later discussion of prior work related to predicting academic performance using personality, the relationship between personality and system usage, and the importance of social factors in MOOC attrition.

## 2 Personality, Behavior, and Predicting Academic Performance

### 2.1 The Big Five Personality Traits

The Big Five personality traits is a model detailing individuals' personality. It describes an individual's personality using five traits: (1) Neuroticism (tendency to be prone to psychological stress), (2) Extraversion (tendency to seek stimulation in the company of others), (3) Openness (appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience), (4) Agreeableness (tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others), and (5) Conscientiousness (tendency to be organized and dependable). These traits have been studied extensively in areas such as learning and job performance, and they have been related to academic achievement [9], academic motivation [7], and job performance [12].

It has been suggested that personality can affect how people behave in an online learning environment as well as the learning approaches they prefer [17]. For instance, while extroverts tend to prefer more dynamic environments like classroom discussions, introverts prefer to work alone or in small groups and in a quiet environment [17]. These tendencies and the above findings partially motivated our use of a measure of the Big Five personality traits as one of the predictors of academic performance in MOOCs.

To study this phenomenon, we must also understand (1) the relationship between personality and academic performance, (2) the relationship between personality and behavior in online settings, and (3) how students are known to behave in MOOCs. The following sections summarize our current understanding of these contributing factors.

## **2.2 Personality and Academic Performance in Traditional Learning Environments**

Most of the work exploring the relationship between personality and academic achievement has been done in traditional learning environments. In [9], they tried to find the relationship between a student's personality traits and his or her academic motivation and achievement. For this purpose, they performed a survey to measure students' Big Five personality traits, their academic motivation, and their Grade Point Average (GPA). This survey also collected socio-demographic information. In the end, they found that the Big Five personality traits (especially conscientiousness, openness, agreeableness, and neuroticism) are significant predictors of GPA. In another project [7], they tried to find the relationship between academic performance, personality and learning styles. They used the Big Five framework for personality traits and they used the four learning styles introduced by [13] which consist of (1) synthesis-analysis (processing information, forming categories, and organizing them into hierarchies), (2) elaborative processing (connecting and applying new ideas to existing knowledge and to the learner's personal experiences), (3) methodological study (what is traditionally emphasized in most academic environments, such as being careful and methodical while completing all assignments on time), and (4) fact retention (processing information so that the main ideas are memorized with the goal of doing well on tests rather than understanding the meaning of what is being learned). After doing regression analysis, correlation analysis, and mediation analysis, they found that personality traits (neuroticism, openness, agreeableness, and conscientiousness) can predict GPA, with the synthetic analysis and elaborative processing learning styles mediating the relationship between openness and GPA.

These findings demonstrate a relationship between personality and performance, but they rely on self-reported grades rather than students' actual performance, and they do not include course features, such as level of collaboration. Our study includes both, and it considers specific student behaviors within the online course environment. To better understand this potential relationship between personality and learner behavior in online course settings we first detail what we know about personality and behavior in everyday online settings.

## **2.3 Personality and Internet Usage: Implications for Online Learning**

The studies detailed above enable us to understand how the Big Five has been applied in a traditional learning setting to explain the effects that personality has on learning performance. Considering that online settings are different from the traditional classroom setting, the above findings may not hold for online learning as different personalities might have different behaviors when they are in online settings than when they are



in a physical location with their classmates and instructor. This behavioral difference is likely if we consider the differences in students' behavior between these settings in their everyday contexts.

Landers et al. [1] investigated the relationship between personality traits and self-reported Internet usage and found that total Internet usage was negatively related to three of the Big Five traits - Agreeableness, Conscientiousness, and Extraversion. The relationship between Internet usage and these personality traits suggests that the personality factors that are predictive of higher performance in face to face classrooms (i.e., agreeableness and conscientiousness [7]) are tied to inaction in online settings. Lander's et al.'s results imply that those with personality types who use the Internet less than others might fare better in traditional learning environments when this environment supports their learning activities. These students may also choose to interact differently with online course materials, which indicates that personality traits may have a relationship with learning in online environments.

## 2.4 Student Behaviors in MOOCs: Social and Personality Factors

Only a few studies have explored the relationship between personality and MOOC usage. For instance, Chen and colleagues [16] tried to find whether personality influences learner behavior and learner success. In order to do so, they sent questionnaires to collect information about students' Big Five personality traits and combined this data with features describing students' activities in the course (e.g., number of forum posts and time watching videos). Their analysis revealed that various features describing system use were correlated with openness and conscientiousness for learners with low prior knowledge. For those with higher prior knowledge, only conscientiousness was related to multiple system usage features (i.e., the amount of time spent watching video lectures and number of quiz questions learners attempted). Another project aimed to predict student success based on their MOOC usage data from the first week of the course [8]. The course they analyzed had two study tracks (basic and scholar track), and the corresponding certificate was given to students based on their activities in the course. They used logistic regression to predict which certificate the learner received and whether the learner would drop the course or receive a normal certificate. Their models revealed that the students who were more connected in the forum in the first week were more likely to receive a certificate with distinction than a normal certificate.

Consistent with student success being tied to their connectedness within a course, social barriers can contribute to course attrition because our personality and course attributes (some courses are more collaborative than others) influence our behaviors. In [3], surveys were used to collect information about student experiences. Student responses revealed that a lack of social interaction was the main barrier to online learning. Similarly, [5] aimed to understand why people dropped out of MOOCs so they sent a questionnaire to the students who had dropped the course. Students said that "having little interaction with others" [5], which can be related to the style of the course as well as the personality of the individual, was a main reason behind their decision to drop the course. This may mean that level of collaboration in a course might have an impact on students' performance in that course.

Further exploring this idea of student interaction and community in MOOCs, Rosé and colleagues [4] tried to find the social factors that contribute to attrition in MOOCs by focusing only on student participation in the class discussion forums. Using a mixed membership stochastic blockmodel, students' transitions between subcommunities were tracked which showed that membership in one of the subcommunities significantly predicted the dropout rate. This finding suggests that being a member of certain subcommunities may increase a student's probability of dropping a course. However, belonging to a certain subcommunity could be related to personality (e.g. people with similar types of personality might enjoy each other's company). Thus, this work partially motivates our study of the effect of personality on academic performance in MOOCs.

Using a broader set of student behaviors across time, Kloft et al. [2] tried to predict MOOC dropout: they applied machine learning techniques (i.e., principal component analysis [PCA] and support vector machines [SVMs]) to students' clickstream and forum data. They computed attributes like number of requests, number of video views, and number of homework page views to predict whether a student would drop the course in any given week. Continuing in this vein, Zheng et al. [6] studied the effects of small group size on students' drop-out rate and learning performance in a MOOC. In their study, the people who responded to their initial questionnaire (asking them about their demographic and personality information) were automatically divided into groups of 10 by applying k-means clustering to this data. Other students were randomly divided into groups of 10. At the end of the course, a second email was sent to students asking them about their satisfaction. Based on these data, Zheng and colleagues were able to see that the students who had been grouped using k-means clustering had lower dropout rates, confirming that the community a student is a member of can influence their dropout, with personality playing a role in community membership. Since community membership and personality influence drop out, it is reasonable to think that they will also influence course scores. However, not all MOOCs have obvious subcommunities and aspects of how the course is designed can influence the development of these subcommunities, which are based on some form of mutual knowledge sharing and support. This type of knowledge sharing and support is fundamental to community development [25], but as the above examples demonstrate, has not been studied at the whole course level. Consequently, we do not yet know how course-wide collaboration and community within MOOCs as well as student personality predict student performance.

To better study how student performance in online courses relates to their personality and factors of the course design, we use the students' automatically logged interactions, a brief measure of their personality traits, information about how collaborative the course is, and a measure of classroom community to predict student grades as recorded by the system.

### 3 Method

Using a learning analytics approach, we aimed to predict student performance by augmenting automatically captured information about their grades and collaborative behavior with perceptual information that was collected from students via

questionnaires. For this research, we focused on the collaboration level in a course. Our hypothesis is that some personality types might do better in a course with certain attributes. For instance, introverts might feel uncomfortable in a course that requires a lot of collaboration with other learners and therefore they might not do very well. To measure this, we introduced two feature sets to predict student grades. Then, we applied machine learning algorithms to both feature sets to see whether adding level of collaboration in a course would increase the accuracy of our models.

Our hypotheses are:

- [H1] Students' MOOC grades can be predicted by the Big Five.
- [H2] The joint use of the Big Five and the level of collaboration within a course can predict student MOOC grades.
- [H3] Prediction of students' MOOC grades will be more accurate when both the Big Five and level of collaboration are used as predictor variables.

### 3.1 Dataset

To test our hypotheses, we performed secondary analyses on data from two MOOCs (Epidemics, Pandemics, and Outbreaks; Disaster Preparedness) that had been offered through the coursera platform. This data consists of a pre-course questionnaire which collected information about learner demographics and personality (using Gosling et al.'s short form of the Big Five personality traits [10]), students' grades in those courses, data about forum posts in each course, and a post-course questionnaire which included student responses to Rovai's classroom community scale (CCS) [22].

The questionnaires and system logs recorded learner data using a common identifier. This allowed us to link students' responses to their activities and performance within the MOOCs. We only predict the performance outcomes of those who responded to the pre-course questionnaire: In total, 323 students responded to this questionnaire. Of these students, 85 were taking the Epidemics, Pandemics and Outbreaks MOOC and 238 were taking the Disaster Preparedness MOOC. All students agreed to the use of their data for research purposes.

### 3.2 Data Pre-processing

**Data Cleaning: Removing Invalid Data.** We removed the students who: (a) had only partly completed the questionnaire since their partial responses mean that the Big Five result was inaccurate, (b) did not have a user ID associated with them as the result of a temporary bug in the survey software, (c) did not yet have a grade and filled the questionnaire recently (less than 6 months ago) since they may be still taking the course.

This cleaning reduced our dataset to include that from 306 students.

**The Big Five.** We followed Gosling's instructions [10] to calculate each student's scores for the Big Five personality traits: extraversion, neuroticism, openness, conscientiousness, and agreeableness. Although in [7] only openness, conscientiousness, agreeableness and neuroticism were found to be related to performance, we included

extraversion in our feature set because it might influence student performance when combined with the level of collaboration of the course.

**Grades.** Since grade is a continuous variable and many machine learning algorithms produce better models when the attributes are discrete rather than continuous [29], we needed to address this mismatch between data format and computational approaches. One of the ways to deal with this problem is by binning or categorizing the variables. For this purpose, we coded grade ranges to make the dependent variable discrete. We labeled grades 90% or higher as A, grades from 80% to 90% as B, grades from 70% to 80% as C, grades from 60% to 70% as D, grades from 50% to 60% as E, grades below 50% as F, and dropout as N. The dropout (N) group consisted of those who had completed the questionnaire more than 6 months ago, had not received a grade, and who had no activity in the last six months. This is possible in MOOCs that are offered through the coursera platform because people can transfer from one offering of a MOOC to the next without losing their progress.

**Course Collaboration.** We defined a measure of course collaboration using a combination of behavioral and perceptual data. We used the forum posts data for each course and divided the total number of questions and answers by the number of active users (an active user is a user who has posted at least one question or answer). This statistic was used as a proxy to for the level of collaboration within each course. The question and answer per active user was 2.7 for the Disaster Preparedness course and 4.6 for Epidemics, Pandemics, and Outbreaks.

We used the Classroom Community Scale (CCS) as a second proxy for the level of collaboration in a course [22]. For this purpose, we calculated CCS for each user in each course (using the post-course questionnaire) and then calculated the CCS average and standard deviation for each course. Then for each student, we added the CCS average and standard deviation according to the course they were enrolled in. The CCS average and standard deviation were 21.42 and 5.94 for the Disaster Preparedness course and 21.87 and 5.91 for Epidemics, Pandemics, and Outbreaks.

### 3.3 Data Analysis

**Model Building.** The independent variables derived from the pre-course questionnaire consisted of students' scores for each of the Big Five personality traits: extraversion, openness, conscientiousness, agreeableness, and neuroticism. The course question and answer per user statistic was the independent variable that was obtained using system logs. The post-course questionnaire provided the average and standard deviation of the CCS score for each course; these statistics provided the final independent variables for our models. Students' coded final grade (described above) served as the dependent variable.

We divided the dataset into training and test data. The training data was used to train our machine learning models, and the test data was used to see how good our trained model was. The training data was 70% of the dataset and the test data was 30% of the dataset. The assignment of data to the training and test data sets was random and it was

performed at the student level: all of the data from a single student was randomly put into either the test or the train set. After that, we ran different machine learning algorithms on the data with regularization (which helps to avoid overfitting and acts as feature selection) and k-fold cross-validation with  $k = 3$  (which avoids overfitting).

We ran different machine learning algorithms on the dataset. Our chosen algorithms were: Support Vector Machines (SVM) [15] and Logistic Regression [14]. SVM builds a model using the training data and uses that model to predict the test data. It is also one of the most common methods used in classification tasks since it is fairly robust and accurate. Logistic Regression uses the sigmoid function (logistic function) to predict the probability of a data point being in each class and assigns the class with highest probability to that data. Logistic regression is useful in this case because it is a simple model that performs well on relatively small amounts of data. Due to the small amount of data, more complex classifiers (e.g., Neural Networks) were not used as they can overfit the model to the data (i.e., create a classifier with high accuracy that generalizes poorly) and thus produce poorer results. Therefore, we chose classifiers that are simpler and are known to perform better on small data sets.

Each classifier was run with different parameters and the parameter with the best accuracy was found. The model was then tested using the test data. As is common in machine learning [28], we chose a set of predefined values for our hyperparameter the inverse of regularization weight,  $C$ , and tested the models with those values. Since each model takes a long time to run, it was only practical to test the models on a subset of values. A wide range of values was selected so that different magnitudes for the parameter could be tested. For SVM this parameter set was  $\{C: [1, 2, 10, 15, 20, 100, 1000], \text{kernel: linear}\}$ , and for Logistic Regression the parameter set was  $\{C: [1, 2, 10, 100, 1000]\}$ .

**Hypothesis Testing.** We ran our machine learning algorithms on the data with two different feature sets:

- `personality_test` - contains student scores for each of the Big Five personality traits. This feature set was used to test H1.
- `personality_collaboration_test` - contains student scores for each of the Big Five personality traits and the level of collaboration in a course (CCS average, CCS standard deviation, and the number of questions or answers per active user). This feature set was used to test H2.

We used the Zero Rule classifier as a baseline and compared the performance of our algorithms to it. We chose the Zero Rule classifier because it will perform better on our dataset since one of the values of performance (“N” which corresponds to dropout) is more frequent than the other grade classifications: 41.5% of our data was “N”. The Zero Rule classifier predicts each entry as “N”. It, therefore, has an accuracy of 0.415 for our data (127 students out of 306 were labeled as N).

To test H3, we used paired t-tests to determine whether the model based on personality alone (`personality_test`) was outperformed by the model that also relies on information about student collaboration (`personality_collaboration_test`).

## 4 Results

### 4.1 Individual Models

For `personality_test` the classifiers were run in 10 distinct runs with only personality traits as a predictor of the grade. The average and standard deviation of the models’ accuracy (percentage of correct predictions) are shown in Table 1. With this feature set, we see little if any improvement over a majority class prediction as represented by the Zero Rule classification model, thus H1 is not supported.

**Table 1.** Classifier accuracy using the Big Five and Collaboration as predictors of student grade

Classifier	personality_test	personality_collaboration_test
	M (SD)	M (SD)
SVM	0.410 (0.0323)	0.518 (0.0587)
Logistic Regression	0.411 (0.0374)	0.503 (0.0542)
Zero Rule	0.415	0.415

Next, we ran our classifiers with the `personality_collaboration_test` feature set. We again used 10 distinct runs. For this feature set, we had personality and collaboration in a course as the predictors of the grade. Collaboration was represented via proxies. These proxies were calculated at the course level and included CCS average, CCS standard deviation, and number of questions or answers per active user. Table 1 reports the average and standard deviation of model accuracy for both models. With this feature set, we see improved classification accuracy over the Zero Rule classifier. Therefore, H2 is supported.

### 4.2 Model Comparison

While the comparison of the Zero Rule classifier to the models based on each feature set suggested H3 would hold because these comparisons supported H2 and failed to support H1, we formally tested H3. To do this, we ran paired t-tests with a 95% confidence interval to determine whether adding the level of collaboration in a course as a predictor had an effect on classifier performance.

For SVM, the `personality_collaboration_test` feature set was more accurate than the `personality_test` feature set:  $t(9) = -5.3713, p < 0.001, d = -0.75$ . This difference is large. For the `personality_test` feature set, 125 out of 306 student grades were predicted correctly; and for `personality_collaboration_test`, 158 students were assigned the correct grade label.

For Logistic Regression, similar results were observed. The model using the `personality_collaboration_test` feature set was more accurate than the one using the `personality_test` feature set:  $t(9) = -0.3841, p < .001, d = -0.70$ . For the `personality_test` feature set, 126 out of 306 students’ grades were predicted correctly; and for `personality_collaboration_test`, 154 students’ grades were labeled correctly.

In both cases, models were more accurate when they used the combined feature set which supports H3. So, learner personality and the collaborative features of a course can be used together to predict learner grades. Furthermore, the similarity in results between the logistic regression and SVM model suggest that the data is linearly separable so other types of algorithms that group data by dividing that data using a line (logistic regression) or hyperplane (SVMs) may also outperform the Zero Rule classifier.

## 5 Discussion

In a MOOC setting, personality is not enough to predict student grades, even though personality has successfully predicted student scores in face to face learning environments [7, 9]. The inability of our models to outperform the Zero Rule classifier may be due to the fact that MOOC dropout rates are much higher than those of traditional classrooms. This difference in attrition means that one grade value (i.e., N - withdrawal) occurs far more frequently than others, making the accuracy of a Zero Rule classifier relatively high, which suggests that excluding all those who withdrew from the course may be appropriate in this context. However, this choice would fail to account for a sizable portion of the potential student population and their outcomes.

In addition to the difference in student attrition, aspects of student background are more variable in MOOCs than they are in most classroom settings. In traditional learning environments students usually have similar background preparation, live in the same cultural milieu, and typically speak a common language with facility. Whereas in MOOCs, students are from all around the world, have diverse cultural and language backgrounds, and have received widely different background preparation [26]. This may increase the number of parameters that affect academic performance. Thus, personality might not be enough to predict the academic performance of students in MOOC settings.

The results of our research show that we can predict grades in MOOCs when using both students' personality and the level of collaboration in a course. This supports findings from other contexts suggesting that the social elements of online courses are tied to academic performance [23, 24]. We build upon these findings that collaboration and community predict students' learning activities [23] and their perceived learning [24], by predicting student grades using two proxies for student collaboration: their sense of community and the amount of interaction within their MOOC discussion forum. Our findings confirm some qualitative work [23] that indicated individual student preferences, such as a desire to learn as part of a community and a desire to avoid interaction - which are both indicative of personality traits, influenced student engagement within their online courses. In our case, a lack of course engagement was represented through student attrition, whereas reduced activity patterns represented disengagement in earlier work that investigated online graduate-level courses [20, 23].

Building on this work, our results show that adding level of collaboration as an attribute of the course design helps improve the prediction of students' academic performance. This finding is aligned with those showing how the discussion forum facilitation method that instructors chose to encourage in their online courses influences student collaboration and learning experiences [18]. Our findings suggest, that while

personality can predict academic performance in classroom settings [7], additional information is needed if we want to accurately predict student performance in online courses, especially MOOCs. This may mean that a person with the same personality will perform differently based on the design of the online course, especially if one course promotes learning activities that align with that learner's personality while another promotes activities that conflict with that learner's personality.

These results hold potential for informing the creation of adaptive features within MOOCs and MOOC design. Given the relationship between features of courses (such as collaborative activities and forum discussion) and student personality, we may be able to design courses so that student activities support multiple personality traits. Essentially, the same learning objectives may need to be supported through multiple activities so that students who may preferentially select certain learning opportunities are not precluded from acquiring the knowledge or skills that the course is meant to develop. For instance, if one learner is more introverted and dislikes working with others, we may have him or her interact with an agent instead of completing work in a group or this learner may be asked to perform additional assignments or quizzes. In contrast, if a learner is more extroverted and is comfortable in larger groups and working with strangers then we can encourage that student to complete course work as part of a group.

## 5.1 Limitations

While our dataset was sufficiently large, collecting more data from a broader set of courses would be beneficial. It would also allow us to test model generalizability since student behaviors and elements of student background or personality are known to differ from one MOOC to the next [26].

Our measure of collaboration, while predictive and based on the work of others (e.g., [18, 22, 23]) could be improved through deeper analyses of the interactions in which students engaged. Moreover, this measure only captures one element of course design. As such, it demonstrates the importance of capturing the role of these type of features. Going forward, it would be beneficial to consider other elements of course design, including how the course was facilitated, its length, and the instructional domain.

## 6 Conclusion

Since little work has explored whether personality and features of course design predict student performance in MOOCs, we built predictive models of student scores based on individual students' Big Five personality traits and the level of collaboration that the student body experienced within a MOOC. These models were tested using two feature sets. The first feature set only consisted of individual students' personality traits and the second set consisted of their personality traits and the level of collaboration in their course. In the end, a students' personality was insufficient for predicting their MOOC score on its own (H1). The discrepancy between this finding and those from classroom-based studies reinforces the importance of validating prior findings from traditional educational settings in MOOC settings. Adding information about course design



features, specifically collaboration, improved the models to a point where they could predict student scores (H2 & H3). This reinforces the idea that MOOC design may be biased towards certain types of students [26, 27]. Further study is needed to see which other features of course design can be used to predict student scores and the extent to which these features of online courses interact to influence both student learning and their learning experiences.

## References

1. Landers, R.N., Lounsbury, J.W.: An investigation of Big Five and narrow personality traits in relation to internet usage. *Comput. Hum. Behav.* **22**, 283–293 (2006)
2. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: Predicting MOOC dropout over weeks using machine learning methods. In: *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (2014)
3. Muilenburg, L.Y., Berge, Z.L.: Student barriers to online learning: a factor analytic study. *Distance Educ.* **26**, 29–48 (2005)
4. Rosé, C.P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., Sherer, J.: Social factors that contribute to attrition in MOOCs. In: *Proceedings of the First ACM Conference on Learning @ Scale Conference - L@S 2014* (2014)
5. Gütl, C., Rizzardini, R.H., Chang, V., Morales, M.: Attrition in MOOC: lessons learned from drop-out students. In: Uden, L., Sinclair, J., Tao, Y.-H., Liberona, D. (eds.) *LTEC 2014. CCIS*, vol. 446, pp. 37–48. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10671-7\\_4](https://doi.org/10.1007/978-3-319-10671-7_4)
6. Zheng, Z., Vogelsang, T., Pinkwart, N.: The impact of small learning group composition on student engagement and success in a MOOC. In: *Proceedings of the 8th International Conference of Educational Data Mining*, pp. 500–503 (2015)
7. Komarraju, M., Karau, S.J., Schmeck, R.R., Avdic, A.: The Big Five personality traits, learning styles, and academic achievement. *Pers. Individ. Differ.* **51**, 472–477 (2011)
8. Jiang, S., Williams, A., Schenke, K., Warschauer, M., O’ Dowd, D.: Predicting MOOC performance with week 1 behavior. In: *Educational Data Mining 2014* (2014)
9. Komarraju, M., Karau, S.J., Schmeck, R.R.: Role of the Big Five personality traits in predicting college students academic motivation and achievement. *Learn. Individ. Differ.* **19**, 47–52 (2009)
10. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003)
11. Barrick, M.R., Mount, M.K.: The Big Five personality dimensions and job performance: a meta-analysis. *Pers. Psychol.* **44**, 1–26 (1991)
12. Costa, P.T., McCrae, R.: *Neo personality inventory-revised: (NEO PI-R)*. Psychological Assessment Resources, Florida (1992)
13. Schmeck, R.R., Ribich, F., Ramanaiyah, N.: Development of a self-report inventory for assessing individual differences in learning processes. *Appl. Psychol. Measur.* **1**, 413–431 (1977)
14. Agresti, A.: *Categorical Data Analysis*. Wiley-Interscience, New York (2002)
15. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995)
16. Chen, G., Davis, D., Hauff, C., Houben, G.J.: On the impact of personality in massive open online learning. In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pp. 121–130. ACM (2016)

17. Rose, D.: Personality as it relates to learning styles in online courses. In: Society for Information Technology & Teacher Education International Conference, pp. 827–831. Association for the Advancement of Computing in Education (AACE) (2012)
18. Demmans Epp, C., Phirangee, K., Hewitt, J.: Student actions and community in online courses: the roles played by course length and facilitation method. *Online Learn.* **21**, 53–77 (2017)
19. Scardamalia, M., Bereiter, C.: Pedagogical biases in educational technologies. *Educ. Technol.* **48**, 3–11 (2008)
20. Demmans Epp, C., Phirangee, K., Hewitt, J.: Talk with Me: student behaviours and pronoun use as indicators of discourse health across facilitation methods. *J. Learn. Anal.* **4**, 47–75 (2017)
21. Rovai, A.P.: Development of an instrument to measure classroom community. *Internet High. Educ.* **5**, 197–211 (2002)
22. Phirangee, K., Demmans Epp, C., Hewitt, J.: Exploring the relationships between facilitation methods, students' sense of community and their online behaviours. Special Issue *Online Learn. Anal. Online Learn. J.* **20**, 134–154 (2016)
23. Richardson, J.C., Maeda, Y., Lv, J., Caskurlu, S.: Social presence in relation to students' satisfaction and learning in the online environment: a meta-analysis. *Comput. Hum. Behav.* **71**, 402–417 (2017)
24. Garrison, D.R., Anderson, T., Archer, W.: Critical inquiry in a text-based environment: computer conferencing in higher education. *Internet High. Educ.* **2**, 87–105 (1999)
25. Baikadi, A., Schunn, C.D., Long, Y., Demmans Epp, C.: Redefining “What” in analyses of who does what in MOOCs. In: 9th International Conference on Educational Data Mining (EDM 2016), pp. 569–570. International Educational Data Mining Society (IEDMS), Raleigh (2016)
26. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: *Learning Analytics and Knowledge (LAK)*, pp. 170–179. ACM, New York (2013)
27. Kizilcec, R.F., Halawa, S.: Attrition and achievement gaps in online learning. In: *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, pp. 57–66. ACM, New York (2015)
28. Kuhn, M.: *Applied Predictive Modeling*. Springer, New York (2016)
29. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: a recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **32**(1), 47–58 (2006)



# Learning by Reviewing Paper-Based Programming Assessments

Yancy Vance Paredes<sup>1</sup>(✉) , David Azcona<sup>2</sup> , I-Han Hsiao<sup>1</sup> ,  
and Alan Smeaton<sup>2</sup> 

<sup>1</sup> Arizona State University, Tempe, AZ 85281, USA  
{yvmparedes, Sharon.Hsiao}@asu.edu

<sup>2</sup> Dublin City University, Glasnevin, Dublin 9, Ireland  
{David.Azcona, Alan.Smeaton}@insight-centre.org

**Abstract.** This paper presents a retrospective analysis of students' use of self-regulated learning strategies while using an educational technology that connects physical and digital learning spaces. A classroom study was carried out in a Data Structures & Algorithms course offered by the School of Computer Science. Students' reviewing behaviors were logged and the associated learning impacts were analyzed by monitoring their progress throughout the course. The study confirmed that students who had an improvement in their performance spent more time and effort reviewing formal assessments, particularly their mistakes. These students also demonstrated consistency in their reviewing behavior throughout the semester. In contrast, students who fell behind in class ineffectively reviewed their graded assessments by focusing mostly on what they already knew instead of their knowledge misconceptions.

**Keywords:** Programming learning · Reviewing behavior  
Educational technology · Educational data mining  
Behavioral analytics

## 1 Introduction

Successful learners monitor their own memory, comprehension, and performance to evaluate their progress. They use this information to adapt their current strategies and behavior [1]. Aside from motivation, metacognition, and resource management strategy, being able to monitor one's progress and understanding is critical to succeed in problem solving in programming learning [2–5]. Unfortunately, novices and experts employ different such self-regulated learning (SRL) strategies [6].

This raises several research questions that are worth investigating as research on SRL in programming learning is still limited. However, due to the complex nature of programming problem solving, research in this discipline involves using qualitative methods, such as questionnaires, think-aloud protocols, and interviews. These are used to code student's behaviors according to corresponding

SRL motivation and strategies. We have begun to see more empirical and qualitative mixed method studies reporting SRL during programming problem solving. For instance, students solve code rearrangement Parson problems using sub-goal labelling [7]; the iterative programming process framework supports SRL activities [3]; adequate prior knowledge affects the searching and evaluating processes in programming problem solving [8].

To address such limitations, researchers have started developing technologies that focus on integrating and modelling physical learning activities while making use of advanced learning analytics. Clickers [9] and multi-touch tabletops [10] are some of the examples. In our case, we developed a system that captures and connects multimodal learning analytics from both the physical and the digital worlds in the programming learning domain. It has the capability of digitizing paper-based artifacts, such as paper assessments, and providing an interface for grading and delivery of feedback to classes with large number of students. It logs how students interact with it (timing, frequency, sequence, attention, and changes of patterns when performing the reflecting actions towards their learning). Most importantly, the system supports students in managing their learning by integrating *assessment content*, *feedback*, and *learning outcome* in classes with blended instruction. Our goal is to systematically track students' learning activities across the physical and the digital spaces. In this paper, we focused on the monitoring and reflecting SRL behaviors.

The rest of the paper aims to answer the following research questions:

**RQ1:** What are the behavioral differences between high-achieving and low-achieving students in terms of monitoring and reviewing? Do high-achieving students review more thoroughly or frequently?

**RQ2:** What is the magnitude of difference in reviewing behavior of students when grouped according to their performance trajectories? Do *improving students* review differently from others?

**RQ3:** Which reviewing behaviors are more effective towards learning?

The paper is organized as follow. First, we discuss the role of feedback and behavioral analytics in programming learning. Next, we provide an overview of our research platform and the data gathering approach. Finally, we present the evaluation results along with its educational implications.

## 2 Literature Review

### 2.1 Feedback in Programming Learning

Feedback has been considered one of the most influential factors that affect educational achievement [11]. In Science, Technology, Engineering and Mathematics (STEM) subjects, such as programming, physics, or math, automated grading of assessment is one of the most popular methods in providing feedback. Such method is particularly pertinent for large classes as it guarantees a short turnaround time. Systems like WEB-CAT [12] or ASSYST [13] apply pattern-matching techniques that verify students' answers by running a set of unit test

cases and comparing them with the correct answers. Unfortunately, in the programming learning domain, these platforms typically check the concrete aspects of the solutions. The logic and reasoning of students are often neglected. As a result, instructors had to manually examine the program quality. Several alternative approaches have been proposed to address the issue of providing semantic and constructive feedback as well as the issue of scaling of the generation of feedback. For example: crowdsourcing code solutions which will then be suggested to students [14]; using parameterized exercises to create a sizable collection of questions to facilitate automatic programming evaluation [15]; PeerGrader [8] and PeerWise [16] utilizing student cohorts to provide peer feedback.

Regardless of the feedback generation methods, all of the above mentioned systems and approaches focused on evaluating digital artifacts, less is discussed in assessing paper-based programming problems. There has been a few relevant early innovations addressing the problem by digitizing exams (e.g. GradeScope [17]). Digitization essentially provides several advantages (e.g. some default feedback can be kept on the digital pages with the predefined rubrics; students' identity can be kept anonymous which eliminates any of the grader's biases, etc.) As our system has the ability to capture how students attend to their graded assessments, we explored these reviewing behaviors to understand their impacts on learning.

## 2.2 Behavioral Analytics in Programming Learning

Modelling student's programming learning is not a new topic. Student models reside in intelligent tutors or any adaptive educational systems. Student's learning is typically estimated based on their behavior logs, such as the interactions with tutors resulting in the updates on the knowledge components. In modelling programming language learning, several parameters are used to estimate students' coding knowledge. For instance, learning can be gauged based on the sequence of programming problem solving success [18], programming assignments progression [19], dialogic strategies [20], programming information seeking strategies [21], assignment submission compilation behavior [22], troubleshooting & testing behaviors [23], code snapshot process state [24], and generic Error Quotient measures [24]. Additionally, Educational Data Mining (EDM) techniques have helped educational researchers to analyze snapshots of learning processes, such as a combination of automated and semi-automated real-time coding to identify meaningful meta-cognitive planning processes in an online virtual lab environment [25]; supervised and unsupervised classification on log files and eye-tracking data to find meaningful events in an exploratory learning environment [26]; the sequences of reviewing and reflecting behaviors Hidden Markov Models (HMM) to predict students' learning performances [27]. In learning analytics literature, Blikstein [28] proposed an automatic analytic tool to access student's learning in an open-ended environment. This considers a range of behavioral analytics to predict learning, such as the amount of code changing, compilation behavior, and code editing.

### 3 Research Methodology

#### 3.1 Research Platform and Data Collection

WebPGA<sup>1</sup> was developed to serve as a platform that connects the physical and the digital learning spaces in programming learning. This system enables the digitization, grading, and distribution of paper-based assessments. Further details regarding the rationale and the design of the platform can be found in [27]. All events (which mostly are students' clickstream) are logged along with their timestamp. Examples of which include: logging in and out, clicking on a question to review, bookmarking a question, navigating through an exam, and taking of notes.

The data were collected from a classroom study conducted in a Data Structure and Algorithms course offered during the Fall 2016 semester. This class had a total of 3 exams and 13 quizzes. Among the 13 quizzes, only 6 were graded while the remaining 7 were recorded only for attendance (full credit was given regardless of the answers). There were 283 students enrolled in the class but only 246 (86.93%) were included in the study as those who dropped the course in the middle of the semester, did not take the three exams, or did not use the reviewing platform at all had to be removed. In this study, we analyzed *review actions* performed by students. A review action is an event where a student examines his or her graded answer. It includes reading the question, the answer, the assigned score and the feedback provided by the grader (see Fig. 1).

The screenshot displays a review interface for a programming problem. At the top, there are navigation options: "Review Tools", "Bookmark", and "I Know How to Solve". The problem statement is: "8. [1 pt] Write a pseudo code for the function `isMaxHeap(A, heapsize)` that checks if the input array `A` is a max heap or not. It should return true if `A` is a max heap and should return false otherwise. In this function, you cannot call any pre-existing function. You will need to utilize case if you want to use any function to call from `isMaxHeap` function. Also, your function should have its running time in  $O(n)$  where  $n$  is the size of the input array `A`." The student's handwritten answer includes a function signature `bool isMaxHeap(A, heapsize)`, a loop `for (i = n/2; i >= 1; i--)`, and conditional checks `if (A[i] < left(i) || A[i] < right(i)) return false;` and `return true;`. A tree diagram shows a root node 1 with children 2 and 3, and node 2 with children 4 and 5. The student's explanation states: "The function checks for all  $i = n/2$  down to 1 if the parent  $A[i]$  is smaller than either of the children: if yes: it returns false immediately and if no: that is, it is a max heap. it returns true." The running time is given as  $O(\log n) < O(n)$ . The score is 11.00. Feedback from the grader includes: "Following condition needs to be checked: whether both  $2i$  and  $2i+1$  (left and right child) are less than or equal to heapsize." The student has rated the feedback with 5 stars and has a personal note: "I should check if the children of the node satisfies the max heap property." There is a "Save Notes" button at the bottom.

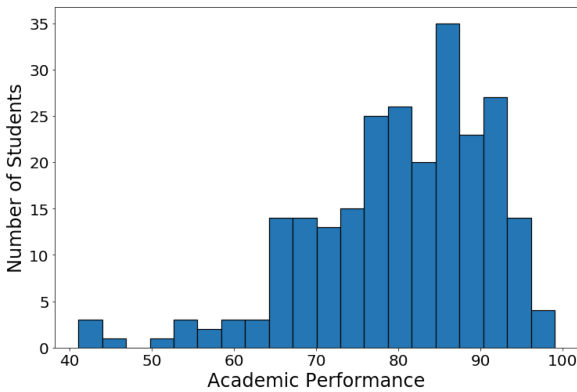
Fig. 1. Screenshot of what the student sees when reviewing his or her graded answer

<sup>1</sup> <https://cidsewpga.fulton.asu.edu/>.

### 3.2 Data Processing

In order to understand how students' monitoring and reviewing behavior affect their learning, students were labeled and grouped in two different ways. First, they were labeled according to their overall academic performance. Second, they were labeled according to their performance trajectory in a given period.

**Overall Academic Performance.** The average of the three exams was used to determine the overall academic performance of a student. Students were divided into two groups: *high-achieving* and *low-achieving*. Figure 2 shows the grades' distribution. Jenks natural breaks classification method [29] was used to identify the optimal break-point (77.60%) to divide the two groups.



**Fig. 2.** Distribution of academic performance of students

**Performance Trajectory.** The exams served as milestones to identify the change in the performance of the students in a given period. There were two time periods in this analysis, namely: *Exam1-Exam2*, between the first and the second exam, and *Exam2-Exam3*, between the second and the third. The difference in the scores between the second and the first exam in a given period is computed. Students are labeled *improving* if the difference is positive; *dropping* if negative; *retaining* if zero.

**Reviewing Behavior.** A total of  $N = 17,518$  review actions were extracted from the logs for this analysis. The score of the student in a particular question determines the label of a review action. The review action is labeled as *r\_correct* if the student got the question right. Otherwise, it is labeled as *r\_incorrect*.

### 3.3 Descriptive Data

An exam is considered reviewed if at least one of its questions is reviewed. Table 1 shows an overview of how students reviewed their exams. This includes the average performance of the class, the number of students who reviewed them, and the average time it took students before their first review attempt (hereinafter referred to as “reviewing delay”). A downward trend can be seen for both the number of students reviewing and their reviewing delay.

**Table 1.** Overview of students’ reviewing behavior

Exam	Avg. score	No. of students who reviewed the exam	Avg. time before first review attempt	Standard deviation
Exam1	81.2%	230 (93.50%)	4.5 days	14 days
Exam2	78.7%	224 (91.06%)	2 days	6 days
Exam3	80.6%	196 (79.67%)	0.8 days	2.4 days

In terms of exam reviewing behaviors, most students reviewed past exams before taking the next one. During the *Exam1-Exam2* time period, there were 217 students (88.21%) who reviewed Exam 1 prior to taking Exam 2. During the *Exam2-Exam3*, it was also 217 students (88.21%) who reviewed Exam 1 or Exam 2 (or both) prior to taking Exam 3. However, these may not necessarily be the same set of students.

## 4 Evaluation Results

### 4.1 Association Between Reviewing Behavior and Learning Performance

To identify the impact of reviewing exams on the learning performance of a student, we compared the effort exerted by high-achieving students and low-achieving students. To answer this question, the contents of the first two exams were manually inspected. Based on the number of questions in the two exams, a student only needs to perform at least 16 review actions to cover all the items. Following the Pigeonhole principle, performing more than 16 review actions could indicate that a question is reviewed more than once. Furthermore, performing less than 16 review actions could indicate that not all questions were reviewed. Review actions for the third exam were omitted since the impacts of these actions cannot be captured and measured anymore (the class has ended).

Table 2 summarizes the average number of review actions performed by the two groups. Using t-test, we found that high-achieving students significantly ( $t = -2.16, p = 0.03$ ) did more reviews than low-achieving students. It is interesting to note that, on average, high-achieving students performed 20.3 review actions. It could indicate that they reviewed their exams after it was made available and possibly prior to taking the next exam. This reflects their effort



in studying the material. On the other hand, low-achieving students, on average, performed 15.4 review actions. This shows how they barely reviewed their exams. This is clearly a bad habit since students are not able to take advantage of learning from the feedback they were provided, which could help them correct any of their misconceptions.

**Table 2.** Average review actions prior to Exam 3

Group	No. of students	Avg. review actions on exams	Standard deviation
High-achieving	158	20.3	19.1
Low-achieving	88	15.4	12.2

Doing more review does not necessarily translate to an effective one. Students may be doing a lot of review but not on the items where they really need to focus—their mistakes. Unfortunately, with the current grouping of students, it would not be surprising to find that majority of the review actions done by high-achieving students would be *r\_correct* (answers they got correctly). This is because it is dependent on their academic performance. Therefore, a different grouping was used to answer this question.

**4.2 Effectiveness of Reviewing Behavior**

To address the issue mentioned above, students were grouped according to their performance trajectory in a given period. Table 3 summarizes the average number of review actions done by the improving and dropping students. The retaining group was omitted since it only has few students. During the Exam1-Exam2 period, there was no significant difference on the number of review actions performed by the two groups. Interestingly, during the Exam2-Exam3 period, dropping students performed significantly more review actions. This possibly happened because during the Exam1-Exam2 period, students only had one exam to review, while during the Exam2-Exam3 they had two. This led us to investigate why there was a drop in the grades of those who reviewed more.

**Table 3.** Review count of improving and dropping students

Group	Exam1-Exam2			Exam2-Exam3		
	n	Mean	SD	n	Mean	SD
Improving	104	9.57	8.88	115	10.23	11.52
Dropping	109	8.93	8.45	100	11.89	12.54

**Table 4.** Reviewing behavior of improving and dropping students

Review action	Group	Exam1-Exam2		Exam2-Exam3	
		Mean	SD	Mean	SD
<i>R_Correct</i>	Improving	0.16	0.18	0.22	0.25
	Dropping	0.22	0.24	0.36	0.28
<i>R_Incorrect</i>	Improving	0.84	0.18	0.78	0.25
	Dropping	0.78	0.24	0.64	0.28

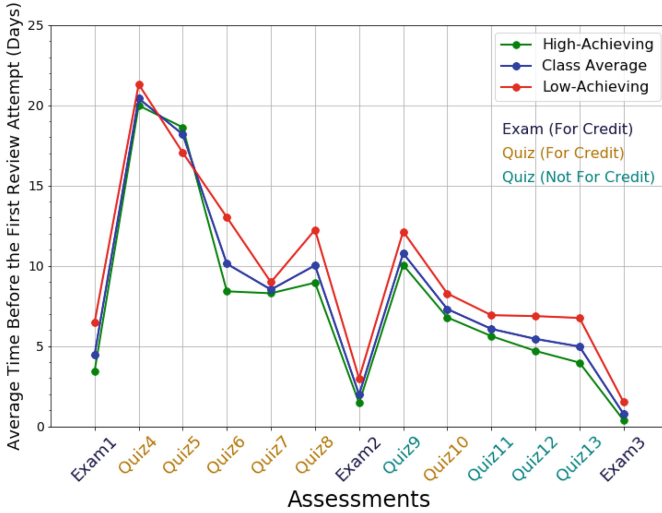
**Improving Group Reviewed Strategically and Effectively.** An improving student may not necessarily be a high-achieving student. It is interesting to investigate what led to the improvement of their exam scores. Table 4 summarizes the reviewing behavior of both the improving and dropping students. Although not statistically significant, improving students during the Exam1-Exam2 period reviewed their mistakes more than the dropping students ( $t = -1.82, p = 0.07$ ). During the Exam2-Exam3 period, a similar trend can be seen, but now statistically significant ( $t = -3.69, p < 0.05$ ). This shows that this strategy, where you focus on your mistakes to get them right, helps in improving your grade.

**Dropping Group Reviewed Ineffectively.** During the Exam1-Exam2 period, dropping students reviewed their correct answers more than the improving students, though not statistically significant ( $t = -1.82, p = 0.07$ ). However, during the Exam2-Exam3 period, the same trend was seen and is statistically significant ( $t = -3.69, p < 0.05$ ). Although dropping students devoted more time in reviewing their mistakes, the effort they spent was not enough. There was no improvement in their grades. It is also possible that they may have overlooked their mistakes. Since this effect was found in both time periods, this demonstrates the persistent ineffectiveness in reviewing of the dropping students. This is concerning especially for students who are struggling or experiencing difficulties in class. Intervention strategies should be developed and applied.

### 4.3 Reviewing Behavior Efficiency

The reviewing delay of students was modeled as a function of their review efficiency and their effort in learning the material. The average reviewing delay for each student (the average of all the delays for each assessment the student reviewed) was computed. Afterwards, it was correlated to their academic performance. It was found that there is a significant negative linear correlation that exists (Pearson's),  $r = -0.24, p < 0.05$ . This means that better performing students tend to attend and review their graded answers sooner.

The trend on how students attended to their assessments throughout the semester was visualized (see Fig. 3). Students were grouped according to their academic performance. The groups' average reviewing delay was computed. The first three quizzes were omitted since the logging feature was only introduced after the third quiz. It can be seen that high-achieving students generally spent less time before they begin to review their assessments (notice that the green

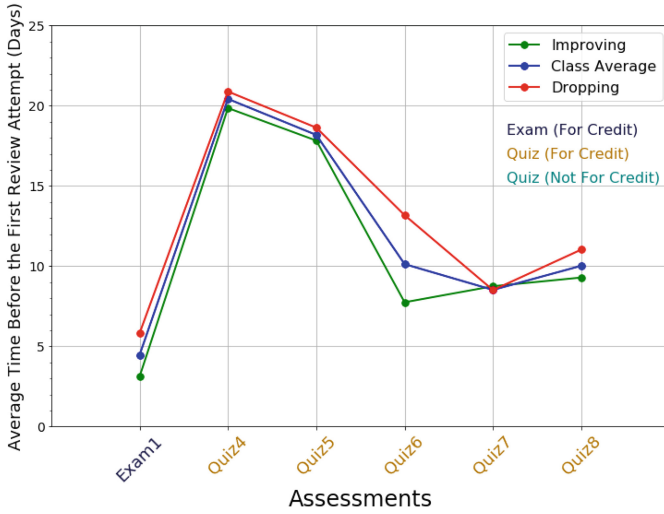


**Fig. 3.** The reviewing delay curve of students when grouped according to their overall academic performance (Color figure online)

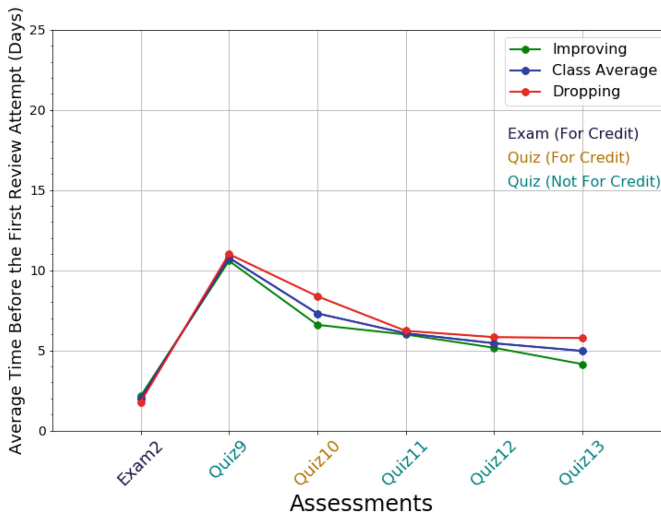
line is generally the lowest line throughout the semester). All students were more attentive in reviewing the three exams (shown by the dips) than the quizzes. This is not surprising. This suggests that the higher the credits that are at stake, the more attentive students become. One possible reason why students took longer time before they reviewed between the fourth to the sixth quizzes is that they did not review it right after it was made available. Students may have only reviewed them prior to taking Exam2. The chart also shows that students learned to use the platform over time as indicated by the downward trend. Interestingly, students started to review assessments sooner, even when the quiz was not for credit. This is an encouraging note and an evidence how students self-regulated their own learning in reviewing assessments.

The same steps were undertaken to investigate if similar findings could be obtained if students are grouped according to their performance trajectory. Unfortunately, there was no significant correlation between their magnitude of change (difference in their exam scores) and their average reviewing delay for both time periods.

Lastly, the trend for the two periods: Exam1-Exam2 (Fig. 4(a)) and Exam2-Exam3 (Fig. 4(b)) were visualized. It was not surprising to see that the improving group attended to their assessments sooner (notice that the green line is generally lower than the red line). This is consistent with the earlier finding which indicates that better performing students review sooner. This showed that being more vigilant in reviewing could potentially be associated to an improvement in grades. Another interpretation is that students who get better grades started seriously preparing for the exam earlier. Students likely reviewed their past exams at the start of preparing for the exam, so the fact that high-performing students did that earlier is not surprising.



(a) During the Exam1-Exam2 Period



(b) During the Exam2-Exam3 Period

**Fig. 4.** The reviewing delay curve of students when grouped according to their performance trajectory (Color figure online)

#### 4.4 Subjective Evaluation

An online survey was administered at the end of the semester to know the experience of students when using the system. Also, to identify possible features that could help them review effectively and efficiently. Only 74 students (30.10%)

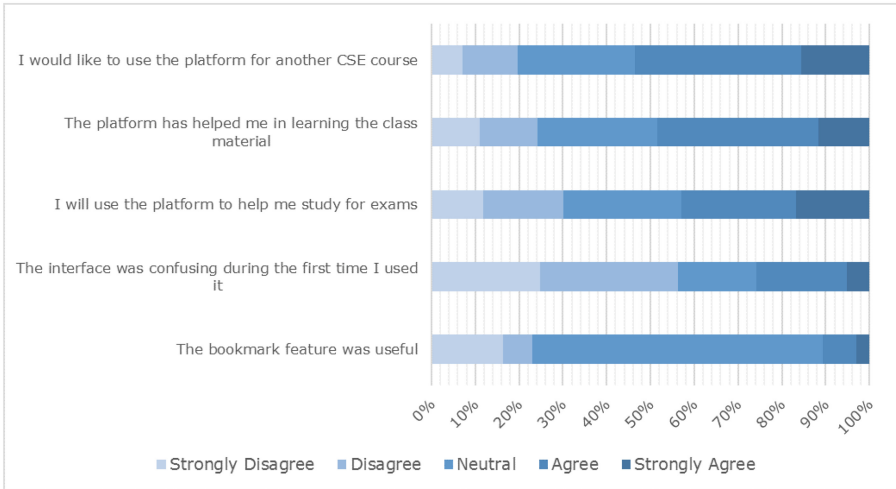


Fig. 5. Selected questions from the online survey

responded to the survey. Figure 5 shows some of the questions and the students’ responses.

**Learning and the Reviewing Platform.** More than half of the respondents (54.1%) believed that the system helped them learn the class material. When asked how they prepare for programming exams, they would review lecture notes (78.2%) or past assignments and assessments (68%). Some even create study guides (45.7%). We also found that 60% would even use our system. All these strategies involve a range of reviewing activities.

**Ease of Using the Platform.** The system enables the students to access and review their graded assessments anytime and anywhere. Majority (60.3%) of the respondents found the system easy to navigate and use as it only took them around 1–2 quizzes to be comfortable using it.

**Awareness of Features.** A color coding scheme was used to display the graded answers of the students, which majority of the respondents were aware of. However, some features, such as bookmarking and filtering were not used as they were not aware of the existence of such features. Finally, we asked for suggestions on how to improve user experience. One popular suggestion was for the system to be able to inform students what content or question to focus on when reviewing (51.2%).

## 5 Conclusion

This study focused on analyzing and understanding student reviewing and learning behaviors captured by an educational tool that enables students to review their paper-based assessments. A classroom study was conducted where data from a Data Structure & Algorithms class were collected. Students were grouped based on their overall performance: high-achieving and low-achieving; and based on their performance in a given time period: improving, retaining, and dropping. By comparing their reviewing behaviors, high-achievers were found to review more and quicker than low-achievers. Both improving and dropping students reviewed their mistakes. However, improving students reviewed and focused on their mistakes more than dropping students. This clearly indicates the effectiveness and the willingness of improving students to learn more from their mistakes. It also indicates a failure on the part of the dropping students to pay enough attention to address their misconceptions.

In addition, this study provides empirical data on how students review their paper-based assessments. This contribution could be used to improve the design of existing educational technologies. Letting the students focus on their mistakes (guided navigation) and advising them to attend to their graded assessments sooner (through prompts) would have a positive impact on their learning.

Finally, reviewing patterns can be extracted from the student behavioral actions and leveraged to train predictive models. These models will enable the further personalization of the feedback and potential interventions that will be provided to future students such as suggested reviewing assessments and material.

## 6 Future Work and Limitations

There are a number of limitations in the current study. The analysis only focused on students' voluntarily reviewing behavior to signify one of the self-regulated learning processes: in the abstract form of monitoring and reviewing their own learning. In the future, a more comprehensive scenario such as planning, comprehension monitoring, and self-explaining will be considered. In addition, the platform was informally introduced to students and no tutorial on how to use it was provided. They had to familiarize it on their own. The usability of the platform is currently being studied. Finally, this study will be further extended to other courses and cohorts to investigate the generalizability of these findings in Computer Science Education.

## References

1. Butler, D.L., Winne, P.H.: Feedback and self-regulated learning: a theoretical synthesis. *Rev. Educ. Res.* **65**(3), 245–281 (1995)
2. Bergin, S., Reilly, R.: The influence of motivation and comfort-level on learning to program. In: Proceedings of the 17th Workshop of the Psychology of Programming Interest Group, Sussex, UK, Psychology of Programming Interest Group, pp. 293–304 (2005)
3. Loksa, D., Ko, A.J.: The role of self-regulation in programming problem solving process and success. In: ICER, pp. 83–91. ACM, New York (2016)
4. Eteläpelto, A.: Metacognition and the expertise of computer program comprehension. *Scand. J. Educ. Res.* **37**(3), 243–254 (1993)
5. Hsiao, I.H., Bakalov, F., Brusilovsky, P., König-Ries, B.: Progressor: social navigation support through open social student modeling. *New Rev. Hypermedia Multimed.* **19**(2), 112–131 (2013)
6. Falkner, K., Vivian, R., Falkner, N.J.: Identifying computer science self-regulated learning strategies. In: Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education, pp. 291–296. ACM, New York (2014)
7. Morrison, B.B., Decker, A., Margulieux, L.E.: Learning loops: a replication study illuminates impact of hs courses. In: Proceedings of the 2016 ACM Conference on International Computing Education Research, pp. 221–230. ACM, New York (2016)
8. Gehringer, E.F.: Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bull.* **33**(1), 139–143 (2001)
9. Trees, A.R., Jackson, M.H.: The learning environment in clicker classrooms: student processes of learning and involvement in large university-level courses using student response systems. *Learn. Media Technol.* **32**(1), 21–40 (2007)
10. 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. *Int. J. Comput. Support. Collab. Learn.* **8**(4), 455–485 (2013)
11. Hattie, J., Timperley, H.: The power of feedback. *Rev. Educ. Res.* **77**(1), 81–112 (2007)
12. Edwards, S.H., Perez-Quinones, M.A.: Web-cat: automatically grading programming assignments. In: ACM SIGCSE Bulletin, vol. 40, pp. 328–328. ACM, New York (2008)
13. Jackson, D., Usher, M.: Grading student programs using assist. In: ACM SIGCSE Bulletin, vol. 29, pp. 335–339. ACM, New York (1997)
14. Hartmann, B., MacDougall, D., Brandt, J., Klemmer, S.R.: What would other programmers do: suggesting solutions to error messages. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1019–1028. ACM, New York (2010)
15. Hsiao, I.H., Sosnovsky, S., Brusilovsky, P.: Guiding students to the right questions: adaptive navigation support in an E-learning system for Java programming. *J. Comput. Assist. Learn.* **26**(4), 270–283 (2010)
16. Denny, P., Luxton-Reilly, A., Hamer, J.: Student use of the peerwise system. In: ACM SIGCSE Bulletin, vol. 40, pp. 73–77. ACM, New York (2008)
17. Singh, A., Karayev, S., Gutowski, K., Abbeel, P.: Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work. In: Proceedings of the Fourth ACM Conference on Learning@ Scale, pp. 81–88. ACM, New York (2017)

18. Guerra, J., Sahebi, S., Lin, Y.R., Brusilovsky, P.: The problem solving genome: analyzing sequential patterns of student work with parameterized exercises. In: Educational Data Mining, EDM, North Carolina (2014)
19. Piech, C., Sahami, M., Koller, D., Cooper, S., Blikstein, P.: Modeling how students learn to program. In: Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, pp. 153–160. ACM, New York (2012)
20. Boyer, K.E., et al.: Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach. *Int. J. Artif. Intell. Educ.* **21**(1–2), 65–81 (2011)
21. Lu, Y., Sharon, I., Hsiao, H.: Seeking programming-related information from large scaled discussion forums, help or harm? In: Proceedings of the 9th International Conference on Educational Data Mining, EDM, North Carolina, pp. 442–447 (2016)
22. Altadmri, A., Brown, N.C.: 37 million compilations: investigating novice programming mistakes in large-scale student data. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 522–527. ACM, NY (2015)
23. Buffardi, K., Edwards, S.H.: Effective and ineffective software testing behaviors by novice programmers. In: Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research, pp. 83–90. ACM, New York (2013)
24. Carter, A.S., Hundhausen, C.D., Adesope, O.: The normalized programming state model: predicting student performance in computing courses based on programming behavior. In: Proceedings of the Eleventh Annual International Conference on International Computing Education Research, pp. 141–150. ACM, New York (2015)
25. Montalvo, O., Baker, R.S., Sao Pedro, M.A., Nakama, A., Gobert, J.D.: Identifying students' inquiry planning using machine learning. In: Educational Data Mining, EDM, North Carolina (2010)
26. Bernardini, A., Conati, C.: Discovering and recognizing student interaction patterns in exploratory learning environments. In: Alevan, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 125–134. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13388-6\\_17](https://doi.org/10.1007/978-3-642-13388-6_17)
27. Hsiao, I.H., Huang, P.K., Murphy, H.: Uncovering reviewing and reflecting behaviors from paper-based formal assessment. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 319–328. ACM, New York (2017)
28. Blikstein, P.: Using learning analytics to assess students' behavior in open-ended programming tasks. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, pp. 110–116. ACM, New York (2011)
29. Jenks, G.F.: The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **7**, 186–190 (1967)





# Which Learning Visualisations to Offer Students?

Susan Bull<sup>1</sup>(✉), Peter Brusilovsky<sup>2</sup>, and Julio Guerra<sup>3</sup>

<sup>1</sup> Consultant, Birmingham, UK

s.bull.consult@gmail.com

<sup>2</sup> School of Information Sciences, University of Pittsburgh, Pittsburgh, USA

<sup>3</sup> Instituto de Informática, Universidad Austral de Chile, Valdivia, Chile

**Abstract.** Research on learning visualisations does not always consider open learner models (OLM), where visualisations support learner decision-making. A range of preferences has been found, but studies mostly compare visualisations within single systems, so some have not yet been contrasted. This paper: (i) offers OLM researchers further results based on screenshots that include a broader range of visualisations than previously; (ii) introduces OLM views for the attention of those in other e-learning fields, as these may be relevant to their context.

**Keywords:** Learning visualisations · Learner preferences  
Open Learner Model

## 1 Introduction

There is a need to better integrate research on learning analytics dashboards (LAD) and open learner models (OLM) [1]. LADs aim to make data actionable, commonly using traditional bar charts, line graphs, tables, pie charts, network graphs [2]. Learner models are dynamic models of learning that allow personalisation; OLMs externalise this model to aid learner decision-making [3]. Whilst some OLMs use traditional visualisations, many use other methods: Fig. 1 outlines examples. *Skill Meters 1&2* show knowledge level in the filled part of the meter. *Bullets* use fill in the bullet. *Graph* has positive data on the right of the axis; problems on the left. *Grid* uses colour to show understanding. *Table 1* lists competencies in columns from weak to strong; a dot in a cell indicates strength of each competency. *Table 2* ranks understanding. *Radar Plot* portrays learning across a curriculum by fill and position. *Histogram* shows knowledge from weak to strong. *Word Clouds* have strong competencies in larger text on the left; weak competencies on the right. *Treemap 1* shows competency by size of the corresponding area; *Treemap 2* uses colour (size shows number of problems). *Circle* also uses colour. *Network* and *Hierarchical Tree* have hierarchical structures similar to that shown by indenting sub-topics in *Skill Meters 1*, *Table 1*, or zooming in *Treemaps 1&2*. *Pre-requisites* and *Concept Map* show corresponding relationships.

In multiple view OLMs, skill meters tend to be viewed more if they are an option, though all views are accessed [4, 5, 11]. Nevertheless, whilst skill meters were popular amongst Fig. 1 screens when considering ‘what to work on next’, pre-requisites and

hierarchical tree were anticipated more useful for that purpose [12]; concept maps were more effective than skill meters to synthesise an overview in a controlled study [13]; and a simple ranked list was favoured over other views (including concept map, pre-requisites, hierarchical tree) in an experimental study [10]. Individuals may also use different views depending on reason for viewing [5]. This suggests a need to further investigate the relative usefulness of different views. Studies have typically been with single systems, so whilst several views have been compared, some have not yet been contrasted. As a first step, we follow approaches where screens were designed to gauge interest in options before deciding which to implement [14, 15], but we instead take visualisations from a range of *EXISTING* OLMs. These are from our own OLMs (for accessibility), but we use views that are similar to those commonly deployed.

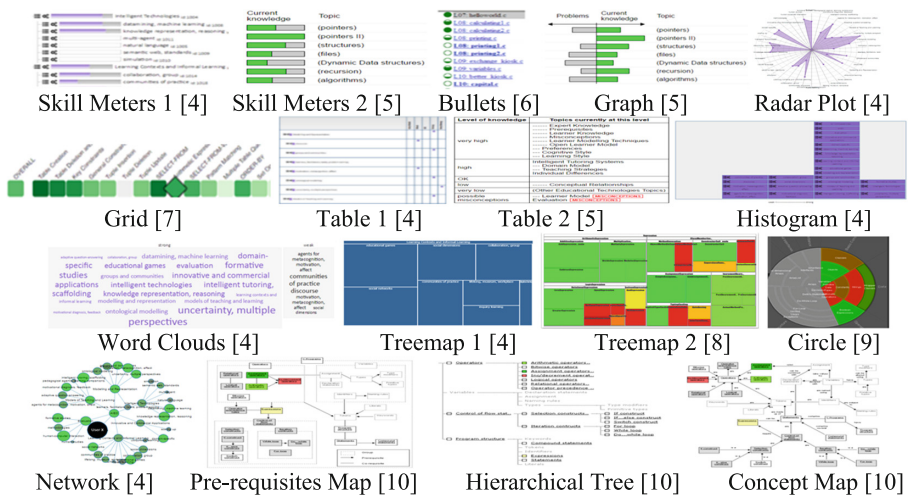


Fig. 1. Outline of layout of some common types of open learner model visualisation

## 2 Evaluation

38 students from School of Information Sciences, University of Pittsburgh, accepted an email invitation and were compensated \$20. 16 had previously used *Grid*. The Fig. 1 screens were shown. Likert scale questionnaires (strongly agree 5-strongly disagree 1) included space for comments. Differences in learning data (concepts, skills, knowledge level, competencies) were not highlighted, to avoid responses based on this.

Results: Each view had some expecting to use it, and some not. All anticipated using multiple views (mean 8.5; median 8; range 3–14). Table 1 shows *Skill Meters 1&2*, *Tables 1&2*, *Treemaps 1&2*, *Bullets*, *Graph*, *Pre-requisites Map* most easily understood: at least 30 claiming ‘I understand the purpose of [VIEW]’ (agree/strongly agree); then *Grid*, *Histogram*, *Network*, *Hierarchical Tree*, *Concept Map*, with 27–29. ‘In a system with many visualisations I would use [VIEW]’ had *Skill Meters 1&2* as most

**Table 1.** Understand/would use view; easily identify well known/not well known.

View	C	Understand/Would use				Identify well known/not well known			
		n. agree	mean	median	range	n. agree	mean	median	range
<u>Skill M 1</u>	+	<u>33/28</u>	<u>4.5/4.1</u>	<u>5/4</u>	<u>3–5/2–5</u>	<u>34/30</u>	<u>4.5/4.2</u>	<u>5/5</u>	<u>2–5/2–5</u>
<u>Skill M 2</u>	+	<u>31/27</u>	<u>4.6/4.0</u>	<u>5/4</u>	<u>3–5/2–5</u>	<u>36/32</u>	<u>4.6/4.4</u>	<u>5/5</u>	<u>2–5/2–5</u>
Bullets	+/-	<u>34/17</u>	<u>4.4/2.4</u>	<u>5/3</u>	<u>2–5/1–5</u>	<u>28/27</u>	<u>4.2/4.1</u>	<u>5/5</u>	<u>2–5/2–5</u>
<u>Graph</u>	+	<u>32/23</u>	<u>4.2/3.8</u>	<u>4/4</u>	<u>2–5/1–5</u>	<u>25/28</u>	<u>4.1/4.1</u>	<u>4/4.5</u>	<u>2–5/2–5</u>
<u>Grid</u>	+/-	<u>29/21</u>	<u>4.1/3.7</u>	<u>4/4</u>	<u>2–5/2–5</u>	<u>29/30</u>	<u>4.1/3.8</u>	<u>4/4</u>	<u>1–5/1–5</u>
Table 1	+/-	<u>33/14</u>	<u>4.4/3.1</u>	<u>4.5/3.5</u>	<u>3–5/1–5</u>	<u>26/26</u>	<u>4.0/4.0</u>	<u>4/4</u>	<u>1–5/2–5</u>
Table 2	+/-	<u>32/21</u>	<u>4.3/3.6</u>	<u>5/4</u>	<u>2–5/2–5</u>	<u>32/27</u>	<u>4.2/4.0</u>	<u>4.5/4.5</u>	<u>2–5/2–5</u>
Radar Plot	+/-	<u>23/18</u>	<u>3.8/3.5</u>	<u>4/3</u>	<u>2–5/1–5</u>	<u>21/22</u>	<u>3.7/3.7</u>	<u>4/4</u>	<u>1–5/1–5</u>
Histogram	–	<u>27/15</u>	<u>4.0/3.1</u>	<u>4/3</u>	<u>2–5/1–5</u>	<u>24/24</u>	<u>3.8/3.7</u>	<u>4/4</u>	<u>2–5/2–5</u>
Word Cl	–	<u>17/12</u>	<u>3.5/2.8</u>	<u>3/3</u>	<u>2–5/1–5</u>	<u>19/12</u>	<u>3.3/3.0</u>	<u>3.5/3</u>	<u>1–5/1–5</u>
Treemap 1	–	<u>30/17</u>	<u>4.0/3.1</u>	<u>4/3</u>	<u>2–5/1–5</u>	<u>23/24</u>	<u>3.6/3.7</u>	<u>4/4</u>	<u>1–5/1–5</u>
Treemap 2	–	<u>31/12</u>	<u>4.1/3.9</u>	<u>4/3</u>	<u>2–5/1–5</u>	<u>23/16</u>	<u>3.6/3.3</u>	<u>4/3</u>	<u>1–5/2–5</u>
Circle	+/-	<u>24/17</u>	<u>3.8/3.2</u>	<u>4/3</u>	<u>2–5/1–5</u>	<u>21/23</u>	<u>3.7/3.6</u>	<u>4/4</u>	<u>2–5/1–5</u>
<u>Network</u>	+/-	<u>29/19</u>	<u>4.2/3.5</u>	<u>4.5/3.5</u>	<u>2–5/1–5</u>	<u>19/21</u>	<u>3.6/3.0</u>	<u>3.5/4</u>	<u>1–5/1–5</u>
<u>Pre-req M</u>	+/-	<u>31/22</u>	<u>4.1/3.6</u>	<u>4/4</u>	<u>2–5/1–5</u>	<u>26/24</u>	<u>4.0/3.7</u>	<u>4/4</u>	<u>2–5/1–5</u>
<u>Hier Tree</u>	+/-	<u>29/25</u>	<u>4.2/3.7</u>	<u>4/4</u>	<u>2–5/1–5</u>	<u>23/26</u>	<u>3.9/4.0</u>	<u>4/4</u>	<u>2–5/1–5</u>
<u>C Map</u>	+/-	<u>28*/19</u>	<u>4.1/3.4</u>	<u>4/3.5</u>	<u>2–5/1–5</u>	<u>20/22</u>	<u>3.6/3.5</u>	<u>4/4</u>	<u>1–5/1–5</u>

*underlined: at least half agree/strongly agree. \*one missing answer for questionnaire item.*

likely (>70%); and at least half chose *Graph*, *Grid*, *Table 2*, *Network*, *Pre-requisites Map*, *Hierarchical Tree*, *Concept Map*. For the two items ‘I could easily identify topics I know well/do not know well using [VIEW]’: at least half responded positively for well known topics for all views; only *Word Clouds* and *Tree map 2* had less than half for topics not known well. *Skill Meters 1&2* scored especially high for both.

Most views attracted positive (+) and negative (–) comments (C). Table 2 provides typical examples, often indicating the positive uses of detail as well as negative perceptions of too much detail. As domains often use a hierarchical structure, to further highlight individual differences in preferences, Table 3 shows the hierarchical views chosen by the ten students expecting to use the least views (3–6) overall (mean 1; median 1; range 0–2). The most popular, *Skill Meters 1*, was selected by only half; *Hierarchical Tree*, by three; a *Treemap*, by two. Three participants anticipated using no hierarchical view, but each of these selected *Table 2*, which CAN show structure in topic labels (e.g. dashes before sub-topics, as in the screen in the study). The other 28 students (7 or more views) chose at least one hierarchical view (mean 3.7; median 4; range 1–6). A combination of *Skill Meters 1*, *Hierarchical Tree* and *Network* would cover these participants’ preferences; omitting any one of these would leave only one student with no preferred hierarchical structure. (Considering all participants together, values for expecting to use hierarchical views were: mean 3; median 3.5; range 0–6.)

Discussion: OLMs use not only traditional methods of information visualisation often found in LADs (see [2]), but also other options, from simple displays for a quick overview, to highly structured views that include information about relationships. This paper identified that, *APART FROM Radar Plot, Word Clouds and Circle*, all views were claimed to be understood by at least two thirds (and only *Word Clouds* by less than half). For a multiple-view OLM, over 70% of participants anticipated using *Skill Meters 1&2*; and at least half, *Graph, Grid, Table 2, Network, Pre-requisites Map, Hierarchical Tree, Concept Map*. These nine visualisations were also considered easily usable to identify well known and less known topics by at least half of the students.

**Table 2.** Typical responses for some of the visualisations claimed as most likely to be used.

View	Positive (+)	Negative (-)
Table 2	It has topics of similar knowledge level blocked together	I don't like reading too much words while they are not clearly categorized
Pre-req M	Can help me to make a study plan to know what I need to learn step by step	Too much information in it
Hier Tree	The hierarchy helps to navigate to particular topic I want to go	So much information [...] tedious to find out what is being told
C Map	Can help me specify the relationship between knowledge	Way too confusing

*Skill Meters 1&2* were the top views for all four questionnaire items. This echoes findings that skill meters are used frequently in practice when amongst the options available [4, 5, 11]. However, initial results looking at the same views to identify 'what to work on next' revealed different preferences (*Pre-requisites Map, Hierarchical Tree*) [12], compared to findings here for well known/less known topics (*Skill Meters 1&2*), though over half also stated they would use *Pre-requisites Map* and/or *Hierarchical Tree* in a multiple-view OLM. Furthermore, of those who expected to use fewer views overall, only half anticipated using the hierarchical *Skill Meters 1*. We therefore recommend considering providing each of the above (*Pre-requisites Map, Hierarchical Tree, one of Skill Meters 1&2*) to the extent that the domain structure allows.

**Table 3.** Hierarchical views expected to be used by those using the fewest views overall (3–6).

	P1(3)	P2(4)	P3(4)	P4(5)	P5(5)	P6(5)	P7(5)	P8(6)	P9(6)	P10(6)
Skill M 1	X	X					X		X	X
Table 1										
Trm 1&2			X							X
Network										
Hier Tree						X	X		X	

A controlled study found a concept map more effective than skill meters for synthesis of an overview [13]; a ranked list was used more often than other views in an experimental study [10]. Our *Concept Map* had three quarters claiming it understandable, and half stated they would use it in a multiple-view OLM. *Table 2* positions topics on five levels, similar to a ranked list, and is considered understandable by as many as for *Skill Meters 1&2*, and over half stated they would use it. We therefore further propose considering *Table 2* and *Concept Map*. However, we also had participants finding the above views difficult. Comments revealed perceptions of too much detail; others found detail useful for specific purposes: identify knowledge (*Table 2*) or conceptual relationships (*Concept Map*), form a study plan (*Pre-requisites Map*), navigation (*Hierarchical Tree*). This supports multiple views for different users, or individuals for different goals. Students can explain why they use different views at different times [5], indicating they understand the relative benefits and limitations *FOR THEM*; and consistent with OLMs where other views are used as well as skill meters [4, 5, 11]. Whilst we largely support using multiple views, when an OLM has a specific purpose, we suggest a view similar to one of the above (e.g. concept map for an overview of understanding of conceptual relationships; pre-requisites to plan activities).

Since the range values show most views had some students expecting to not use them, we also consider alternatives from the remaining views that had at least half anticipating using them and stating they would be able to identify known and less known topics: *Graph*, *Grid* and *Network*. We do not suggest *Network* or *Graph* if only one view is to be employed, since these were not the most used views in their respective deployed OLMs [4, 5]; and *Network* shares the structure of *Hierarchical Tree*, with *Graph* very similar to *Skill Meters 2* - both already recommended above (with higher scores). However, these may be useful as additional views to provide greater choice. Thus, we propose the above 8 visualisations be considered as options in a multiple-view OLM. *Grid* is more difficult to compare, since 16 participants were already familiar with it. This may have inflated the values, but *Grid* has been successfully used in practice in a single-view OLM [7], and may be helpful when space in an interface is limited (e.g. where the OLM is displayed together with course content). *Bullets* are an interesting case: they were claimed to be more easily understandable than any other visualisation, scored well for identification of known/less known topics, but had less than half expecting to use them. For this reason, we suggest *Bullets* for cases where space is especially restricted as they will likely be understood, and offer information in a similar manner to *Skill Meters 2*. However, we would not propose *Bullets* as a replacement where *Skill Meters 2* (or *1*) can be used. Finally, *Treemap 1* could be considered for a large, many-layered hierarchical domain, also to efficiently use available space. (These three views may also be useful as additional options in a multiple-view OLM.) *Table 4* summarises our initial suggestions, to be supplemented with any additional visualisations appropriate for the specific context.

Because domains are often structured hierarchically, we further consider responses relating to these views. The median for anticipated use of hierarchical views was 3.5. *Skill Meters 1*, *Hierarchical Tree* and *Network* together would satisfy all 28 participants anticipating using at least 7 views; removing any would leave only one person with no preferred option. However, amongst the ten choosing only 3–6 views overall, four would have no favoured hierarchical view. Three of these expected not to use any

of the hierarchical views, but these all opted for *Table 2*, which *CAN* be configured to show hierarchy levels - albeit not within the hierarchical structure (as (sub-)topics are simply ranked). Nevertheless, including *Table 2* may raise awareness of the hierarchal structure, and lead to a better understanding of the full hierarchical view(s). In this case, only one student (who chose a *Treemap*) would have no preferred view.

Participants predicting use of a feature does not necessarily mean they will use it in practice [16], but offering visualisations that students can anticipate easily using may lead them to try to interpret the information in context. This paper therefore offers a *STARTING POINT* for OLM and LAD designers who are selecting learning visualisations to incorporate. Further research into these combinations in use can then be undertaken.

**Table 4.** Visualisations suggested for consideration for single and multiple-view OLMs.

Views	SM1&2	Bullets	Graph	Grid	T2	TM1	Net	Pre-r	HT	CM
Single view	X				X			X	X	X
Single/restricted		X		X		X				
Main mult views	X				X			X	X	X
Additional views		(X)	X	(X)		(X)	X			

### 3 Summary

Building on findings that students may use different views when there are multiple options in an OLM, we studied reactions to typical OLMs to explore relative benefits and drawbacks of view combinations. The screenshots were largely judged understandable, though there were differences in expected use. Combining views designed to fit the purpose of viewing with ones previously successfully used, also taking into account our findings here, is a step towards providing useful alternatives in future e-learning systems, as well as for considering options when only one view is preferred.

### References

1. Bodily, R., Kay, J., Alevan, V., Davis, D., Jivet, I., Xhakaj, F., Verbert, K.: Open learner models and learning analytics dashboards: a systematic review. In: LAK. ACM (2018)
2. Schwendimann, B.A., et al.: Understanding learning at a glance: an overview of learning dashboard studies. In: LAK, pp. 532–533. ACM (2016)
3. Bull, S., Kay, J.: SMILI©: a framework for interfaces to learning data in open learner models, learning analytics and related fields. IJAIED **26**(1), 293–331 (2016)
4. Bull, S., Johnson, M.D., Masci, D., Biel, C.: Integrating and visualising diagnostic information for the benefit of learning. In: Reimann, P., Bull, S., Kickmeier-Rust, M., Vatrappu, R.K., Wasson, B. (eds.) *Measuring and Visualizing Learning in the Information-Rich Classroom*, pp. 167–180. Routledge Taylor & Francis, New York (2016)

5. Bull, S., Mabbott, A.: 20000 inspections of a domain-independent open learner model with individual and comparison views. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 422–432. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_42](https://doi.org/10.1007/11774303_42)
6. Brusilovsky, P., Yudelson, M.V.: From WebEx to NavEx: interactive access to annotated program examples. *Proc. IEEE* **96**(6), 990–999 (2008)
7. Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V.: The value of social: comparing open student modeling and open social student modeling. In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) UMAP 2015. LNCS, vol. 9146, pp. 44–55. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-20267-9\\_4](https://doi.org/10.1007/978-3-319-20267-9_4)
8. Brusilovsky, P., Baishya, D., Hosseini, R., Guerra, J., Liang, M.: KnowledgeZoom for Java: a concept-based exam study tool with a zoomable open student model. In: ICALT. IEEE (2013)
9. Hsiao, I.-H., Bakalov, F., Brusilovsky, P., König-Ries, B.: Progressor: social navigation support through open social student modeling. *NRHM* **19**(2), 112–131 (2013)
10. Mabbott, A., Bull, S.: Student preferences for editing, persuading, and negotiating the open learner model. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 481–490. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_48](https://doi.org/10.1007/11774303_48)
11. Duan, D., Mitrovic, A., Churcher, N.: Evaluating the effectiveness of multiple open student models in EER-tutor. In: ICCE, APSCE, pp. 86–88 (2010)
12. Bull, S., Brusilovsky, P., Guerra, J., Araujo, R.: Individual and Peer Comparison Open Learner Model Visualisations to Identify What to Work On Next, UMAP-Ext LBR4 (2016)
13. Maries, A., Kumar, A.: Effect of Student Model on Learning. In: ICALT, pp. 877–881 (2008)
14. Guerra-Hollstein, J., Barria-Pineda, J., Schunn, C.D., Bull, S., Brusilovsky, P.: Fine-grained open learner models: complexity versus support. In: UMAP. ACM (2017)
15. Law, C.-Y., Grundy, J., Cain, A., Vasa, R.: A preliminary study of open learner model representation formats to support formative assessment. In: COMPSAC, p. 887. IEEE (2015)
16. Nielsen, J.: First Rule of Usability? Don't Listen to Users. Nielsen Norman Group (2001). [nngroup.com/articles/first-rule-of-usability-dont-listen-to-users](http://nngroup.com/articles/first-rule-of-usability-dont-listen-to-users). Accessed 23 June 2018



# Detection of Student Modelling Anomalies

Sergey Sosnovsky<sup>1(✉)</sup>, Laurens Müter<sup>1</sup>, Marc Valkenier<sup>1</sup>, Matthieu Brinkhuis<sup>1</sup>,  
and Abe Hofman<sup>2</sup>

<sup>1</sup> Utrecht University, 3584 CC Utrecht, The Netherlands

{s.a.sosnovsky,l.h.f.muter,m.p.valkenier,m.j.s.brinkhuis}@uu.nl

<sup>2</sup> University of Amsterdam, 1001 NK Amsterdam, The Netherlands

a.d.hofman@uva.nl

**Abstract.** As the modern TEL tools gain wider adoption in real educational contexts, they start facing important practical problems. One such problem for adaptive educational systems is the reliability of their student modelling mechanisms. Even when such a mechanism has been tested and calibrated to represent students' knowledge reasonably well, the student herself can become a source of problems. Students can use the system in a non-intended way, exhibit long periods of off task behaviour, try gaming the system, seek help of parents or peers, etc. Such usage patterns will manifest themselves in sequences of activity that do not represent student abilities and will result in student modelling anomalies causing subsequent suboptimal adaptive interventions from the system. This would be very important for a system that is used in real classrooms with younger children, especially, when it is also available at home as a supporting tool for independent work. This paper reports a study of such a system – Math Garden. Several user modelling anomalies have been detected in its logs. First steps towards building an automated tool for on-the-fly student modelling anomaly detection are reported.

**Keywords:** Student modelling · Adaptive educational system  
Educational data mining · Student modelling anomaly

## 1 Introduction

Student modelling is one of the central mechanisms of any adaptive educational system (AES). Without the correct estimation and timely updates of students' knowledge, an AES would not be able to optimize its content and behaviour towards individual students' strengths and weakness. When these estimations are incorrect, the entire adaptivity workflow is compromised. In such a situation, the adaptive interventions of the system can be suboptimal, inefficient, or even harmful. Hence, accuracy of modelling a user is important for any adaptive system, yet for an AES, this importance is magnified by the delicacy of learning as information activity and a student as a category of users. Unlike other users, students are by definition novices in the domain (and, often, even the tasks) supported by the adaptive system<sup>1</sup>. Besides, unlike users of other adaptive systems,

---

<sup>1</sup> Compare a student supported by AES with a user of an adaptive search system or an adaptive recommender.



students are more often children. This means, students are very susceptible to systems' errors; they have hard time detecting them and recovering from them on their own. The importance of these issues has been long recognized by the AES research community. Several methods for evaluating accuracy and predictive validity of student models have been proposed [1, 2]. It has become a standard to subject newly developed student modelling mechanisms of AESs to thorough examination (see [3] for a comprehensive example).

Another interesting aspect of students separating them from other types of users is their tendency to engage in non-productive behaviour, often due to a lack of motivation. This can further harm the student modelling process, the consecutive adaptive interventions, and at the end, the students themselves as has been shown, e.g. in [4]. Several categories of such behaviour have been described in literature. A large body of research focused on various types of "gaming the system" [5], which can be described as abusing system's features in order to achieve results other than learning. Examples of gaming the system can range from excessive usage of the help functionality of an AES [5] to systematic guessing of a correct answer. Students also engage in other types of off-task non-productive behaviour, for example, by playing with the content they have already mastered [6].

While successfully researched in the lab, these phenomena are rarely addressed in commercial adaptive educational software. This is unfortunate, as in a *real* classroom, failure of student modelling components to address instances of "abnormal" learning activities lead to *real* learning problems. In real setting, absence of control (available for researchers during lab studies) is likely to produce more often occurrences of these phenomena. In addition, the actual usage context brings new potential causes of erroneous student modelling, such as incidental help from parents that can shortly increase student performance or an illness that can shortly lower it. Such situations result in abnormal V-shaped patterns of learning activities that bring about anomalies in student modelling. Then, the system over- or underestimates the current level of student proficiency and starts conducting the learning process on an either too high or too low level of difficulty, which in turn can cause frustration or boredom [7]. In this paper, we analyse usage logs of a commercial AES Math Garden to discover instances of such student modelling anomalies.

The rest of the paper is organized as follows. Section 2 gives a short description of the Math Garden system and the student mechanism it employs. Section 3 briefly describes types of student modelling anomalies analysed in this paper. Section 4 outlines the conducted data mining experiment. Section 5 concludes the paper with the summary of the results and plans for future work.

## 2 Adaptive Learning Support with Math Garden

The Math Garden system [8] supports learning by adaptively providing practice items tailored to a specific student's ability at a specific time [9]. The relative difficulty of items can be adjusted to individual students, so regardless of the practice domain and ability, about 70% of all adaptively selected items are answered correctly.

Students answering items in a practice context can be regarded as paired comparisons for which we like to estimate dynamic ratings, since both item difficulty and student ability are expected to change over time (e.g., [10, 11]). The Elo Rating System (ERS)

has a history in the chess community, where matches can be considered as paired comparisons and where dynamically changing abilities of chess players are expressed in Elo ratings [12]. In an educational context, some modifications are required to both continually estimate the difficulty of the items and the ability of the students [13].

User  $p$  responding to an item  $i$  is considered a match between the user and the item, which can be won (correct response) or lost (incorrect response) with as outcome as score  $S_{pi}$ . Ratings for user ability  $\theta_p$  and for item difficulty  $\delta_i$  are updated after every score as follows:

$$\theta_p \rightarrow \theta_p + K(S_{pi} - E(S_{pi})), \quad \delta_i \rightarrow \delta_i - K(S_{pi} - E(S_{pi}))$$

where  $K$  is a scaling factor and the expected score  $E(S_{pi})$  is a simple logistic function on the difference between the current ability estimate  $\theta_p$  and item difficulty estimate  $\delta_i$ . Simply put, winning a match always adds some points to the user rating  $\theta_p$ , while losing subtracts some points. The amount of points at stake is determined by the difference between user ability  $\theta_p$  and for item difficulty  $\delta_i$ , where  $E(S_{pi}) = .5$  if these are equal. The ERS allows to track the development of person ability estimates  $\theta_p$  and item difficulty estimates  $\delta_i$ , in real time. The availability of current ratings can be used, amongst others, to adaptively select items tailored to the user, to provide teachers with feedback on development in subdomains, and to calculate reference groups.

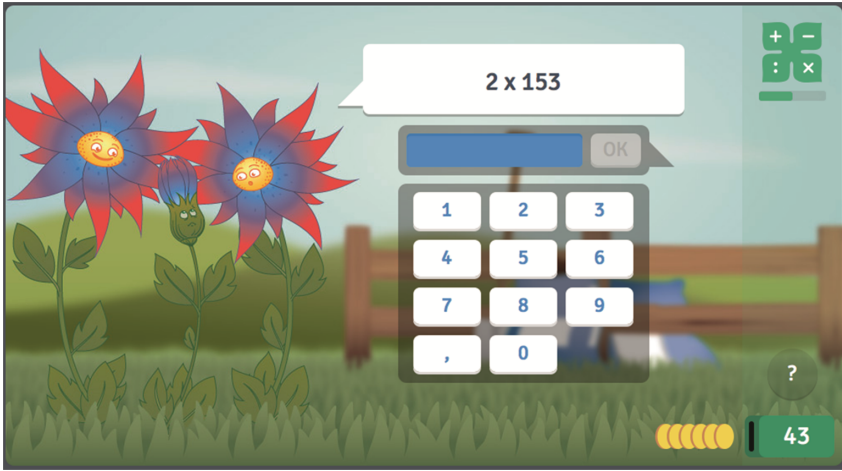
### 3 Student Modelling Anomalies

As a student learns a domain, the level of her knowledge steadily grows. In Math Garden, this will be reflected in a gradually rising Elo rating curve. Sometimes, a student masters a certain skill (e.g. how to multiply numbers greater than 10). This will result in a steep rise of her Elo rating as she starts successfully solving items of higher difficulty. Such patterns of student model estimations growth correspond to normal learning behaviour.

However, when we observe a sudden increase in Elo rating followed by a sudden drop, we consider this a student model anomaly representing an abnormal student behaviour followed by a period of correction. This can be the result of a session with a parent, sibling, teacher or a friend, when a more knowledgeable peer assists the student in solving learning tasks. If this assistance does not translate into learning, once it is removed, the student receives a sequence of tasks that are too difficult for her. Another possible reason for such a pattern is a short account take over by a parent who can try the software unaware of the effect it will have on the following adaptive behaviour of the system. In the next section, we refer to this anomaly as the low-high-low (LHL) pattern.

Another possible anomaly is the sudden drop in Elo rating followed by a step rise. This can happen due to a student practicing during illness or due to a student gaming the system. The later has been known to happen in Math Garden as the students try to deliberately lower the difficulty of tasks in order to be able to answer easy tasks faster.

As a result, the system granted them more “coins” (see Fig. 1). In the next section, we refer to this anomaly as the high-low-high (HLH) pattern.



**Fig. 1.** An example Math Garden item from the domain of multiplication. The on-screen buttons or a keyboard can be used to provide an answer. The question mark button can be used to skip the item. The item is adaptively selected to match its difficulty to the ability of the student. The amount of coins gradually decreases the more time a student spends to provide an answer.

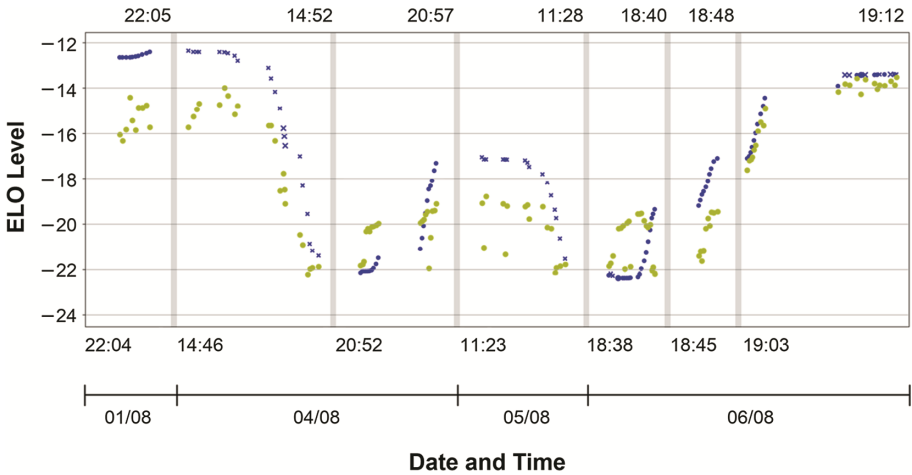
## 4 Experiment

For this study, a dataset collected by Math Garden has been used. It contains more than 10 million records of more than 90,000 students practicing multiplication exercises over the course of one year. Student accounts associated with too much activity (more than 1000 attempts) have been excluded from the consecutive analysis. Often, such accounts belong to teachers, or represent logins shared by entire classrooms if a teacher does not follow authentication guidelines. Students with too little activity (less than 20 attempts) have been also filtered out as they did not accumulate enough history for the student modelling mechanism to even start reliably predicting their knowledge. The resulting dataset contained 8718520 records from 89457 students.

The remaining data has been processed by the pattern detection algorithm. This algorithm continuously computes an Elo rating delta score for every sequence of attempts to decide if a particular sequence might belong to a student modelling anomaly. If the algorithm detects a combination of consecutive rises and drops with a steepness above a threshold, a pattern is discovered.

The algorithm has detected HLH patterns in the learning logs of 310 students and LHL patterns in the logs of 413 students and both types of patterns in the logs of 64 students. Figure 2 visualises a part of the Elo curve of one of those 64 students. Grey lines indicate periods of inactivity that have been cut out of the timeline. Each period of activity has its starting and ending times displayed. Overall, two distinct HLH patterns

can be observed. The first starts at 14:46 on 04/08 – a student submits a long sequence of incorrect answers and drops to the minimum Elo-rating of  $-22$ . Then, at 20:52 on the same day, she starts a climb followed by a drop at 11:23 the next day. Finally, on 06/08, in three consecutive sessions, the student climbs back to the original level. It is important to notice that, the drops exclusively consist of incorrect attempts (X-shaped dots), while climbs are sequences of only correct answers (regular dots). This is a very likely example of a student gaming the system – deliberately dropping to the easiest possible content in order to get the maximum number of “coins” as fast as possible.



**Fig. 2.** Changes in Elo-scores of a student (blue) and Elo scores of items (yellow) (Color figure online)

### 5 Summary and Future Work

This paper presents the results of the datamining experiment we have conducted on the logs of a commercial AES Math Garden. We have been able to detect a considerable number of anomalous student modelling patterns that resulted in suboptimal adaptive sequences of learning tasks. The developed algorithm has a promise to reliably recognise such sequences in the logs of Math Garden. The next step for this project is to develop a component that can detect abnormal patterns of learning activities on the fly and inform the adaptation mechanism of Math Garden that a certain sequence of attempts is likely to not reflect the actual level of students’ abilities. Such a component will need to be able to distinguish between the sequences that represent the beginning of HLH and LHL patterns and sequences corresponding to normal learning trajectories. We plan to use classification based on a number of features that can be easily extracted from students’ attempts, e.g. the time of the day, the day of the week, holiday/working day, time a student spends on an attempt, etc.

## References

1. Fogarty, J., Baker, R., Hudson, S.: Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *Proc. Grap. Interface* **2005**, 129–136 (2005)
2. Martin, B., Mitrovic, A., Koedinger, K., Mathan, S.: Evaluating and improving adaptive educational systems with learning curves. *User Model. User Adapt. Interact.* **21**(3), 249–283 (2011)
3. Sosnovsky, S., Brusilovsky, P.: Evaluation of topic-based adaptation and user modeling in QuizGuide. *User Model. User Adapt. Interact.* **25**(4), 371–424 (2015)
4. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students “game the system”. In: *Proceedings of ACM CHI 2004, Computer-Human Interaction*, pp. 383–390 (2004)
5. Wood, H., Wood, D.: Help seeking, learning, and contingent tutoring. *Comput. Educ.* **33**, 153–159 (1999)
6. Mostow, J., et al.: A La Recherche du Temps Perdu, or As time goes by: where does the time go in a reading tutor that listens? In: *Sixth International Conference on Intelligent Tutoring Systems*, pp. 320–329 (2002)
7. Woolf, B.P., et al.: Affect-aware tutors: recognizing and responding to student affect. *Int. J. Learn. Technol.* **4**(3–4), 129–164 (2009)
8. Brinkhuis, M.J.S., et al.: Learning As It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System. *PsyArXiv*, 22 March 2018. <https://doi.org/10.17605/osf.io/g4z85>
9. Hofman, A.D., Jansen, B.R.J., de Mooij, S.M.M., Stevenson, C.E., van der Maas, H.L.J.: A solution to the measurement problem in the idiographic approach using computer adaptive practicing. *J. Intell.* **6**(1), 14 (2018)
10. Pelánek, R.: Application of time decay functions and the Elo system in student modeling. In: Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds.) *Proceedings of EDM’2014*, pp. 21–27. International Educational Data Mining Society, London (2014)
11. Wauters, K., Desmet, P., Van den Noortgate, W.: Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *J. Comput. Assist. Learn.* **26**(6), 549–562 (2010)
12. Elo, A.: *The Rating of Chess Players, Past and Present*. B. T. Batsford Ltd, London (1978)
13. Klinkenberg, S., Straatemeier, M., van der Maas, H.L.J.: Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educat.* **57**(2), 1813–1824 (2011)



# Expanding the Curricular Space with Educational Robotics: A Creative Course on Road Safety

Andri Ioannou<sup>(✉)</sup>, Chrysanthos Socratous, and Elena Nikolaedou

Cyprus Interaction Lab, Cyprus University of Technology, 3036 Limassol, Cyprus  
andri.i.ioannou@cut.ac.cy

**Abstract.** While initiatives worldwide continue to place pressure on schools to improve STEM education, the already overcrowded curriculum often leaves little space for the integration of new courses or topics. Numerous benefits are reported in the literature about the use of educational robotics; yet, their integration in school contexts requires time that cannot be taken from other important courses. In the end, most educational robotics activities are done outside the curriculum such as in after-school programs and summer camps. The major contribution of this work is the presentation of a case of creative and non-intrusive integration of educational robotics to support the current school curricula. We present an example of expanding the curricular space, by integrating educational robotics in an existing course unit. In the absence of formal educational robotics curriculum and courses, the study presents an exemplar case of educational robotics integration in a creative and non-intrusive way. The lesson design and implementation are presented; the creative infusion can be realized and holds benefits for the students. Through educational robotics, students can practice new skills such as problem solving and teamwork, while they gain knowledge in the specific domain of the course unit.

**Keywords:** Educational robotics · Technology integration · K-12  
Expanding the curricular space · Bee-Bot

## 1 Introduction

In the recent years, there has been an increased interest in the educational use of robotics with numerous attempts to integrate the technology from kindergarten to university level worldwide [1]. Within a (social) constructivism spirit, the use of educational technology aims to enable students to engage in problem-solving, collaborative learning and creative thinking; educational robotics is considered one such technology, whether it is used to teach specific content in a domain such as engineering or is designed to work as a construction and programming tool for promoting problem solving and computational thinking [2]. Today, educational robotics is seen as an innovative, progressive and versatile educational tool for teaching and learning, that is also fascinating for students of all ages [3]. Several authors have reported learning gains as a result of student engagement with various robotics projects, including the development of problem solving, creativity and collaboration skills [1].

Overall, educational robotics has emerged as a unique educational tool that can provide hands-on activities in an attractive learning environment, boosting students' interest and curiosity [4, 5]. Yet, despite the great interest developed around this topic, formal educational robotics curricula and courses are currently lacking in K-12 schools around the world. Simply, the overcrowded K-12 curriculum leaves little time for dedicated courses or units. Therefore, most educational robotics activities are done outside the curriculum such as in after-school programs and summer camps [6]. The present study aims to investigate how educational robotics can be integrated in existing school subjects, in a creative and non-intrusive way, therefore expanding the curricular space (i.e., learning about robotics and computational thinking while learning language). That is, the authors sought to present a case of technology integration which expands the curricular space in that it allows students to practice skills such problem solving and teamwork while they work on a subject of the school curriculum. The overarching research question of the study was:

*RQ: How educational robotics may be realized as means for expanding the curricular space via their creative integration in current school subjects?*

## 2 Background Work

Literature reveals that educational robotics is a growing sector with the potential to significantly affect the nature of science and technology education (i.e., STEM education), from kindergarten to tertiary education [1, 7]. There are a number of reports resulting from various educational robotics programs about educational robotics improving the performance of students in mathematics, physics and engineering [1]. Moreover, researchers [8] have found that students who attended robotics courses developed powerful logic and critical thinking skills, oral presentation and teamwork skills. When dealing with robotics, students are stimulated to identify the problem, to analyze and explore possible solutions to achieve the objective, and to check their solution with the appropriate control procedures e.g., evaluating the solution in terms of functionality [6]. In general, the role of educational robotics should be considered broadly, as a tool that can support the development of a variety of skills, including cognitive skills, personal development, and collaboration skills. Researchers have argued that educational robotics offer special educational advantages, because the technology is interdisciplinary in nature and includes a synthesis of many technical issues, including algebra and trigonometry, design and innovation, electronics and programming, the forces and the laws of motion, as well as other materials and hands-on processes [9]. It is for this reason – the interdisciplinary nature of the technology – that the present investigation considers ways to integrate robotics in the existing school curricula, as opposed to suggesting dedicated educational robotics courses and curricula.

Research on educational robotics is typical seen through the lens of constructionism [10]. Constructionism argues that learning occurs when the student creates a physical structure that reflects the experience of solving problems, relying on incentive received from the construction of the object itself [10]. Generally speaking learning that is driven by problems (problem based learning) allows the learner to build his/her own

knowledge. In this spirit, educational robots are essentially a constructionist tool, with which students interact and utilize their knowledge and experience to solve real problems by developing and testing their solutions [11, 12]. A typical lesson plan in educational robotics includes an initial introduction to programming the robot, followed by student practice on applying their knowledge to make the robot work [13].

One of the main weaknesses in the area of educational robotics is the absence of clearly defined curricula, educative material for teachers and learners, as well as a repository of available kits and their capabilities [2]. What's more, educational robotics is most often seen as an extra-curricular activity and as part of informal education. Efforts should be made to design educative material for educational robotics linked to existing school curricula and taking advantage of the capabilities of available (and affordable) educational robotics kits. With no doubt, teacher professional development on the integration of educational robotics is imperative; teachers need to see educational robotics as a teaching tool to enhance the learning process, complement the learning experience, and provide incentives for students, while the role of the teacher remains of great importance in supporting and scaffolding the learning experience [13].

Overall, while initiatives worldwide continue to place pressure on schools to improve STEM education, the already overcrowded curriculum leaves little space for the integration of new courses or topics [14], such as that of educational robotics. The present study aimed to investigate how educational robotics can be integrated in existing school subjects, in a creative and non-intrusive way, therefore expanding the curricular space. Similar initiatives have been previously considered by others. Although not in the area of educational robotics, the GlobalEd 2 project also builds on the idea of expanding the curricular space; it builds upon the interdisciplinary nature of the social studies course in the schools of USA and integrates technology (i.e., a simulation web-based environment) aimed at increasing the instructional time devoted to science and persuasive writing [14]. On the other hand, from an interest and gender differences point of view, researchers [15] have suggested that educational robotics should be integrated into the curriculum of subject areas such as art, music and literature, to meet the interests of a diverse population of students. The authors' [15] argument was based on the premise that meeting students' personal interests allows them to persist more when they encounter problems and to continue to expand their exploration to new directions. For example, in one of their studies, the topic was the park; students had to remember their experiences at a park (e.g. seeing a dog chasing a cat) and they had to use educational robots to represent their experience. By incorporating the history narrative (storytelling) in the robotics activity, the students improved their reading and writing skills [15].

### 3 Method

This work aimed to present a case of creative and non-intrusive integration of educational robotics in the overcrowded school curricula. Student questionnaires were administered to understand the experience from the students' point of view, whilst teacher interviews were conducted to understand the experience from the teachers' perspective, strengthening the evidence and enhancing the validity of the findings.



### 3.1 Participants

Participants were 43 students and 3 educators, coming from one private and two public schools in Northeastern Europe. Specifically 10 second-graders (3 girls and 7 boys) with their (female) teacher came from a private school, and 33 second-graders (15 girls and 18 boys) and their two teachers (females) came from two classes in a public school. None of the participants (teachers and students) had previous experiences with educational robotics.

### 3.2 Procedures

Teaching “road safety” is part of the country’s teaching requirements, found as a unit in the subject of “general citizenship and wellbeing”. In this study, “road safety” was addressed (and is typically addressed) during the first two weeks of October, both in the public and private schools.

We used Bee-Bots, a commercial programmable floor robot kit. Based on BeeBot.us, the robot’s friendly layout appeals to children and can be a starting point for teaching control, directional language and programming to young children.

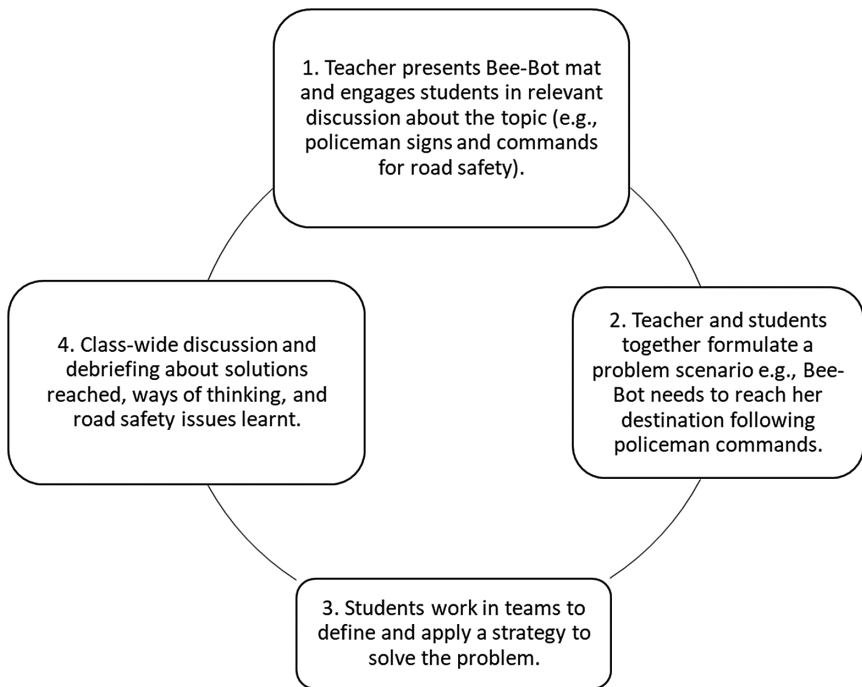
There was a preparation phase, during which the researchers and teacher of the private school worked closely together to co-design lesson plans to integrate Bee-Bot in the “road safety” unit. Several lesson plans were designed using freely available Bee-Bot mats and other images (e.g., policeman, stop-sign, pedestrian cross) located online such as at <http://www.twinkl.co.uk/>, printed in A4, plasticized, and assembled on the classroom floor for group activities (i.e., one mat for each group).

The first lesson used a testing mat (see Fig. 1) and aimed to familiarize students with Bee-Bot by practicing with the following: Bee-Bot tabs, directional commands, termination; decoding tabs to understand the resulting movement and verifying the answer. This lesson was completed as a class-wide experience in front on a single mat and using one Bee-Bot.

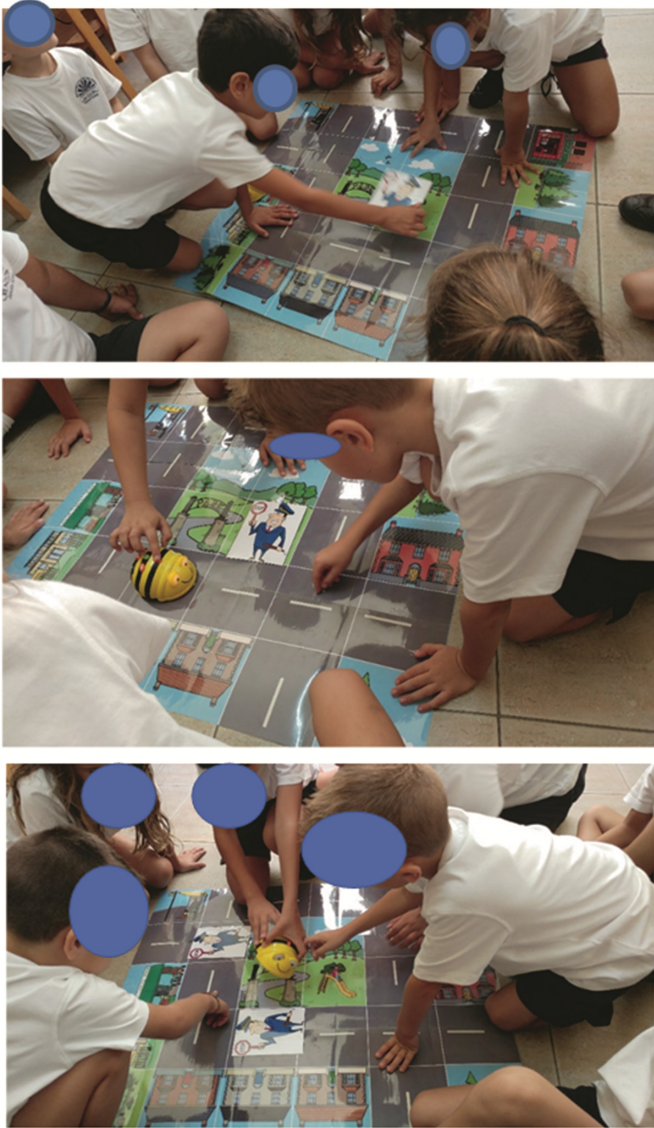


**Fig. 1.** First lesson with Bee-Bot and testing mat.

Following the first lesson, all Bee-Bot activities focused on “road safety”. Students worked in groups of 3–5 students, in front of a mat. Typically, the teacher together with the students formulated a problem scenario e.g., Bee-Bot needs to reach a hotel, following policeman commands such as stop signs placed in various places. Setting the starting point and ending point, including direction of the Bee-Bot on the mat, was part of the problem definition. Figure 2 illustrates the typical structure of a lesson plan. Lesson plans became progressively more difficult (i) in terms of problem solving e.g., Bee-Bot had to follow a more complicated path in reaching the ending point via obstacles and only a single trial was allowed for the team, and (ii) in terms of knowledge about “road safety” i.e., Bee-Bot had to understand commands by the policeman such as a stop sign or diversion and had to consider pedestrian crosses. Figure 3 presents some episodes from the school implementation. In sum, the curricular space was expanded in that the lessons targeted problem solving (e.g., computational thinking) and teamwork skills together with knowledge about “road safety”.



**Fig. 2.** Typical structure of a lesson plan.



**Fig. 3.** Lesson plan implementation with 2nd graders

### 3.3 Data Collection

For the duration of two weeks in October 2016, the researchers observed 10–11 teaching sessions (45 min each session) implementing the series of lesson plans in each participating school. All participants signed informed consents and were aware of the roles of the researcher-observers in the field. At the end of all lessons, a 30-min semi-structured interview was conducted with each of the participating teachers with the double scope

of understanding: (i) how the experience was good (or not), and (ii) what can we learn about the integration of educational robots in existing school topics and curriculum. Moreover, at the end of the experience, the participating students (N = 43) completed a 7-item attitudinal questionnaire regarding their overall experience (see Fig. 4).

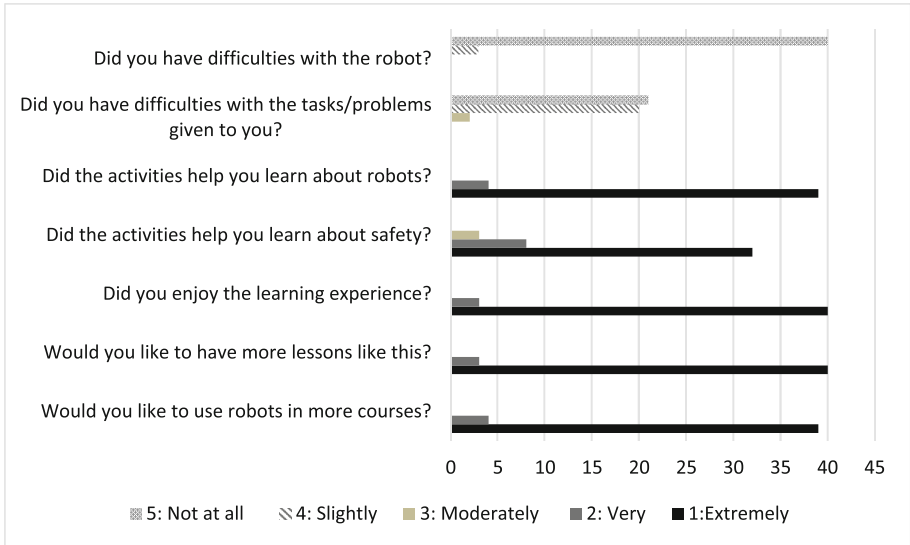


Fig. 4. Students’ perceived experiences (N = 43)

## 4 Analysis and Results

All students (N = 43) completed the questionnaire supporting our understanding of the experience from the students’ point of view. The teachers’ perspective (via interview data) was analyzed thematically to extend our understanding of the experience.

### 4.1 Students’ Perceived Experiences

All students (N = 43) completed the questionnaire. Results demonstrated that the integration of educational robotics in the existing curriculum was fully endorsed by the participants. As illustrated in Fig. 4, all students thought that the lessons were enjoyable whilst they allowed learning about robots as well as “road safety”.

### 4.2 Teachers’ Perceptions

Teachers’ semi-structured interview data were transcribed and analyzed. A thematic analysis was conducted by two researchers, working closely together to identify core consistencies and meanings (themes) in the pool of qualitative data (Patton, 2014). In general, no variations were noted in the three teachers’ perceived experiences across

public (2 classrooms) and private schools. We report on these themes next, organized within the double scope of the interview.

*How was the experience good or not?*

All three teachers fully endorsed the integration of Bee-Bots in the school lesson on “road safety” as well as the overall idea of using educational robots to expand the curricular space to address skills (e.g., problem solving) beyond knowledge on the matter. The teachers deemed this approach of technology integration as *non-intrusive with valuable learning benefits* (theme 1). As one of the teachers argued:

“The students were not destructed by the playfulness of the Bee-Bot, but rather they exhibited learning gains from this experience as they discussed about “road safety”, namely stop signs and pedestrian crosses, problem-solved with their Bee-Bot, evaluated and improved their solutions, and explained their thinking to others during our class-wide discussions.”

Moreover, the teachers endorsed the *emerging gameful character* (theme 2) of the overall experience, which was considered valuable for collaborative learning and teamwork. As the teachers explained, students, within their team, engaged in collaborative learning and problem-solving targeting a common goal; collaboration was better and better as lessons progressed. Between teams, competition dynamics emerged naturally and were enriched by the teacher’s praise and rewards for good problem-solving and collaboration. The gamefulness of the activity was further promoted by social rewards or peer pressure by teammates of the owned group or the competing group. The benefits of gameful learning have already been discussed in [16], consistent with the teachers’ perspectives in the present study.

Furthermore, the learning experience was perceived as *engaging and embodied* (theme 3). Students were present, mentally and physically and while planning their strategy, they often used their bodies to support their thinking. For example, students stood up and performed the steps and turns on the mat, before enabling the Bee-Bot (especially when only one trial was allowed), or after they realized an unexpected outcome (i.e., to help decoding the error). In the teacher’s own words:

“The activity engaged their bodies and minds and motivated participation even from the quietest students. They often stood up and ‘tested’ the steps using their bodies to support their thinking [...] They enthusiastically planned their Bee-Bot path solution which involved domain knowledge, for example, Bee-Bot as vehicle stops at stop signs, and they had a lot of fun seeing the result of their planning. And if the solution was not correct, they decoded their solution to understand what went wrong, which helped them practice their problem-solving skills.”

*What can we learn about the integration of educational robots in existing school topics and curriculum?*

Not surprisingly, the teachers explained that *careful planning and access to resources* (theme 4), such as lesson plans and mats for the robot, are needed for successful integration of education robots. As they noted, understanding the functionality of the technology is imperative, but a good knowledge of the daily curriculum and school topics is also required, before the educator can think of effective learning activities around educational robotics. That is, the curriculum goals need to be fully addressed, whilst additional opportunities for learning are mediated by educational robotics, such as the development of problem solving and computational thinking skills.

According to the teachers, this planning might take quite longer than typical lesson preparation and might not be something a novice teacher would undertake unless s/he has support and access to relevant, open educational resources. In the teacher's own words:

“You need to make sure the objectives of the curriculum are met and that the robotic activities will not drift attention away from these objectives, but rather, will add to it, by enabling additional types of skills such problem solving. This planning is not easy to do before you are well familiar with the daily curriculum and school topics and you also understand the technology [...]. Open educational resources can help a lot; for example, although I can now think of amazing lessons plans for the upcoming units, I don't have the skills to design the Bee-Bot mats.”

Nevertheless, given good preparation took place in this study, all teachers agreed that that series of lessons were successful in meeting the curriculum goals on “road safety” as well as expanding students' opportunities to engage in problem solving and teamwork. Moreover, they stated how, upon this experience, they could already think of numerous lessons for expanding the math and language curricula using Bee-Bot, such as for example, using a mat with shapes, numbers, and symbols for addition, subtraction, multiplication, and division in math.

In terms of implementation, after the first lesson, no guidance was needed in using the robot. Teamwork was a challenge only in the first couple of lessons during which, students exhibited lack of cooperation (e.g., all wanted to handle the Bee-Bot). Perhaps this was due to the enthusiasm caused by the novelty of the task; *teamwork and collaboration around the robot got better as the lessons progressed* (theme 5) and as students realized that group work had value into getting the task completed successfully. Students learnt to divide responsibilities within the group using a rotation pattern; this practice was realised in all three classes, after the first couple of lessons, and it mostly the students' owned initiative (e.g., deciding and assigning roles), as initial evidence of teamwork skills being developed. All three educators agreed that small groups (3–5 students) worked well, which is consistent with previous practice and findings in educational robotics [3].

## 5 Discussion and Conclusions

While initiatives worldwide continue to place pressure on schools to improve STEM education, the already overcrowded curriculum often leaves little space for the integration of new courses or topics. Numerous benefits are reported in the literature about the use of educational robotics; yet, their integration in school contexts requires time that cannot be taken from other important courses. In the end, most educational robotics activities are done outside the curriculum such as in after-school programs and summer camps [6]. Yet, the increasing availability of educational robotics kits and the growing interest in their use by researchers and practitioners, presents an opportunity to examine issues of technology integration in creative and non-intrusive ways. The present study aimed to investigate how educational robotics can be integrated in an existing school subject, expanding the curricular space by allowing the development of robotics, problem solving and teamwork skills together with domain knowledge.

Findings from this study support the researchers' standpoint about the value of using educational robotics to expand the curricular space. The study presents evidence that this approach is non-intrusive, but rather engaging [8], embodied [12], and gameful [16] with valuable learning benefits around problem solving and teamwork. Although, the infusion of educational robotics requires some extra preparation on behalf of the teacher, the benefits seem to be rewarding. With careful design, a cognitive bridge is created between curriculum objectives and the educational robotics experiences, encouraging students to acquire content knowledge in addition to other types of skills [6]. These findings, although preliminary, are strengthened in terms of consistency across data sources (i.e., student questionnaires and teacher interviews) and school settings (i.e., implementation in one private and one public school with three educators involved).

Overall, the study demonstrated a case of creative and non-intrusive infusion of educational robotics in the existing curricula, in the absence of time for dedicated educational robotics courses. The approach was deemed appropriate and beneficial for students, showcasing educational robotics as means for expanding the curricular space to allow for the development of robotics, problem solving (e.g., computational thinking) and teamwork skills together with domain knowledge. We understand that our data are preliminary and rely on self-reported measures and observations; yet, a major contribution of this work is the realization of a method toward creative and non-intrusive integration of educational robotics in the overcrowded school curricula. Future work should aim to objectively document student learning gains in an expanded curricular space using educational robotics. Future work should also aim to create open educational resources relevant to this idea via the infusion of educational robotics in existing school curricula and topics. Furthermore, while enabling teachers to develop and teach educational robotics as a core curriculum course might be an important goal [17], we would further argue that teacher professional development should present successful examples and provide inspiration for the creative integration of educational robotics in the existing school curricula. We hope that the encouraging findings of this work will motivate further research and practice in this area.

**Acknowledgement.** Authors acknowledge funding from the European Union's Horizon 2020 Framework Programme through the NOTRE Project (H2020-TWINN-2015, Grant Agreement Number: 692058).

## References

1. Benitti, F.B.V.: Exploring the educational potential of robotics in schools: a systematic review. *Comput. Educ.* **58**(3), 978–988 (2012)
2. Ioannou, A., Makridou, E.: Exploring the potentials of educational robotics in the development of computational thinking: a summary of current research and practical proposal for future work. *Educ. Inf. Technol.* (2018). <https://doi.org/10.1007/s10639-018-9729-z>
3. Atmatzidou, S., Demetriadis, S.N.: Evaluating the role of collaboration scripts as group guiding tools in activities of educational robotics: Conclusions from three case studies. In: 2012 IEEE 12th International Conference on Advanced Learning Technologies, pp. 298–302. IEEE, July 2012

4. Alimisis, D.: Educational robotics: Open questions and new challenges. *Themes Sci. Technol. Educ.* **6**(1), 63–71 (2013)
5. Eguchi, A.: Educational robotics for promoting 21st century skills. *J. Autom. Mob. Robot. Intell. Syst.* **8**(1), 5–11 (2014)
6. Bilotta, E., Gabriele, L., Servidio, R., Tavernise, A.: Edutainment robotics as learning tool. In: Pan, Z., Cheok, A.D., Müller, W., Chang, M. (eds.) *Transactions on Edutainment III*. LNCS, vol. 5940, pp. 25–35. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-11245-4\\_3](https://doi.org/10.1007/978-3-642-11245-4_3)
7. Spolaôr, N., Benitti, F.B.V.: Robotics applications grounded in learning theories on tertiary education: a systematic review. *Comput. Educ.* **112**, 97–107 (2017)
8. Mosley, P., Kline, R.: Engaging students: a framework using LEGO® robotics to teach problem solving. *Inf. Technol. Learn. Perform. J.* **24**(1), 39–45 (2006)
9. Petre, M., Price, B.: Using robotics to motivate ‘back door’ learning. *Educ. Inf. Technol.* **9**(2), 147–158 (2004)
10. Papert, S., Harel, I.: Situating constructionism. *Constructionism* **36**(2), 1–11 (1991)
11. Danahy, E., Wang, E., Brockman, J., Carberry, A., Shapiro, B., Rogers, C.B.: Lego-based robotics in higher education: 15 years of student creativity. *Int. J. Adv. Rob. Syst.* **11**(2), 27 (2014)
12. Socratous, C., Ioannou, A.: A study of collaborative knowledge construction in stem via educational robotics. In: Kay, J., Luckin, R. (eds.) *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 2008 13th International Conference of the Learning Sciences (ICLS)*, vol. 1, pp. 496–503. ISLS, London, UK (2018)
13. Mubin, O., Stevens, C.J., Shahid, S., Al Mahmud, A., Dong, J.J.: A review of the applicability of robots in education. *J. Technol. Educ. Learn.* **1**, 0015–209 (2013)
14. Lawless, K.A., Brown, S.W., Brodowinska, K., Field, K., Lynn, L., Riel, J., Le-Gervais, L., Dye, C., Alanazi, R.: Expanding the science and literacy curricular space: the GlobalEd 2 project. In: Paper presented at Eastern Educational Research Association Annual Meeting, Jacksonville, FL, February 2014
15. Rusk, N., Resnick, M., Berg, R., Pezalla-Granlund, M.: New pathways into robotics: Strategies for broadening participation. *J. Sci. Educ. Technol.* **17**(1), 59–69 (2008)
16. Ioannou, A.: A model of gameful design for learning using interactive tabletops: Enactment and evaluation in the socio-emotional education classroom. *Educ. Technol. Res. Develop.* (2018). <https://doi.org/10.1007/s11423-018-9610-1>
17. Hamner, E., Cross, J.: Arts & Bots: Techniques for distributing a STEAM robotics program through K-12 classrooms. In: *Proceedings of the Third IEEE Integrated STEM Education Conference*, March 2013



## **Poster and Demo Papers**



# New Approaches to Training of Power Substation Operators Based on Interactive Virtual Reality

Rinat R. Nasyrov<sup>1(✉)</sup> and Peter S. Excell<sup>2</sup>

<sup>1</sup> National Research University Moscow Power Engineering Institute, Moscow, Russia  
nasyrov.rinat@gmail.com

<sup>2</sup> Wrexham Glyndwr University, Wrexham, UK  
p.excell@glyndwr.ac.uk

**Abstract.** A novel approach to the training of power substation operators based on virtual reality is presented. The VR is extended by incorporation of a range of options for interactivity which permit the trainee to take actions in the simulated substation, including incorrect actions, with realistic consequences simulated. Any real substation may be simulated visually and functionally in the virtual environment. The system enables substation operators to gain realistic operational experience without the anxieties of causing blackouts and damage in a real grid.

**Keywords:** Substation · Operators · Learning · Training · Virtual reality  
Interactivity

## 1 Introduction

Virtual reality (VR) technologies routinely allow (for example) pilots, drivers and astronauts to experience realistic training, without risk of injury or equipment damage. On the other hand, it has to be recognized that the majority of uses of VR are for entertainment. However, lessons learned from entertainment systems could be useful for new non-entertainment applications: the present work reports development of a VR training system for power substation operators, who have to be aware of correct reactions to a wide range of both routine and fault or emergency situations.

Major issues that have to be handled include: (1) health and safety of substation operators, (2) operational integrity of substation equipment, (3) interruption of customers' power supply as consequences of substation operators' errors during switching. In addition, the Russian grid statistics show that 30–35% of faults and blackouts are due to errors during switching [1] and this indicates a need for substantial improvements in the training of substation operators.

The range of possible faults and of operator actions mean that VR training of substation operators could be a very beneficial and innovative technology.

## 2 Strengths and Shortcomings of Currently Used Simulators

There are many simulators currently used for training substation operators. They have some strengths, e.g.:

- Control of simulated operation in routine and non-routine events.
- Estimation and recording of the substation operators' decisions made during routine and non-routine events.
- Rapid evaluation of parameters of steady-state mode of network models.

At the same time, currently used simulators have shortcomings, such as:

- Two-dimensional (2D) display of the main control room and substation equipment does not impart realistic scenarios to develop correct skills in substation operators.
- Routine normal operations (verification actions, lockout-tagout, collective protection) tend to be perfunctory and without deep insight for the trainees.
- Navigation around the substation is not a realistic representation of real time navigation.

Thus, it can be seen that the currently used simulators are imparting only a limited set of skills to substation operators. A VR simulator for substation operators has the potential to be vastly more meaningful and hence was developed to rectify the deficiencies of currently used simulators.

### **3 New Approaches to Training Substation Operators, Based on Virtual Reality**

There are two main technologies of VR implementation: 3D CAVE [2] and 3D Helmet [3]. Both of them can be used for a simulator and both provide complete immersion in VR. The selection of the technology of VR implementation has to be made alongside consideration of the conditions of real use. If there is sufficient space both the CAVE technology and 3D Helmet may be used. However if the space is limited, only 3D Helmet technology is viable.

#### **3.1 3D Model of Substation**

Regardless of the chosen technology of VR implementation it is necessary to create a 3D model of the substation. The model must meet requirements such as: visual similarity, spatial analogues and functional similarity of the substation. Models should be implemented for a range of substations in order to justify the cost of the system.

The visual similarity of a substation 3D model leads the user (operator) to develop relevant skills of operation on a particular substation during the training. After finishing the training on a certain substation the operator does not need to adapt to the real substation because he or she has already interacted with it in VR. Spatial analogue representations of objects in a substation 3D model are also very important. Operators must know how long it takes to get from one point of the substation to another one. This could help in emergency cases when the speed of action is a decisive factor. The reactions to the user's actions in the VR model must be the same as on the real substation. Otherwise, operational experience gained in VR will be of little use.

### 3.2 Scenarios of Training

The operation order of routine switching is highly critical. A training simulator must control the order of routine switching, comparing the user's actions to the correct order of switching. The correct order of switching is contained in the programmed scenarios of training. The scenarios of training must include both regular and emergency cases. Scenarios of regular operations act to hone the skill of routine switching and understanding of the basic operation of the substation. Scenarios of emergency training start in the same way as regular training but suddenly an emergency appears. The user has to take correct decisions as fast as possible after the emergency moment [4]. The user can choose between training and test mode. If he or she chooses a training mode, they can select the type of scenario: emergency or regular training in selected specific scenarios. In test mode, the user does not know what type of training scenario he or she will encounter: it will be a random choice of the simulator. System estimate consist of comparison of right steps and user done steps. If user made some number of uncritical or one critical mistake it is not allowed to him work on real substation.

## 4 Implementation of Virtual Reality Simulator

In the Power Electrical Systems Department of MPEI (Russia) a 3D VR simulator for substation operators has been developed. The department receives ongoing research funding from the PJSC Interregional Distribution Grid Company of Central Russia to develop the simulator. As a prototype of the simulated substation a typical 110/35/10 kV substation in Russia was chosen. It contains: six 110 kV overhead line connections; six 35 kV overhead lines and one 35 kV cable line connection; thirty 10 kV cable line connections; outdoor switchgear for 110 kV; outdoor switchgear for 35 kV; indoor switchgear for 10 kV; three three-phase transformers. The area of the prototype is about 18000 m<sup>2</sup>. As can be appreciated, the prototype contains many kinds of equipment and hence many scenarios can be implemented on it.

As mentioned before, the choice of technology depends on the conditions of use. In the present case, there were significant constraints on space: to alleviate this, it was decided to include an "Omniroad" treadmill device in the project. Consequently, the 3D Helmet was chosen as the VR technology.

Creation of training scenarios is the last step of implementation of the virtual reality. As regular scenarios were chosen about taking out of/into service 110/35/10 kV: transformers, circuit-breakers, disconnectors, overhead and cable lines.

As emergency taking scenarios were chosen regular scenarios above with sudden powering off of transformer, emergency collapse of column insulation of disconnector, emergency SF<sub>6</sub>-gas depressurizing, emergency line-to-earth fault of 35 kV overhead line, emergency voltage transformer fuse failure, emergency current transformer explosion. These scenarios cover about 90% of the operations on the substation.

## 5 Brief Technical Description

The game engine part of the simulator is based on Unreal Engine 4. A mathematical model of a significant prototype has been written in the C language: this enabled the latency to be reduced to less than 0.01 s, which is essential for a convincing experience. The headset of the simulator uses an umbilical cable to link to the host computer because currently there is no single technology of wireless link that can transmit 2 channels of HD plus 3 channels of gyroscopic information and 1 usd-3 channel with acceptable latency. The mathematical modelling of the prototype and the 3D modelling took approximately 1200 person-hours in total, for six persons: one modeller (mesh creator), two programmers (mathematical model creators), one unifying programmer (connecting mathematical and 3D models) and two electrical power engineers.

## 6 Results

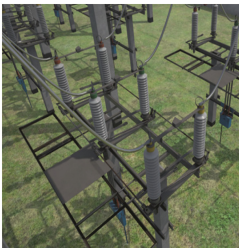
A mathematical model of a significant prototype has been completed with protection and automation. A 3D model of a 10 kV indoor switchgear group has been created (Fig. 1). Models of an outdoor 110 kV SF6 circuit-breaker and a disconnector are shown in Figs. 2 and 3 respectively. The interaction of operators with the 3D model is shown in Fig. 4. The system has been tested with a number of trainee operators and results have been successful, in the sense of perceived realism and relevance of the training. Work is now proceeding to implement other switchgear environments.



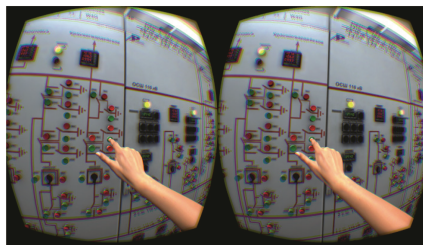
**Fig. 1.** 3D model of 10 kV indoor switchgear group.



**Fig. 2.** 3D model of a 110 kV SF6 circuit-breaker.



**Fig. 3.** 3D model of a 110 kV disconnector.



**Fig. 4.** Trainee operator interacting with a 3D model.

## References

1. Russkih, A.A., Nasyrov, R.R.: Electrical Safety System. PCS, Moscow (2015). ISBN 978-5-905485-81-7
2. Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V., Hart, J.C.: The CAVE: audio visual experience automatic virtual environment. *Commun. ACM* **35**(6), 64–72 (1992). <https://doi.org/10.1145/129888.129892>
3. Thompson, J.I.: A three dimensional helmet mounted primary flight reference for paratroopers. Thesis AFIT/GCS/ENG/05-18. United States Department of The Air Force, Air Force Institute of Technology, Dayton, Ohio (2005)
4. Nasyrov, R.R., Suleimanov, I.R., Churkin, A.I., Pilyugin, A.V., Marchenkov, D.V.: Switching training simulator based on virtual reality. *Elektrichestvo (Electricity)* **3**, 27–32 (2016)



# Enabling Systematic Adoption of Learning Analytics through a Policy Framework

Yi-Shan Tsai<sup>1</sup>(✉) , Maren Scheffel<sup>2</sup> , and Dragan Gašević<sup>1,3</sup> 

<sup>1</sup> University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK  
yi-shan.tsai@ed.ac.uk

<sup>2</sup> Open Universiteit Nederland, Valkenburgerweg 177, 6419 AT Heerlen, Netherlands  
maren.scheffel@ou.nl

<sup>3</sup> Monash University, 19 Ancora Imparo Way, Clayton 3800, Australia  
dragan.gasevic@monash.edu

**Abstract.** Learning analytics (LA) has shown a great potential in improving learning experience and enhancing pedagogical effectiveness. However, the adoption of LA in higher education involves various social, cultural, and technical issues that need to be addressed strategically. We present a study that aims to assist with the development of institutional LA policies to ensure effective and legitimate adoption of LA. The study takes an action research approach and involves key stakeholders directly, so as to incorporate a wide range of perspectives in the policy formation. Ethics and privacy issues were considered the priority element in a LA policy and the top concern for students. A sense of uncertainty about the returns in investment was observed among senior managers, whereas teaching staff were mostly worried about time pressure and the potential of LA to be used for performance appraisal. This poster presents a policy framework that can be used to support institutional readiness assessment, strategy formation, and policy development.

**Keywords:** Policy · Learning analytics · Higher education

## 1 Introduction

The field of learning analytics (LA), with its associated methods of online student data analysis, is able to provide novel and real-time approaches to assessing critical issues such as student progression and retention, thereby informing decisions related to teaching and learning. While LA has gained much attention and has been/is being adopted by many higher education institutions (HEIs) in Europe and other parts of the world, the maturity levels of HEIs in terms of being ‘student data informed’ are only in the early stages. Literature has identified that the adoption of LA in complex educational systems requires a systematic approach to bring about effective changes [2]. Moreover, some common challenges that beset the adoption at a wide scale need to be addressed by involving all

relevant stakeholders [3]. Our research project sets out to tackle the identified problems by building a policy framework that is based on findings of various consultations with a diversity of stakeholders. The study aims to answer four questions: (1) what is the state of the art in terms of LA adoption in Europe, (2) what are the key challenges that impede institutional adoption of LA, (3) how do expectations of LA vary among different stakeholders, and (4) how can we address LA related actions and challenges through policies.

The goal of the study is to incorporate existing experiences of institutional adoption with key stakeholders' perspectives regarding opportunities for LA and concerns about it, thereby developing a policy framework to support effective and responsible adoption at an institutional scale.

## 2 Methods

The policy framework is developed using mixed methods. Between 2016 and 2017, various datasets have been collected through online group concept mapping (GCM), interviews, surveys, and focus groups. With the online GCM, we have collected 99 statements from 29 LA experts across the world. With the interview method, we had in-depth conversations with 64 institutional leaders from 51 HEIs across Europe. With the survey method, we have reached out to institutional leaders from 46 European HEIs, 3,263 students from six European HEIs and 210 teaching staff from four European HEIs. With focus groups, we have carried out in-depth conversations with 74 students and 59 teaching staff from four European HEIs. The development of protocols for the above mentioned activities were driven by the research questions listed above. The methods used for analysis include cluster analysis, exploratory factor analysis, confirmatory factor analysis, and thematic content analysis. The development of the policy framework was inspired and guided by the Rapid Outcome Mapping Approach (ROMA) [1,2,5] model that begins by defining an overarching policy objective, followed by six steps designed to provide policy makers with context-based information: (1) map political context, (2) identify key stakeholders, (3) identify desired behaviour changes, (4) develop engagement strategy, (5) analyse internal capacity to effect change, and (6) establish monitoring and learning frameworks.

## 3 Results

### 3.1 Essential Features of a LA Policy

The group concept mapping activity received 99 statements in response to the prompt "an essential feature of a higher education institution's LA policy should be...". Six key themes emerged from these statements including (1) privacy & transparency, (2) roles & responsibilities (of all stakeholders), (3) objectives of LA (learner and teacher support), (4) risks & challenges, (5) data management, and (6) research & data analysis. The rating results of these statements show an obvious drop of ratings for the ease of implementation level of these



themes, compared to their importance level. One of the implications is that the six features could potentially be challenges to deal with in the policy development process. Moreover, issues around privacy and transparency were considered the most important elements, but also the easiest to include in a policy.

### 3.2 State of Adoption – Senior Managers’ Perspective

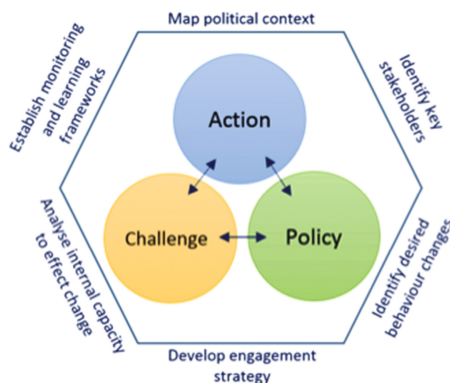
The interview data showed that 21 out of 51 institutions were already implementing centrally-supported LA projects, 9 of which had reached institution-wide level, 7 partial-level (including pilot projects), and 5 were at the data exploration and cleaning stage. Meanwhile, 18 institutions were in preparation to roll out institutional LA projects, and 12 did not have any concrete plans for an institutional LA project yet. The survey data revealed that 15 institutions had implemented LA, of which 2 had reached full implementation and 13 were in small scale testing phases. Sixteen institutions were in preparation for LA projects, and 15 were interested but had no concrete plans yet. One of the implications of the two data sets is that there was high interest in LA among HEIs in Europe, but the maturity of adoption was low.

From the survey, we identified that five top drivers for institutions to adopt LA were to improve student learning performance and satisfaction, teaching excellence, student retention, and to explore what LA can do for the institution/staff/students. These drivers were also repeatedly mentioned by the interview participants. In particular, for those who were driven by the fifth reason, their adoption was predominately experimental and exploratory. As a result, there was a sense of uncertainty about the return of investment in these institutions given that the contextual relevance and benefits of LA were still unclear.

### 3.3 Interests and Concern – Perspectives of Students and Staff

The result of the student survey that compared ideal and realistic expectations of LA identified two factors: ethical expectations and service expectations. Students held strong beliefs toward the university securely holding all collected data, whilst the belief that a university should seek consent before the collection, use, and analysis of educational data appeared to elicit the lowest average response for each sample of students. Moreover, students appeared to show strong interest in receiving regular updates on their learning, but low interest in receiving early interventions if LA services found them to be at-risk. The result suggests a student preference over a LA service that facilitates independent learning rather than one which would impede their self-direction.

Consultations with students and teaching staff through focus groups [4] revealed a strong interest in using LA to enable personalised support and provide an overview of learning progress, so as to improve pedagogical effectiveness and learning experience and success. Despite their interest in LA, both students and teaching staff expressed various concerns about adopting LA. Among these, ethical and privacy issues, such as access, security and anonymity, appeared to be



**Fig. 1.** The policy framework structure

the top concerns for students. As for teaching staff, time pressure and potential use of LA in judging teaching performance particularly concerned them.

## 4 Conclusion

This research project has reached out to nearly half of the European countries, and observed high interest in LA among HEIs. However, few HEIs have taken a systematic approach to LA with defined strategy and policy. Our preliminary findings have identified prominent challenges that need to be tackled through an overarching policy. Up to now, the research team has developed the first draft of a policy framework primarily based on the interview data. This policy framework maps out 51 HEIs' experience to the six dimensions of the ROMA model and presents key actions to take towards systematic adoption of LA, key challenges to address in the adoption process, and key questions to answer when developing an institutional learning analytics policy (Fig. 1). The policy framework will be updated with findings from other datasets and connected to detailed case studies as a reference model and can then be used to guide the development of institutional policies and strategic planning for learning analytics.

## References

1. Ferguson, R., et al.: Research evidence on the use of learning analytics - implications for education policy. JRC Science for Policy Report (2016)
2. Macfadyen, L., Dawson, S., Pardo, A., Gašević, D.: Embracing Big Data in Complex Educational Systems: The Learning Analytics Imperative and the Policy Challenge. *Res. Pract. Assess.* **9**, 17–28 (2014)
3. Tsai, Y.S., Gašević, D.: Learning analytics in higher education - challenges and policies: a review of eight la policies. In: *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, pp. 233–242. ACM, New York (2017)

4. Tsai, Y.S., et al.: Teacher and Student Perspectives on Learning Analytics. SHEILA Executive Summary (2018). [http://sheilaproject.eu/wp-content/uploads/2018/06/Teacher-and-student-perspectives\\_3.pdf](http://sheilaproject.eu/wp-content/uploads/2018/06/Teacher-and-student-perspectives_3.pdf)
5. Young, J., Mendizabal, E.: Helping researchers become policy entrepreneurs - how to develop engagement strategies for evidence-based policy-making. Briefing paper, Overseas Development Institute (2009)



# Diversity Profiling of Learners to Understand Their Domain Coverage While Watching Videos

Entisar Abolkasim<sup>1</sup>(✉), Lydia Lau<sup>1</sup>(✉), Vania Dimitrova<sup>1</sup>(✉),  
and Antonija Mitrovic<sup>2</sup>(✉)

<sup>1</sup> School of Computing, University of Leeds, Leeds, UK  
{sc10ena, L.M.S.Lau, V.G.Dimitrova}@leeds.ac.uk

<sup>2</sup> Intelligent Computer Tutoring Group,  
University of Canterbury, Christchurch, New Zealand  
tanja.mitrovic@canterbury.ac.nz

**Abstract.** Modelling diversity is especially valuable in soft skills learning, where contextual awareness and understanding of different perspectives are crucial. This paper presents an application of a diversity analytics pipeline to generate domain diversity profiles for learners as captured in their comments while watching videos for learning a soft skill. The datasets for analysis were collected from a series of studies on learning presentation skills with Active Video Watching system (AVW-Space). Two user studies (with 37 postgraduates and 140 undergraduates, respectively) were compared. The learners' diversity and personal profiles are used to further understand and highlight any notable patterns about their domain coverage on presentation skills.

**Keywords:** Diversity profiling · Domain coverage · Diversity analytics pipeline  
Video-based learning · Presentation skills

## 1 Introduction

Videos are widely used by learners and tutors as a prime medium for learning and teaching [4]. Videos can be especially powerful for soft skills training. If carefully chosen, it can provide opportunities for self-regulated learning and for exploring different perspectives on the same situation. The paper presents a novel computational approach to automatically detect the domain coverage in learner comments by deriving diversity profiles for learners, and investigates how this may relate to individual learner's characteristics. The Semantic-Driven Diversity Analytics Tool (SeDDAT) presented in [1] has been extended and applied on new datasets, obtained from two user studies with postgraduate and undergraduate students respectively, on learning presentation skills from videos.

By adapting the Stirling Diversity Framework [6], domain diversity profiles for the learners are generated in terms of variety, balance and disparity. *Variety* refers to breadth of domain coverage. This is useful for learning to gather the learners' overview of the domain. *Balance* goes further and captures the extent of domain coverage. This is useful to see the degree of consistency in the level of understanding across domain categories.

*Disparity* refers to the density of domain coverage, i.e. measures how spread out the domain concepts are covered by the pool of comments. The learners' diversity and personal profiles are then analysed to address the following research question:

*Are there any notable differences between user groups with regards to domain diversity and individual profiles?*

## 2 Overview of the Diversity Profiling Pipeline

The pipeline consists of: **(a)** *Input preparation* for SeDDAT- including the appropriate ontology, an entry point and content file(s) annotated using this ontology. **(b)** *Diversity measurements* using SeDDAT to create diversity profiles. **(c)** *Diversity Analysis*, where the analyst inspects the diversity profiles for diversity patterns. **(d)** *Fine-tuning of profiling*, if interesting patterns are spotted, a new entry point for SeDDAT can be specified for further diversity analysis. More details are described in [2].

## 3 Datasets for Profiling

The diversity pipeline was applied on two user studies using the Active Video Watching (AVW-Space). **Study A** had 37 postgraduate students (PGs) and **Study B** 140 undergraduate students (UGs); details about the design of studies can be found in [3]. The following collected data were specifically relevant to this research: **(i)** data about the videos; **(ii)** user profiles, such as demographics, background experiences, Motivated Strategies for Learning Questionnaire (MSLQ) [5]; and **(iii)** user comments. The total number of comments was 744 from Study A and 1129 from Study B.

A Presentation Skills Ontology (PreSON) was developed (as described in [2]) to automatically tag the user comments and assist diversity profiling. The semantic tagging resulted in a total of 1,217 annotations for Study A and 2,070 for Study B; with 197 and 220 distinct entities, respectively. The average number of annotations and distinct entities per video covered by the comments are: Study A - 152.1 (*std.* 30.1) and 66 (*std.* 8.7); and Study B - 258.8 (*std.* 65.0) and 78.5 (*std.* 9.1).

## 4 Domain Diversity Profiling for Learners

The user (learner) diversity profiles and other associated data (e.g. demographics, MSLQ, knowledge, etc.) are analysed below to address the research question.

**Comparing the Study Groups.** Learners' personal and diversity profiles were compared to see if PGs differ from UGs in their background knowledge, attitudes towards learning and their behavior during video watching. Table 1 reports the profile items with significant difference between the two studies.

**Table 1.** Significant differences between learners from Studies A and B; † denotes Likert scale was used - 1 (lowest) to 5 (highest); \*\* significance at  $p < .005$ , and \* at  $p < .05$ .

User profile items	Study A (37)	Study B (140)	Significance
Domain variety	3.65 (.79)	2.91 (1.13)	$t = 4.55^{**}$
Domain balance	.49 (.24)	.28 (.22)	$t = 4.98^{**}$
Domain disparity	10.23 (4.43)	7.88 (5.4)	$t = 2.73^*$
Comments (no. of)	19 (12.87)	7.89 (9.59)	$t = 4.91^{**}$
Training <sup>†</sup>	2.14 (.95)	1.7 (.78)	$t = 2.57^*$
Experience <sup>†</sup>	2.86 (.79)	2.34 (.85)	$t = 3.53^{**}$
Task Value <sup>†</sup>	4.50 (.38)	3.97 (.59)	$t = 6.69^{**}$
Academic control <sup>†</sup>	3.91 (.47)	4.14 (.57)	$t = 2.64^*$
Intrinsic <sup>†</sup>	4.05 (.53)	3.77 (.59)	$t = 2.82^*$
Extrinsic <sup>†</sup>	3.41 (.81)	4.19 (.63)	$t = 5.48^{**}$
Effort regulation <sup>†</sup>	2.95 (.43)	3.56 (.66)	$t = 6.84^{**}$
Organisation <sup>†</sup>	3.85 (.95)	3.22 (.89)	$t = 3.63^{**}$
Elaboration <sup>†</sup>	4.13 (.55)	3.67 (.66)	$t = 4.33^{**}$
Self-Regulation <sup>†</sup>	3.61 (.39)	3.28 (.52)	$t = 4.19^{**}$

PGs seem to have higher domain diversity (variety, balance and disparity) compared to UGs – i.e. PGs had more diverse domain coverage with regards to presentation skills. PGs reported significantly more training and experience on presentation skills. This may explain the higher diversity scores. PGs scored higher on MSLQ Task Value, Intrinsic Motivation, Organisation, Elaboration and Self-Regulation, whereas UGs scored higher on Extrinsic Motivation, Academic Control and Effort regulation. These figures seem to correlate to the fact that Study A was on a voluntary basis (more comments) whereas Study B was part of a course assessment (fewer comments).

**Comparing Personal Attributes.** Each diversity property was compared across all learners' personal attributes to see if the learners' personal characteristics, such as gender or language (native/non-native speaker), will influence diversity scores. The only significant difference was in gender for Study A. The domain balance ( $U = 208$ ,  $p < .05$ ) and domain disparity ( $U = 205$ ,  $p < .05$ ) were significantly higher for female learners ( $n = 26$ ) than male learners ( $n = 11$ ). It is surprising that the language attribute did not have an impact on the coverage of the domain (i.e. diversity scores).

**Comparing the Most and Least Diverse Learners.** To further understand if any of the learners' profile items contribute to the domain coverage, their diversity scores are ranked for each study - variety, then balance, and then disparity. The top and bottom quartile were analysed versus the middle two quartiles. There are significant differences on the number of comments in both studies (Table 2) between these three subgroups of learners (all pairwise comparisons significant at  $p < .005$ ). In Study A, there was also significant difference between the subgroups based on presentation experience. Although it was expected that experienced learners should have higher diversity properties, this was not the case.

**Table 2.** Comparing quartiles defined on domain variety, balance and disparity (all pairwise comparisons significant at  $p < .005$ ).

	Top quartile Study A (9) Study B (35)	Middle quartiles Study A (19) Study B (70)	Bottom quartile Study A (9) Study B (35)	Significance
Experience study A	2.56 (1.13)	3.16 (.5)	2.56 (.73)	W = 7.664*
Comments study A	34.33 (13.19)	17 (6.77)	7.89 (7.42)	W = 20.64**
Comments study B	18.06 (13.94)	5.83 (3.61)	1.83 (1.36)	W = 83.46**

**Comparing Correlations.** For both studies, domain variety, balance and disparity are strongly correlated (with correlations ranging from .494 to .904, all at  $p < .005$ ). Also, the number of comments is strongly correlated with domain variety, balance and disparity in both studies ( $p < .005$ ). This indicates that learners should be triggered to write more comments, as the more they write the more they notice with regards to the domain while watching the videos.

## 5 Conclusion

This paper presented an instantiation of a novel semantic driven analytics pipeline for understanding domain diversity in learners' comments when watching videos. SeDDAT was applied on two studies about presentation skills to generate domain diversity profiles. PGs had significantly higher domain diversity than UGs; the former were more intrinsically motivated while the latter were more extrinsically motivated. It was surprising that the native language did not impact on the domain diversity, which indicated that cognitive understanding of presentation skills was orthogonal to language. This work contributes to future intelligent learning environments that address the needs of the learners and their diverse background in the modern society, which would require automated ways to capture and compare different domain perspectives.

**Acknowledgements.** Support by EU-FP7-ICT-257184 ImREAL, Ako Aotearoa and teaching development grant at University of Canterbury. We thank Amali Weerasinghe and Alicja Piotrkowicz for helping with the creation of PreSO<sub>n</sub>, and the expert trainers from Skills@Library at Leeds University for validating the ontology. Finally, we thank all participants in the user studies.

## References

1. Abolkasim, E., Lau, L., Dimitrova, V.: A semantic-driven model for ranking digital learning objects based on diversity in the user comments. In: Verbert, K., Sharples, M., Kloboučar, T. (eds.) EC-TEL 2016. LNCS, vol. 9891, pp. 3–15. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45153-4\\_1](https://doi.org/10.1007/978-3-319-45153-4_1)
2. Abolkasim, E., Lau, L., Dimitrova, V., Mitrovic, A.: Ontology-based domain diversity profiling of user comments. In: Proceedings of the 17th Conference on Artificial Intelligence in Education (2018). (in press)

3. Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., Weerasinghe, A.: Using learning analytics to devise interactive personalised nudges for active video watching. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 22–31. ACM (2017)
4. Hua, K.: Education as Entertainment: YouTube Sensations Teaching The Future. Forbes (2015)
5. Pintrich, P.R., De Groot, E.V.: Motivational and self-regulated learning components of classroom academic performance. *J. Educ. Psychol.* **82**(1), 33 (1990)
6. Stirling, A.: A general framework for analysing diversity in science, technology and society. *J. R. Soc. Interface* **4**(15), 707–719 (2007)





# Using Digital Medical Collections to Support Radiology Training in E-learning Platforms

Félix Buendía<sup>1</sup>, Joaquín Gayoso-Cabada<sup>2</sup>, and José-Luis Sierra<sup>2</sup>(✉)

<sup>1</sup> Universitat Politècnica de València, 46071 Valencia, Spain  
fbuendia@disca.upv.es

<sup>2</sup> Universidad Complutense de Madrid, 28040 Madrid, Spain  
{jgayoso, jlsierra}@ucm.es

**Abstract.** This work is focused on using the huge amount of medical cases available in medical digital collections to support specific radiology training courses, particularly addressed to medical residents. Such support is based on retrieving information items from these extant digital collections and generating instructional resources that can be deployed in the resident training context. The key element for this information management is a tool called *Clavy*, which retrieves pieces of content in medical collections and allows hospital tutors to generate educational resources easily under standard specifications and work with them in the most popular e-learning platforms. An example of a radiology course was implemented in *Moodle* to demonstrate the *Clavy* approach to the generation of training resources and their use in e-Learning platforms.

**Keywords:** Medical knowledge · Learning objects · E-learning platforms

## 1 Introduction

Medical knowledge has been growing over the last years in an exponential way. Such growth is particularly significant in the area of radiology, where multiple medical digital collections related to radiology topics have been developed. The current work is focused on using the huge amount of medical cases available in these collections, to support specific training courses, particularly addressed to medical residents who combine the practice of medicine and instruction. To this aim, we have developed an experimental tool called *Clavy* [2, 3], which can help to organize these repositories and contribute to improving the knowledge gathered by radiologists during their residency period in hospitals. A group of physicians at the *la Fe* hospital (Valencia, Spain) has recently started to practice with a set of medical case examples in the radiology area to test their training potential and the suitability of information management tools for processing them. Assessment results from the process promoted by *Clavy* involving these physicians are very positive.

The remainder of the work is structured as follows. Section 2 introduces the *Clavy* approach. Section 3 exemplifies this approach. Finally, some conclusions and further work are drawn in Sect. 4.

## 2 The *Clavy* Approach

*Clavy* supports a three-step workflow:

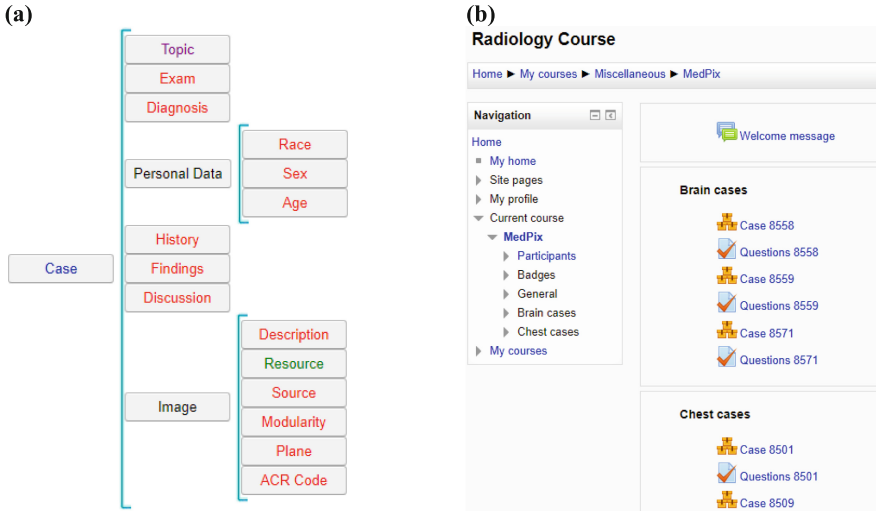
- In the first step, instructors discover and import digital resources from different sources with a high educational value suitable to be transformed into learning objects. For this purpose, *Clavy* enables the aggregation of the content of multiple collections using *plug-ins*. In the case of simple medical collections (e.g., unstructured sets of DICOM images) it can be possible to use a general-purpose *plug-in* to perform the importation (e.g., in this case, a *plug-in* able to extract the information from DICOM records). However, more complex collections (e.g., *MedPix* or *EuroRad*) will already exhibit a collection-specific structure that must be adequately preserved by the importation process. In this case, the most typical situation is to provide a collection-specific *plug-in* able to connect to the external source in order to ingest relevant learning objects together with all the associated information.
- In the second step, instructors can curate all the information imported, ensuring a coherent and unified structure and reorganizing the repository to meet the specific needs of the target users (medical residents, in our case).
- In the third step, objects can be exported in standard e-learning formats like IMS-CP, SCORM or IMS Common Cartridge to be published in suitable learning management systems or in other e-learning platforms. For this purpose, *Clavy* provides a second kind of *plug-in* to export the complete repository, or a part of it, to third-party platforms.

## 3 Applying the *Clavy* Approach to *MedPix*

In order to exemplify the different aspects of the *Clavy* process, we will outline how it was used on the aforementioned *MedPix* medical collection on clinical cases [1]:

- Importation was carried out using a collection-specific *plug-in*. This *plug-in* lets residents' instructors recover *clinical cases* as learning objects. In *MedPix*, *clinical cases* (comprising clinical images and additional descriptive information) cover different *clinical topics*, since both types of elements are cross-referenced. Therefore, once the instructor indicates the clinical cases to ingest the following steps are performed: (i) the *plug-in* uses the *MedPix* REST API to recover the URLs in *MedPix* for these clinical cases; (ii) in turn, each case can be recovered by using the REST API again; (iii) then, by scraping each case, the *plug-in* is able to discover the set of related topics; (iv) the actual information for the topics can be retrieved by using the REST API a third time; (v) topics are in turn scraped to retrieve additional related cases, which are then ingested and analyzed until all the relevant information has been retrieved; and (vi) once all the relevant information is ingested, the *plug-in* makes all this information persistent as a *Clavy* repository.
- Once the learning objects were imported into *Clavy*, instructors of residents curated these objects by using a *schema* editor and a *learning object* editor. In particular, the *schema* editor was very useful for adapting the initial organization produced by the

importation *plug-in* to specific educational settings. Indeed, the initial *MedPix* schema contained 72 attributes, many of which are not excessively interesting from an educational point of view. After editing it, these attributes were reduced to 28, the most useful from an educational point of view, plus some oriented to enhancing structure (Fig. 1a).



**Fig. 1.** (a) Excerpt of the reconfigured *Clavy* schema for learning objects imported from *MedPix*; (b) snapshot of the sample *MedPix*-based course deployed on *Moodle*.

- The resulting learning objects, associated to *MedPix* medical cases, were used to implement a sample course on *Moodle*. For this purpose, these learning objects were exported as IMS Content Packages using a suitable *Clavy* exportation *plug-in*. The course was organized using a simple structure (Fig. 1b): (i) an *Introduction* forum that explained its main features, inviting participants to ask questions about the course objectives; and (ii) a main corpus of *MedPix* medical cases with their structured description and attached MCQs (Multiple Choice Questions) to be answered by volunteer residents.

The course finally implemented allowed us to assess the approach promoted in this work in two different dimensions:

- On one hand, the course let us assess the extent to which the approach can suit the needs of instructors (the staff in charge of tutoring residents). For this purpose, we actively involved instructors in the design of the course. They found the simple instructional structure proposed, based on the intercalation of clinical cases and related MCQs, adequate, and helped to select the corresponding items.
- On the other hand, the course was used to explore the access to the instructional resources by medical residents at the *la Fe* hospital and to check their answers to questions extracted from the *MedPix* collection. Opinions gathered from the interactions

of the residents with course resources and questionnaire items revealed a general satisfaction with their instructional usefulness. One of the main features they observed is the potential of those instructional resources to link image information with radiology text reports and the way such links can be explored and evaluated by means of test questionnaires and other similar activities (e.g. forum posts). On the negative side, most users highlighted that better image visualization was required and a stronger relationship between descriptive cases and questionnaires should be established. Nevertheless, course outcomes were mostly positive, which made the educational potential of the approach apparent.

## 4 Conclusions and Future Work

The current work has presented the *Clavy* platform as a key element in the process of collecting, transforming and generating instructional resources in the radiology area. Through the development of a course oriented to the training of residents in radiology at the *la Fe* hospital, *Clavy* has proved to be a useful tool for tutors, not only for collecting data from these multiple and diverse information sources in a versatile way, but also to process such data by transforming the associated semantic structure and generating useful contents with instructional purposes. The outcomes concerning the participation of residents in the course have been very positive, highlighting the degree of engagement of radiology residents who enrolled in the course.

Currently we are working on supporting the exportation of *Clavy* learning objects to other e-Learning formats that support interaction (e.g., in particular, SCORM packages). We are also working on the importation of MCQs from *MedPix* and on the embedment of these MCQs in SCORM packages. Further works plan to support the IMS Common Cartridge, and also to implement new courses to assess users who could participate in a residency hospital program as part of their training (*Clavy* would help hospital tutors to generate their own contents in this training program based on the extraction of medical cases in which they are involved).

**Acknowledgements.** Thanks to the support of the Research Projects TIN2014-52010-R and TIN2017-88092-R and residents and tutors of *la Fe* hospital (Valencia, Spain).

## References

1. MedPix. <https://medpix.nlm.nih.gov/>. Accessed 03 Apr 2018
2. Gayoso-Cabada, J., Rodríguez-Cerezo, D., Sierra, J.-L.: Browsing digital collections with reconfigurable faceted thesauri. In: Proceedings of International Conference on Information Systems Development (ISD), Katowice, Poland, pp. 378–389 (2016)
3. Gayoso-Cabada, J., Rodríguez-Cerezo, D., Sierra, J.-L.: Multilevel browsing of folksonomy-based digital collections. In: Cellary, W., Mokbel, M.F., Wang, J., Wang, H., Zhou, R., Zhang, Y. (eds.) WISE 2016, Part II. LNCS, vol. 10042, pp. 43–51. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-48743-4\\_4](https://doi.org/10.1007/978-3-319-48743-4_4)



# Temporal Analytics of Workplace-Based Assessment Data to Support Self-regulated Learning

Alicja Piotrkowicz<sup>1</sup>(✉), Vania Dimitrova<sup>1,2</sup>, and Trudie E. Roberts<sup>1</sup>

<sup>1</sup> Leeds Institute of Medical Education, University of Leeds, Leeds, UK  
A.Piotrkowicz@leeds.ac.uk

<sup>2</sup> School of Computing, University of Leeds, Leeds, UK

**Abstract.** One of the most effective ways to develop self-regulated learning skills in higher education is to include work placements. Workplace-based assessment (WBA) provides opportunities for students to gain feedback on their practical skills, reflect on their performance, and set goals and actions for further development. This requires identifying temporal patterns, as placements usually span extended periods of time. In this paper we explore two intelligent computational methods (burst detection and process mining) to derive temporal patterns. We apply both methods on WBA data from a cohort of first-year medical students. Through this we identify interesting temporal patterns, and gather educators' feedback on their usefulness for self-regulated learning.

## 1 Introduction

Preparing lifelong learners, who develop self-regulation skills and continuously grow as professionals, is a key challenge in higher education. A way to develop self-regulation skills is to include work-based activities to allow students to engage in the professional practice for a substantial period of their studies. However, it is challenging to gain a holistic view of placement experience and to link placement to continuous personal development. This requires noticing temporal patterns related to placement engagement, which can trigger reflection and promote self-regulation [1]. A common way to support the discovery of temporal patterns is by using visual dashboards. The effectiveness of visualisations depends on the capability to discover relevant patterns and link them to self-regulated learning. Unlike (usually static) visualisations, intelligent data analysis allows automatic discovery of temporal patterns, so that interactive nudges can be provided to trigger reflection. This calls for the identification of aspects and patterns of the learners' data which are beneficial for self-regulation. Our work investigates this in a case study of medical education. At our institution students are fairly independent in their choice of WBA topic and timing, which necessitates developing SRL skills. An interactive visualisation backed by analytics could help students with sense making and action planning for their learning.

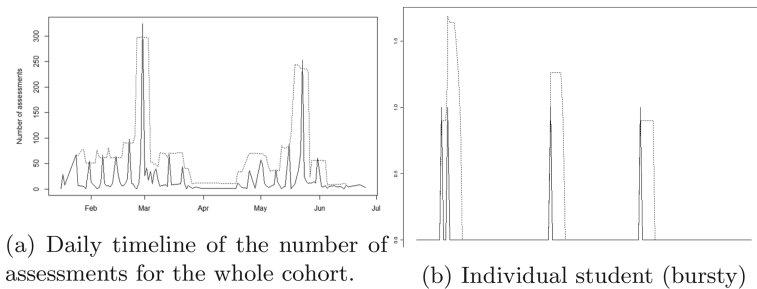
We investigate two temporal analytics methods: burst detection and process mining, and address the following research questions: (i) *What patterns can be derived from WBA data using burst detection and process mining?*, and (ii) *Which patterns identified using temporal analytics are judged as useful by educators?*

## 2 Methods

The goal of analysing the temporal aspect of workplace-based assessment data is to identify patterns and processes which can support students' self-regulated learning. We conduct both cohort-level and individual-level analysis.

**Data.** We use WBA data for a cohort of 1<sup>st</sup> year medical students, consisting of 2360 unique assessments undertaken by 228 students between January and June 2017. During work placements students are assessed on a list of mandatory and optional clinical skills that they need to acquire throughout their degree. As students could freely choose the number of assessments they wish to undertake, there are considerable differences in assessment counts between students.

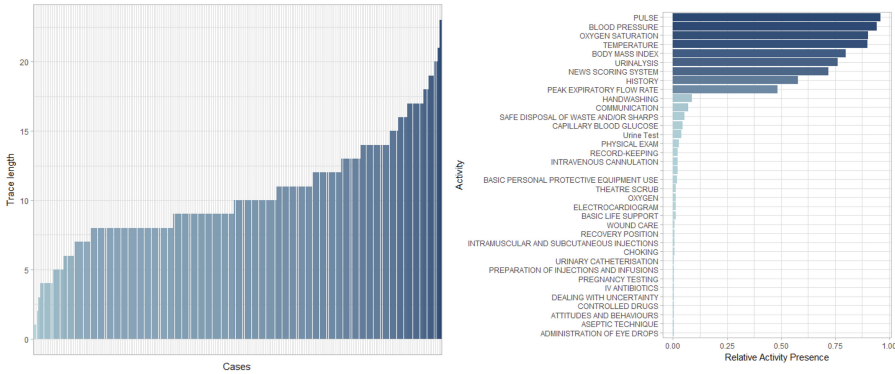
**Burst Detection.** Students can freely choose when they will undertake a WBA during their placements. Although they are encouraged to do the assessments regularly, it is often the case that students decide to do a number of assessments in a very short period of time (usually on the same day), resulting in a 'burst', i.e. a spike in assessment activity. Such burstiness might be a possible parameter for identifying at-risk students. We implemented the burst detection algorithm by [2] and applied it to the whole cohort and to each student separately (cf. Fig. 1). At the cohort level we identified 10 bursts. We noted that the most noticeable bursts corresponded to the end of placements.



**Fig. 1.** Burst detection examples. Solid line = # assessments, dotted line = cutoff. A burst occurs at any time point where the solid line is above the dotted line.

**Process Mining.** Process mining transforms temporal data into an event log which generalises unique individual paths through a task into common pathways. It originated in the business domain and is used extensively in healthcare (e.g. [3]). It has been applied to a limited extent in education, particularly in the

field of education data mining (e.g. [4]). Until now its applicability has not been investigated for WBA data. We used the *bupaR* process mining package in R for the processing<sup>1</sup>. The WBA event log yielded almost no common processes (225 unique paths for 228 students). A coarser granularity (e.g. skill category, assessor role) might result in more common pathways. We were still able to use the event log to obtain some pre-defined metrics, including: a summary of the trace lengths (i.e. the number of assessment per student; cf. Fig. 2a), and the percentage of students that have completed a given clinical skill (cf. Fig. 2b).



(a) Number of assessments per student for the whole cohort. (b) Percentage of students to complete a clinical skill.

**Fig. 2.** Example visualisations of queries against the WBA event log.

### 3 Evaluation and Conclusions

Our initial evaluation involved semi-structured interviews with two educators (one clinical education expert responsible for developing and running the clinical skills education programme, and one technology-enhanced learning expert responsible for the e-portfolio and TEL outreach). We asked: (i) *Is a particular analytics method producing any useful insights?*, and (ii) *Is it useful for students, or educators?* The materials used in interviews are made available<sup>2</sup>.

**Feedback on Burst Detection.** It is important to know when students are completing assessments, and whether they are consistent. Burst detection would be useful from an administrative perspective, especially if mapped against the beginning and end of placement. It could also be used for quality assurance of individual placements (e.g. whether students are given a range of opportunities for assessment). The method would be less useful for students. Furthermore, the

<sup>1</sup> <https://cran.r-project.org/web/packages/bupaR/bupaR.pdf>.

<sup>2</sup> <https://bit.ly/2r93nJs>.

results should consider that some clinical skills are commonly assessed together, so several assessments in a day might not reflect a true burst.

**Feedback on Process Mining.** In general, process mining would be useful from an administrative perspective, such as assessing placement quality. The skills type analysis was judged as particularly useful, as it shows that students are not engaging with optional skills. As the students move through the degree, they are expected to recognise WBA as a learning opportunity and engage with the optional skills more. Additional information about the expected entrustability level, skill category, and comparison to the cohort, would be useful.

**Issues Surrounding Temporal Analytics.** Both educators pointed out that temporal analytics are useful, but they do not provide enough context of the assessment. One educator said that it is important to look into the textual feedback from the assessor and the student's response to it, for a more holistic picture of the students' learning process. Generally, the temporal analytics considered in this paper were judged to be useful for placement quality assessment. The analytics could be used to visualise to students, however it raises the question whether students would be able to interpret and act on the information shown to them. Data interpretation would need to be integrated within the curriculum, so that students would be able to use to support their self-regulation.

**Conclusions.** We applied burst detection and process mining to workplace-based assessment data, and found notable patterns: (i) at the cohort-level the most noticeable bursts corresponded to the end of placements, (ii) the number of completed assessments per student varied considerably, and (iii) students rarely chose to complete assessment for optional clinical skills. The analytics were evaluated by two educators as particularly useful for assessing the quality of clinical placements. Two issues were identified: (i) lack of context provided by the count data, and (ii) potential difficulty in interpreting this kind of data visualisations by students. In future work we want to address these issues by incorporating text analytics, and by adding data interpretation to the curriculum.

**Acknowledgements.** This research was conducted as a part of the myPAL project, which involves a large team of educators, software developers, and researchers (<http://mypalinfo.leeds.ac.uk/people/>).

## References

1. Winne, P.: Learning analytics for self-regulated learning. In: Lang, C., Siemens, G., Wise, A., Gašević, D. (eds.) *Handbook of Learning Analytics*, pp. 241–249. Society for Learning Analytics Research (2017)
2. Vlachos, M., Meek, C., Vagena, Z., Gunopoulos, D.: Identifying similarities, periodicities and bursts for online search queries. In: *SIGMOD* (2004)



3. Baker, K., Dunwoodie, E., Jones, R.G., Newsham, A., Johnson, O., Price, C.P., Wolstenholme, J., Leal, J., McGinley, P., Twelves, C., et al.: Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *Int. J. Med. Inf.* **103**, 32–41 (2017)
4. Romero, C., Cerezo, R., Bogarín, A., Sánchez-Santillán, M.: Educational process mining: a tutorial and case study using moodle data sets. In: ElAtia, S., Ipperciel, D., Zaïane, O.R. (eds.) *Data Mining and Learning Analytics: Applications in Educational Research*. Wiley (2016)



# Learning Analytics Dashboard Analysing First-Year Engineering Students

Jonas Vaclavek<sup>1</sup>(✉), Jakub Kuzilek<sup>1</sup>, Jan Skocilas<sup>2</sup>, Zdenek Zdrahal<sup>1</sup>,  
and Viktor Fuglik<sup>1,3</sup>

<sup>1</sup> CTU in Prague, CIIRC, Jugoslavskych Partyzanu 1580/3,  
166 00 Prague 6, Czech Republic  
jonas.vaclavek@cvut.cz

<sup>2</sup> CTU in Prague, FME, Technicka 4, 166 07 Prague 6, Czech Republic

<sup>3</sup> Faculty of Education, Charles University, Magdaleny Rettigove 4,  
116 39 Prague 1, Czech Republic

**Abstract.** Nowadays, the higher education institutions experience the problem of the student drop-out. In response to this problem, universities started employing analytical dashboards and educational data mining methods such as machine learning, to detect students at risk of failing their studies. In this paper, we present interactive web-based Learning Analytics dashboard - *Analyst*, which has been successfully deployed at Faculty of Mechanical Engineering (FME), Czech Technical University in Prague. The dashboard provides academic teaching staff with the opportunity to analyse student-related data from various sources in multiple ways to identify those, who might have difficulties to complete their degree. For this purpose, multiple analytical dashboard views have been implemented. It includes summary statistic, study progression graph, and credit completion probabilities graph. In addition, users have the option to export all analysis related graphs for the future use. Based on the outcomes provided by the *Analyst*, the university successfully ran the interventions on the selected at-risk students and significantly increased the retention rate in the first study year.

**Keywords:** Analytical dashboard · Educational data mining · Student retention  
Data visualisation

## 1 Introduction

In recent years, universities face a problem of low retention of the students, especially in the first year of a university degree. In EU countries between 20% and 54% of students fail to complete their degree [1]. At the same time, higher education has experienced extensive growth of ICT based educational systems. These systems allow universities to collect a vast amount of data, which can be further analysed.

Learning Analytics (LA) seems to be one of the most promising fields regarding the potential of the student data analysis [2, 3]. One of the visualisation tools for displaying the results of the analysis is learning dashboards [4].

In our previous work [5], we have proposed a technique of discovering students at risk of failing in an academic year. These students can be offered an assisted help to increase their chances to finish the academic year with fewer problems.

In this paper, we present a learning dashboard which has been successfully deployed at the FME [5]. The application implements the approaches proposed in [5] and other analytical methodologies to support teachers while making decisions about learning processes.

## 2 Data

The FME uses a university system to export anonymised data. It contains personal and demographic data of the students along with their performance in courses. Additionally, for each performance record, there is a timestamp representing a date in which student has been given a grade. The date is vital information used in the analysis to calculate students' performance in an academic year [5].

The Analyst has been primarily developed for analysing the first-year students. Their data is available for four consecutive academic years (starting from 2013). In each year, approximately four hundred students have registered to the degree, and between 15% to 20% of them failed in the first year. Students are divided into five performance groups for each academic year based on the criteria defined by the FME staff concerning their teaching expertise.

## 3 Analyst

The dashboard has been implemented as a web application using Shiny technology<sup>1</sup> which is available for R language<sup>2</sup>. The combination of Shiny and R creates an environment where a web application with R outputs can be easily created.

All graph components of the application are created using the ggplot2<sup>3</sup> library and converted using the Plotly<sup>4</sup> library to make all the graphs in the application interactive and provide an option to store them locally.

The user interface of the application is divided into three parts (Fig. 1). Upper part provides a data filter that affects input data of all analytical tools. The user can filter by academic year, gender, a form of study and course type. It is also possible to analyse only first-year students. Whenever a change occurs in the filter, the dashboard recalculates the data and renders the results automatically. Left part contains a navigation menu for dashboard views and administration tools. Central part shows a content based on the selection in the menu. Administration tools are used to upload exported data, edit general course details or manage other parameters of the application. Following sections explain the dashboard views.

---

<sup>1</sup> <http://shiny.rstudio.com>.

<sup>2</sup> <https://www.R-project.org/>.

<sup>3</sup> <https://CRAN.R-project.org/package=ggplot2>.

<sup>4</sup> <http://ggplot2.org/>.



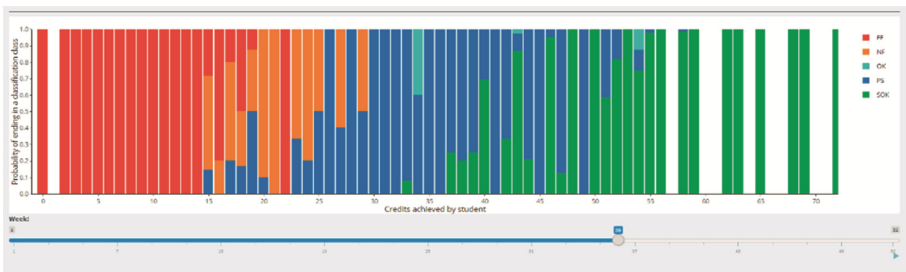
**Fig. 1.** Analyst overview. SOK - no problems, OK - problems in one of semesters, PS - problems in winter semester, NF - failing in summer semester, FF - failing in winter semester.

### 3.1 Study Progress

The view contains a graph (central part of the Fig. 1) with students divided into performance groups. Each group is represented by a line which shows the average number of credits earned by students in the group for each week of the academic year. The dotted lines split the graph into several segments for better orientation in the academic year.

### 3.2 Study Probabilities

The view consists of two components (Fig. 2). The first component allows the user to select a week of the academic year. The second component takes that week, filters data until the selected week and creates a graph where the horizontal axis shows numbers of credits and the vertical axis displays the probability of ending up in a specific performance group based on the number of credits earned up to the specified week. The probabilities are calculated using the Bayes’ formula.



**Fig. 2.** Study probabilities graph

### 3.3 Summaries

The view (Fig. 3) gives an overview of the filtered data. In the top part, it shows a number of students in each performance group and percentage with respect to the whole cohort.

The histogram displays all courses in the selected dataset on horizontal axis and probability of achieving the corresponding mark on the vertical axis.

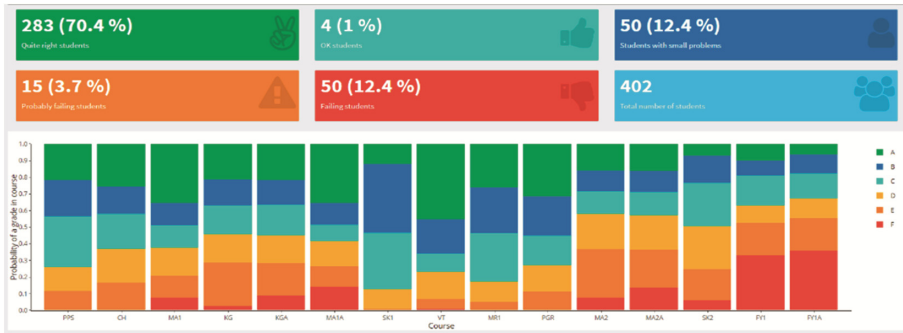


Fig. 3. Summaries view

## 4 Conclusion

We have developed a web application with multiple analytical tools which has been used by the FME staff to increase the student retention by more than 49%<sup>5</sup>. In the future, the *Analyst* will be further extended to provide predictive modelling tools for estimation of students' outcomes.

**Acknowledgement.** This work was supported by junior research project by Czech Science Foundation GACR no. GJ18-04150Y.

## References

1. Quinn, J.: Drop-out and completion in Higher Education in Europe (2013). <https://edudoc.ch/record/110174/files/dropout.pdf>
2. Shacklock, X.: From bricks to clicks: the potential of data and analytics in higher education (2016)
3. Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., Rienties, B., Ullmann, T., Vuorikari, R.: Research Evidence on the Use of Learning Analytics: Implications for Education Policy, Seville (2016)
4. Jivet, I., Scheffe, M., Specht, M., Drachsler, H.: License to evaluate: Preparing learning analytics dashboards for educational practice. In: Proceedings of the Eight International Conference on Learning Analytics and Knowledge, Sydney (2018)
5. Zdrahal, Z., Hlosta, M., Kuzilek, J.: Analysing performance of first year engineering students. In: Proceedings of the Data Literacy for Learning Analytics Workshop, Edinburgh (2016)

<sup>5</sup> [https://analyse.kmi.open.ac.uk/resources/documents/letter\\_of\\_recognition.pdf](https://analyse.kmi.open.ac.uk/resources/documents/letter_of_recognition.pdf).



# The Learning Analytics Indicator Repository

Daniel Biedermann<sup>1(✉)</sup>, Jan Schneider<sup>1(✉)</sup>, and Hendrik Drachsler<sup>1,2,3(✉)</sup>

<sup>1</sup> German Institute for Educational Research, Frankfurt, Germany  
{biedermann,schneider.jan,drachsler}@dipf.de

<sup>2</sup> Open Universiteit, Valkenburgerweg 177, 6419 AT Heerlen, Netherlands

<sup>3</sup> Goethe University Frankfurt, Frankfurt am Main, Germany

**Abstract.** This paper presents the Learning Analytics Indicator Repository (LAIR), an interactive web-based application that allows the exploration of learning analytics approaches. From scientific publications in the field of learning analytics, we extracted the stakeholders, metrics, platforms and indicators, and transformed them into a directed graph representation. The LAIR allows filtering by these components and provides a list of publications where the approaches can be found. We invite other researchers to contribute to this repository.

**Keywords:** Indicators · Review · Learning analytics

## 1 Introduction

In Learning Analytics (LA), there is heterogeneity in the way data is collected and analyzed, as well as in the goals that these approaches aim to achieve. Some infer performance measures using institutional data such as grades, and scores [1], while other approaches try to reach the same goal by utilizing natural language processing methods on written texts [2]. In other cases, the data retrieved from the students remains the same, but the goals of LA are different. Approaches may use Learning Management Systems (LMS) data as part of a classifier to infer learning strategies [3], while others use the same data again for performance prediction [4]. Moreover, various approaches utilize the same measurements in different combinations.

Those faced with the task of researching a particular subset of LA, and those wishing to implement LA at their institution, may have difficulty finding out which approaches actually exist for their particular endeavor, as well as the constraints and requirements that this may entail.

Information about the various approaches taken in LA can be found in reviews on LA in general [5], and in the more specialized reviews of LA dashboards [6]. While traditional reviews do report on the several parts of the LA approaches, it can often be difficult to find out where and how to find these results if the results are just listed as numbers in a table.

We have therefore created the Learning Analytics Indicator Repository (LAIR), which provides an overview of the LA landscape, where the approaches are visualized in a comprehensible and explorable way with links to their sources. It is intended to users to quickly find resources related to their topic, to identify what to measure, and how to measure it.

## 2 Method

From 122 research papers in the field of LA (we are continuously adding more approaches to the LAIR), we extracted the approaches and split them into the following components:

**Subjects.** Those that the data is collected from.

**Activity.** Activities are what the collected data is about, i.e. what the learner actually does while data is collected.

**Platform.** The technology or the place where the data collection occurs while the data subjects are performing their activities.

**Metrics.** The metrics are the measurements that are collected on the platform about the activity.

**Indicator.** An indicator shows if and to what extent a particular concept can be derived from the metrics.

**Inference.** This is what the LA concept aims to achieve. We use this term when the use of indicator is too specific and a more general description is needed, e.g. for the explorative studies which are not looking for particular indicators and just try to see what they can infer from the data.

We put the extracted components into categories, merged all the terms that are only syntactically different (e.g. ‘course grade’ and ‘final exam grade’), and clustered them together under common umbrella terms. We chose the terms such that they could later aid in the discoverability.

We interpret each LA approach as paths from the creation of the data by the subjects to the indicators as the outcomes. This suggests that an approach can be represented as a directed graph (see Fig. 1). The directed graph representation enables searching, traversing and, moreover, allows a very intuitive visualization to quickly perceive the relation between the various components of an approach.

## 3 Learning Analytics Indicator Repository

We put all of the extracted approaches into their graph representation and collected them in an interactive web-based application that we call the Learning Analytics Indicator Repository (LAIR), which can be found online <sup>1</sup>.

In the LAIR, users can filter for the components that they are interested in. The set of research papers which contain at least one of the selected filters is listed. For example, clicking on the metric category “Affective State” lists all

<sup>1</sup> <http://lair.edutec.guru>.

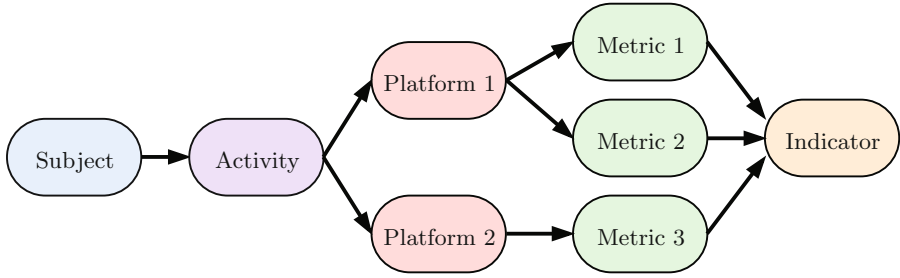


Fig. 1. LA approaches as a directed graph

papers which have used affective state as part of their approaches. Selecting one of the papers from this list reveals the visualization of the graph for this paper (see Fig. 2).

The LAIR also offers a visualization that displays the selected filters combined in one sankey diagram, where the size of a node reflects how often a particular component was part of an approach, and the size of a link reflects how often a particular component is used in conjunction with another component.

The LAIR is intended as an ongoing effort where we will continue to incorporate and submit our findings. It is also explicitly conceptualized as a project for the research community, and we added a form that allows submission of scientific publications that are not yet listed.

The screenshot shows the LAIR interface with three main sections:

- Filters (left):** A sidebar with 'Subject' and 'Activity' sections. Under 'Subject', 'Student' is selected. Under 'Activity', 'Collaborative Learning' is selected.
- Lists of Papers (center):** A table listing various papers with their titles and IDs. The paper 'Modeling Learning & Performance: A Social Networks Perspective' is highlighted.
- Selected Approach Visualization (right):** A graph showing the flow from 'Student' to 'Online Course' to 'Forum', which then branches into 'Social Interactions' and 'Network Analysis', both leading to 'Performance Prediction'.

Fig. 2. The LAIR showing the filters (left), the list of papers (center), and the graph for the selected paper.



## 4 Conclusion and Future Work

With the LAIR, we have created an explorable and comprehensible overview of the LA approaches that have been researched in the literature. The LAIR provides the approaches as directed graphs and visualizes them in a dedicated web-based application. Approaches can be filtered by their components and visualized in a graph, making their composition more transparent and understandable.

The LAIR is created as an ongoing effort and becomes increasingly useful the more research is actually captured in it. We therefore designed it such that other researchers can contribute to it, and hope that it enables other researchers to find related works.

Visualizations of learning analytics approaches as graphs could also be used as part of LA dashboards, where these graphs could help to explain learners how the analytics results have been achieved. This can be a step towards algorithmic transparency and help the field in the efforts towards consolidation.

## References

1. Rogers, T., Colvin, C., Chiera, B.: Modest analytics: using the index method to identify students at risk of failure. In: ACM Press, pp. 118–122 (2014)
2. Robinson, C., Yeomans, M., Reich, J., Hulleman, C., Gehlbach, H.: Forecasting student achievement in MOOCs with natural language processing. In: ACM Press, pp. 383–387 (2016)
3. Jovanović, J., Gašević, D., Dawson, S., Pardo, A., Mirriahi, N.: Learning analytics to unveil learning strategies in a flipped classroom. *Internet High. Educ.* **33**, 74–85 (2017)
4. Waddington, R.J., Nam, S.: Practice exams make perfect: incorporating course resource use into an early warning system. In: ACM Press, pp. 188–192 (2014)
5. Moissa, B., Gasparini, I., Karczinski, A.: A systematic mapping on the learning analytics field and its analysis in the massive open online courses context. *Int. J. Distance Educ. Technol.* **13**(3), 1–24 (2015)
6. Schwendimann, B.A., et al.: Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Trans. Learn. Technol.* **10**(1), 30–41 (2017)



# Eye-Tracking for User Attention Evaluation in Adaptive Serious Games

Alexander Streicher<sup>(✉)</sup>, Sebastian Leidig, and Wolfgang Roller

Fraunhofer IOSB, Karlsruhe, Germany

{alexander.streicher,sebastian.leidig,wolfgang.roller}@iosb.fraunhofer.de

**Abstract.** The Ideal Path Score (IPS) developed in this work is able to improve adaptivity of serious games by more accurately estimating performance and need for help based on players' interactions and eye movements. The automatic personalization of adaptive e-learning systems supports effective learning for users with varying levels of knowledge and skills. Particularly in games, indicators informing adaptivity, like attention and performance of the player, should be assessed non-invasively to avoid interrupting the player's flow experience and to keep up the immersion. Passive sensors like eye tracking can solve this challenge. This paper presents the concept of the IPS and its integration in an adaptive serious game for image interpretation training. The realized IPS-adaptive game assesses performance and attention of players based on eye movements and interactions with the game.

**Keywords:** Adaptive games · Eye tracking · Ideal path  
Serious games

## 1 Introduction

The problem statement in this paper deals with the question of when an adaptive serious game needs to adapt, e.g., when to automatically personalize or customize the learning experience, and how to effectively assess user progress. The correct timing is important to match the players' needs [5]. Digital game based learning needs to constantly motivate the users and sustain a constant flow experience to achieve an effective learning outcome [1]. This flow, a balance of challenges and skills, can keep the learner self-motivated and is an important aspect of effective serious gaming. Adaptive serious games for learning try to personalize the gaming and learning experience to keep the user in the flow channel and to maximize to learning outcome. Effective adaptivity is based upon sound user or learner models which contain all the necessary information to adaptively guide the learner. The user models can include information on the users' abilities, which can be measured either implicitly or explicitly. However, in serious gaming each intervention by an explicit measurement, e.g., via user questionnaires, could have negative impacts on the users' flow experience. Hence, implicit measurements try to estimate the users' current learning progresses or cognitive states. For effective adaptivity, ideally the adaptive interventions would be guided by a measure of the users' progress. One possibility to measure the users' progress

is to look at the purposefulness or goal-orientedness of their actions. A user working efficiently towards the goal obviously does not need further assistance, whereas a user who is lost or moving in a wrong direction should be adaptively assisted. An approach to measure such a goal-orientedness is the definition of a metric to measure the distance between an “ideal path” and the users’ observed action [3]. The scientific research question is, if there is a correlation between gaze, ideal path and attention. We are asking, when the users follow the ideal path do their gazes also follow that path, and can the attention level be inferred (estimated) from that. We contribute the concept and work-in-progress of the *Ideal Path Model* (IPM) and the *Ideal Path Score* (IPS) for its application for attention level detection with eye tracking.

## 2 Ideal Path Score for Adaptivity

The IPS can improve adaptivity of serious games by more accurately estimating performance and need for help based on players’ interactions and supported by eye movements. The IPS is especially helpful in combination with gaze or eye tracking. Eye tracking can give insights on the cognitive states of the users by tracking their visual attention. A typical example would be that attention is turned to the first area of interest, moving the fovea to this point. Once the movement is complete, the feature is inspected with higher attention before moving to the next area of interest [2]. This gaze data can make an adaptive system more robust, i.e., a high correlation between gaze direction and pointing coordinates (mouse clicks or touch events) could indicate a high user attention level. To evaluate the attention level in regard to a goal-orientedness the Ideal Path Model as a reference model has been developed.

### 2.1 Ideal Path Model

The IPM describes all necessary steps to reach the game’s goals without any unnecessary detours. In essence it is a sequence of episodes and interactions in a

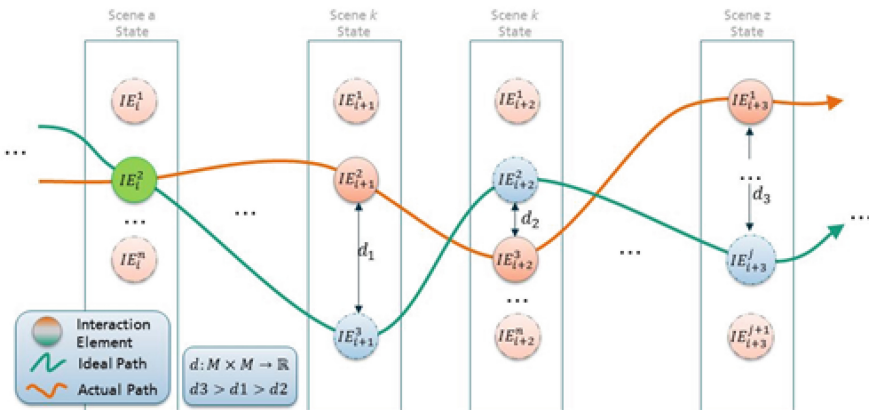


Fig. 1. Ideal Path Model with scene states, interaction elements and distances.

virtual environment that most directly leads to the next goal [3]. For example, in an adventure game, the ideal path would be the optimal walk-through, i.e., the optimal sequence of interactions from the game start to the game ending. The building blocks of the IPM are (Fig. 1): (1) scene manifestations which capture the current state of a scene; (2) interaction elements which are all game elements a player can interact with; (3) an ideal path through the sequence of all scene manifestations and interaction elements; (4) the actual path which reflects the actual sequence one player has taken. A scene can have multiple manifestations for each possible interaction a user can undertake. The IPM can be built manually or automatically by recording the steps an “optimal player” would undertake. The recording of both the ideal path and all actual paths can be implemented using the *Experience API* (xAPI) data format.

## 2.2 Ideal Path Score

The IPS supports the computation of user progress. The score is normalized to  $[-1; 1]$  to be invariant of varying game genres or different users. A value of  $IPS = 1$  means a perfect game move (congruent with the ideal path); a value of  $IPS = 0$  is a game move without significant progress; and  $IPS = -1$  is a degrading game move (negative progress), e.g., moving in the utter opposite direction. For games with continuous movements IPS could be in  $\{x|x \in R, -1 \leq x \leq 1\}$ . While the Ideal Path Model is generic and can be modeled game independent, the IPS and its metric are typically game specific. For example, for step-by-step games this could be a string similarity distance; or for a 3D shooter-type game the metric could be a distance between waypoints. For our game, a 2.5D seek-and-find game, the metric is the euclidean distance between optimal and actual direction.

## 3 Application

The seek and find game *SaFIRa* [4] has been extended with an eye tracking plugin and the IPS (Fig. 2). The game itself is implemented with the game engine Unity. The adaptivity for SaFIRa has been realized with the “E-Learning A.I.” (ELAI) adaptivity framework [4]. The ELAI’s interpretation engine and heuristic adaptivity score computation (a weighted linear equation formula with so called Didactic Factors) has been extended by the IPS as a new factor. The concept has been successfully implemented. Preliminary evaluation results indicate an improvement of the adaptive behavior.

## 4 Summary

The presented *Ideal Path Model* and its linked *Ideal Path Score* (IPS) enable attention-driven adaptivity for serious games. The IPS can be used for more precise estimations of players’ performance and their need for adaptive assistance. It targets the problem statement of when an adaptive system should



**Fig. 2.** Eye-tracking and the *Ideal Path Score* applied to adaptivity for a serious game.

engage. This is of particular importance in games where interrupting the players' flow experience should be avoided to keep up the immersion. The realized IPS-adaptive game assesses performance and attention of players based on eye movements and interactions with the game. In future work an evaluation will target multiple hypotheses, including e.g., correlation between measured attention or goal-orientedness and related subjective answers of study participants; or influence of eye tracking on IPS evaluation and on adaptivity.

**Acknowledgments.** The underlying project to this article is funded by the Federal Office of Bundeswehr Equipment, Information Technology and In-Service Support under promotional references. The authors are responsible for the content of this article.

## References

1. Chen, J.: Flow in games (and everything else). *Commun. ACM* **50**(4), 31–34 (2007). <http://doi.acm.org/10.1145/1232743.1232769>
2. Duchowski, A.: Eye tracking methodology: Theory and practice. pp. 1–328 (2007)
3. Streicher, A., Roller, W.: Towards an interoperable adaptive tutoring agent for simulations and serious games. In: *International Conference on Theory and Practice in Modern Computing, MCCSIS 2015*, pp. 194–197. IADIS Press (2015)
4. Streicher, A., Roller, W.: Interoperable adaptivity and learning analytics for serious games in image interpretation. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017. LNCS*, vol. 10474, pp. 598–601. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_71](https://doi.org/10.1007/978-3-319-66610-5_71)
5. Streicher, A., Smeddinck, J.D.: Personalized and adaptive serious games. In: Dörner, R., Göbel, S., Kickmeier-Rust, M., Masuch, M., Zweig, K. (eds.) *Entertainment Computing and Serious Games. LNCS*, vol. 9970, pp. 332–377. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46152-6\\_14](https://doi.org/10.1007/978-3-319-46152-6_14)



# Development of a Learning Economy Platform Based on Blockchain

Masumi Hori<sup>1</sup> , Seishi Ono<sup>1</sup> , Toshihiro Kita<sup>2</sup> , Hiroki Miyahara<sup>3</sup> ,  
Shiu Sakashita<sup>4</sup> , Kensuke Miyashita<sup>5</sup> , and Kazutuna Yamaji<sup>6</sup> 

<sup>1</sup> NPO CCC-TIES, Nara, Nara, Japan  
{hori,ono}@cccties.org

<sup>2</sup> Kumamoto University, Kumamoto, Kumamoto, Japan

<sup>3</sup> University of Yamanashi, Kofu, Yamanashi, Japan

<sup>4</sup> Acutus Software, Inc., Tokyo, Japan

<sup>5</sup> Kyoto Women's University, Kyoto, Kyoto, Japan

<sup>6</sup> National Institute of Informatics, Chiyoda, Tokyo, Japan

**Abstract.** Knowledge-based societies require more than just the accumulation of existing knowledge. Knowledge must be continuously updated and new knowledge must be created. In this paper, we propose a learning economy model that incorporates market economy principles. The model contributes to the production, dissemination, and monetization of new knowledge based on a blockchain technology. This study aims to propose the learning economy model and an example of implementing this model.

**Keywords:** Blockchain · Content capsule · E-book · Learning economy  
Online education

## 1 Introduction

Over the last half-century, there has been a paradigm shift in the society. The industrial society has become the knowledge-based society [1]. In this new paradigm, the capital represented by knowledge is recognized as a key resource in social development. At the same time, however, this knowledge capital has rapidly become obsolete. Knowledge-based societies, therefore, require more than just the accumulation of existing knowledge. Knowledge must be continuously updated, and new knowledge must be created [2].

The objective of our study was to realize a system in which citizens can constantly acquire new knowledge and update their existing skills, to allow them to function effectively in the knowledge-based society.

## 2 Learning Economy Model

### 2.1 Information, Knowledge, and Learning

In fields such as cognitive science and knowledge management, the words “information,” “knowledge,” and “learning” have several different meanings. We adopted the

Brookes definition. Brookes [3] defined self-knowledge that is being internalized as  $s$  and the knowledge structure of  $s$  as  $k[s]$ . In Brookes’s model,  $k[s]$  is updated to  $k[s + \Delta s]$  by the action of information  $\Delta i$ . This yields the following equation:

$$k[s] + \Delta i = k[s + \Delta s]. \tag{1}$$

The knowledge structure  $k[s]$  to which the information  $\Delta i$  can be added to yield  $k[s + \Delta s]$  is also information. Finally, we define “learning” as the generation of new information when  $\Delta i$  acts on the knowledge structure  $k[s]$  from Eq. (1).

### 2.2 Flow of Learning Economy

In the learning economy model, market economics drive the circular flow of learning.

Knowledge plays the role of a resource in the model, and two markets are expected to emerge: one for information fragments and one for systemized information. The flow of learning can be described using the SECI model [4]. This is shown as Fig. 1.

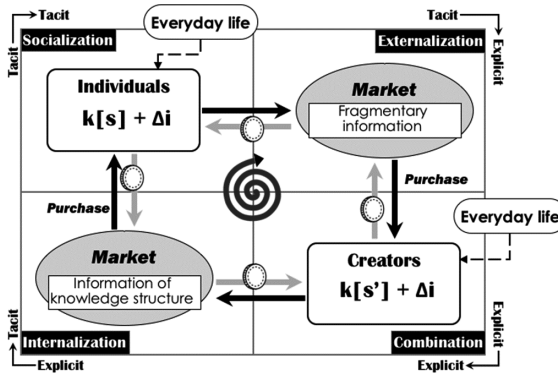


Fig. 1. Flows within the learning economy model.

In the learning economy model, transactions involve exchanges of knowledge, in which both the recipient and the creator add information  $\Delta i$  to their knowledge structure  $k$ , with the goal of creating and launching new knowledge onto the market. In this paper, we call those who create and release new fragments of information an individual and those who combine the information to form systemized knowledge a creator. Both can create new knowledge from the fragments of information acquired in everyday life, or by combining fragments of knowledge purchased in the marketplace.

Systemization also creates new knowledge, so that a flow is created. In the learning flow, participants share a common belief that acquiring, updating, and creating new knowledge is a valuable process and that benefits accrue to those involved in the activity.

To allow information to be traded and to establish a market, information must be a product. This is done by introducing a “knowledgization” capsule, which can encapsulate any information that is available in a digital form. The capsules transfer knowledge

to the other participants, who are able to reconstruct knowledge and to create a capsule enclosing the new information.

### 3 The Proposed System

#### 3.1 Capsule Technology

The “knowledgezied” capsule is based on a technology developed in [5]. It can encapsulate multiple open resources and tools, including videos, quizzes, and live functions, and offer them as learning materials.

The capsule is created in EPUB3 format, as this is a widely used e-book standard, allowing the capsules to be distributed through e-book stores and to be accessed using an e-book reader. An embedded output function allows them to be read on a web browser.

The capsule comprises a metadata part and an engine part. The metadata part uniquely locates the resource on the web and defines the display of the metadata in a structured or systemized form. These metadata are, respectively, called the Metadata of Resources (MR) and Metadata of Resource Configuration (MRC). The engine embeds an INPUT and OUTPUT. The INPUT converts the MR and MRC, whereas the OUTPUT interprets the metadata and displays the resources on a web browser.

#### 3.2 Demonstration System of Learning Economy Model

We developed a demonstration system of a learning economy model that make a deal with “knowledgeziation” using blockchain as a virtual currency with smart property and social network service (SNS) as a user interface. In this system, we used Hyperledger Fabric as blockchain and Mastodon (<https://github.com/tootsuite/mastodon>) as SNS, which are both offered as open source. The demonstration system was designed to create “knowledgeziation” capsules and trade them using a blockchain.

We have demonstrated the implementation of knowledge encapsulation and have shown that the capsules can be used in transactions.

The demonstration system confirmed that transactions occur only when the story posted to Mastodon is output in an EPUB3 format as a “knowledgeziation” capsule, so that value has been added. However, the format of the “knowledgeziation” capsule needs further development. The demonstration capsule technology used the EPUB3 format. However, this format places restrictions on the scripts and media that can be embedded. A new format may need to be developed.

### 4 Conclusions

The learning economy model offers three main benefits.

First, the learning economy model enables individuals or creators to circulate new knowledge by combining fragments of the information acquired in the course of everyday life with the information available in the market.



Second, the use of market economy principles provide incentives to learners, guarantees an autonomously flowing learning environment, and enable the needs of learners and society to be addressed.

Finally, knowledge can be evaluated. By introducing a virtual currency, a common value is placed on the shared knowledge.

The learning economy model that was proposed in this paper is designed to allow anyone in the society to participate in the creation and distribution of new knowledge. In contrast with traditional schooling, no tuition fees need to be paid. Lifetime learning can be sustained by launching new information onto the market, making the proposed system a form of open education.

We implemented and demonstrated a system for creating and trading “knowledge-ization” capsules, based on stories submitted to a specified SNS. The same mechanism could be used to enable transactions based on information collected from the web. By allowing value to be added, an ecosystem that can provide a new foundation for learning is created.

**Acknowledgments.** This work was supported by JSPS KAKENHI Grant Number JP7H01844 and NII Joint Research Grant. The authors also would like to thank Enago ([www.enago.jp](http://www.enago.jp)) for the English language review.

## References

1. Drucker, P.: *The Age of Discontinuity: Guidelines to Our Changing Society*. Warner Books, New York (1988)
2. Lundvall, B.Ä., Johnson, B.: The learning economy. *J. Ind. Stud.* **1**, 23–42 (1994)
3. Brookes, B.C.: The foundations of information science. Part I *Philos. Asp. J. Inf. Sci.* **2**, 125–134 (1980)
4. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2009). <https://bitcoin.org/bitcoin.pdf>
5. Hori, M., Ono, S., Kobayashi, S., Yamaji, K., Kita, T., Yamada, T.: Our microcontent approach: learning content encompassing learning platform. In: Poster Session, Learning with MOOCs III, Philadelphia, 6 October 2016



# Enhancing Human Learning of Motions: An Approach Through Clustering

Quentin Couland<sup>(✉)</sup>, Ludovic Hamon<sup>(✉)</sup>, and Sébastien George<sup>(✉)</sup>

Laboratoire d'Informatique de l'Université du Mans,  
LIUM - EA 4023, Le Mans Université,  
Avenue Olivier Messiaen, 72085 Le Mans Cedex 9, France  
{quentin.couland,ludovic.hamon,sebastien.george}@univ-lemans.fr

**Abstract.** More and more software applications use human motions to improve the information retention. Some virtual environments are especially built to support the learning of human motions. However, these kinds of applications and their pedagogical feedback are rarely made from the analysis of 3D captured motions. This can be explained by the heterogeneity, the complexity and the high-dimensional nature of such data. However, machine learning techniques could be used to overcome these issues. This paper presents a first step towards the improvement of the human learning process of a motion, thanks to the analysis of clusters representing user profiles. In the context of the Bottle Flip Challenge and using raw captured motions, descriptors based on speed and acceleration are extracted. The motions are then automatically analyzed, according to two different approaches: one with the ground truth, and one without constraints on the number of clusters. The results suggest that the data are separable using the computed descriptors.

**Keywords:** Human motion · Human learning · Machine learning  
Clustering

## 1 Related Work

Captured motions have been used in various fields related to human learning of motions. Indeed, it is possible to extract cinematic and dynamic data from low level raw data (*i.e.* the evolution of joints position through time) [2,4]. These methods require an expert to analyze the data and give a feedback to the learner. Some works used supervised and unsupervised algorithms to analyze facial expressions, movements and actions. Among them 3D captured data were studied with a set of expert rules relating to the learner displacement [1]. Although these approaches are efficient, the motions do not require a cognitive effort in terms of human learning. Furthermore, the goal was not to evaluate the degree of learning and/or success of the motion, and the descriptors could not be used to give a pedagogical feedback.

The use of supervised machine learning algorithms assumes that: (a) labeled data exist for the specific problem, and (b) there is a sufficient amount of data to

allow the algorithm to converge. In practical cases, (a) is often in contradiction with (b). Furthermore, the results are not always explainable as the descriptors can be modified through the algorithm (*e.g.* using PCA), and the separation model is not humanly interpretable, or with difficulty (*e.g.* SVN or Neural Networks). Unsupervised learning approaches have other requirements and, in a human learning context, could help the teachers to regroup the learners in different clusters identified by their observation needs. This would allow (i) seeing if recurring behaviors appear regarding to these needs, and (ii) adapting the learning process to each cluster *i.e.* learner profile.

## 2 Motion Analysis: A Clustering Based Approach

Establishing the relevant features that allow telling if a motion is successful (or not) is a non-trivial task. For a given task, there is not one or several perfect motions. The use of supervised machine learning algorithms assumes that large databases of labeled motions exist; yet, there are very few specialized database containing the same kind of motions with several degrees of success and that requires a cognitive effort in terms of human learning. Taking into account all of these elements, the automatic analysis of motions through clustering techniques is the chosen approach.

### 2.1 Protocol

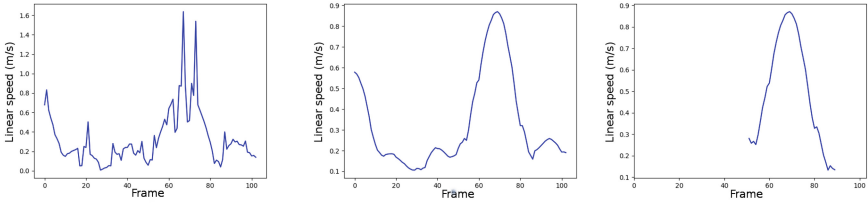
A database made of non-trivial motions to learn (in terms of human learning) was created. For this study, the Bottle Flip Challenge<sup>1</sup> was chosen as the learning task. This task requires some dexterity, and the execution time is short. To capture the motions, a MOCAP suit based on Inertial Measurement Units (IMU) was chosen<sup>2</sup>. Figure 1a shows the hand's captured motion. One can see that there are a lot of artifacts, due to the capture system. These data were filtered using the Savitzky-Golay filter, in order to eliminate the errors and noises related to the MOCAP system (Fig. 1b). For each motion, the throwing part of the motion was automatically segmented (Fig. 1c), then rebuilt with a fixed number of frames. The speed and acceleration data, as well as the corresponding directions along the x, y and z axis, were extracted from this rebuilt motion from: (i) the beginning of the throw, (ii) the maximum speed value and (iii) the end of the throw.

## 3 Results

The tests were made on a set of 13 people's data, consisting of 1300 throws in total. Different sets of joints have been used: hand (H), forearm (FA), arm (A), shoulder (S), these body parts being the most solicited during the motion.

<sup>1</sup> [https://en.wikipedia.org/wiki/Bottle\\_flipping](https://en.wikipedia.org/wiki/Bottle_flipping).

<sup>2</sup> <https://neuronmocap.com/>.



**Fig. 1.** (a) Speed of the captured motion through time of the right-hand of an user (b) Initial speed filtered (c) Extracted throwing part.

There are two main objectives here: (i) determine if the data can be partitioned and (ii) determine if it’s possible to obtain a clustering based on the degree of success of the task.

### 3.1 Cohesion and Separation of the Clusters

This approach was based on the hypothesis that there are different types of motions that can be gathered in separable clusters. In this context, the computed metric is the Average Silhouette Score (ASS) [6]. The Silhouette Score (*SS*) gives a value indicating how a sample is well-fitted to the assigned cluster, compared to other clusters. The Average Silhouette Score (*ASS*) is the mean of every sample’s *SS*. This value ranges from  $-1$  to  $1$ , with  $1$  indicating that in average, every sample best belongs to their cluster, and  $0$  indicating that the clusters are overlapping. An *ASS* above  $0.5$  indicates that a reasonable structure is found in the data, while an *ASS* above  $0.7$  indicates that a strong structure is found [7]. Table 1 shows the results for the five best data combinations, with the highest *ASS* score being  $0.7038$ , suggesting that our data are separable.

**Table 1.**  $max(ASS)$  (for  $k$  varying from 2 to 10) for various joints and data combinations.

Data type	H	H, FA	H, FA, A	H, FA, A, S
BegMaxEnd Speed[x/y/z]	<b>0.7038</b> ( $k = 2$ )	0.6734 ( $k = 2$ )	0.6677 ( $k = 2$ )	0.6650 ( $k = 2$ )
BegMaxEnd Speed Norm	<b>0.5147</b> ( $k = 2$ )	0.3992 ( $k = 2$ )	0.3869 ( $k = 2$ )	0.3803 ( $k = 2$ )
BegMaxEnd Speed Norm,Dir[x/y/z]	<b>0.3355</b> ( $k = 2$ )	0.3271 ( $k = 2$ )	0.2665 ( $k = 2$ )	0.2496 ( $k = 2$ )

### 3.2 Ground Truth Approach

The second hypothesis relies on the fact that each cluster corresponds to a success degree of the task. The k-means algorithm was run with  $k = 2$  and  $k = 3$ , on the data. The clusters were then analyzed, in order to verify their contents in terms of motions leading to a “successful/failed” throw for  $k = 2$

and leading to a “successful/almost successful/failed” throw for  $k = 3$ . For this experiment, the Adjusted Rand Index (ARI) was chosen [5], to verify if the obtained labeling was similar to the ground truth. This metric is a measure of the similarity between two data partitioning. This approach yielded low scores ( $ARI \approx 0$ ), indicating that the labeling did not match the ground truth, showing that the speed combined to the partitioning strategy is not an indication of the degree of success of the task.

## 4 Discussion and Future Work

An approach of 3D captured motion analysis based on clustering was proposed. The goal is to assist the learner in their motion learning task, by allowing the expert to analyze the learner’s motion, and giving them a way to adapt the learning process. This approach was based on two hypotheses: (i) it is possible to find an explainable partitioning of the data, and (ii) it is possible to automatically separate the motions based on the degree of success of the task. Results shown in Sect. 3.1 suggest that the combination of the speed vectors on each axis is a good separation criterion. Having the best *ASS* values for the hand shows that the hand’s descriptors were the most significant. In addition, the most discriminant features were the speed at the throw moment, in both forward and up direction (regarding to the body of the person throwing), in terms of relative distance: 2.82 and 2.78, respectively (between 0.04 to 0.3 for the other speed values). Consequently, the motions were indeed separable, which validated the chosen indicators in terms of discriminant features. However, the results in Sect. 3.2 show that the *ARI* value is close to 0 in every case, suggesting a random data assignment. Hence, the current descriptors, as well as the considered approach, seem to be irrelevant to assess the degree of success of the performed motion. Considering the chosen features, it seems that the considered task does not have a significant variation. The chosen descriptors are low-level ones (kinematic/dynamic) [2], and using higher-level ones could allow separating the motions on more meaningful features, thus allowing a better analysis. The use of Dynamic Time Warping algorithm, computing a distance between time series [3], would provide another similarity measures between the motions. Performing recursive clustering on well separated clusters is another lead, as it could allow to determine more accurate learner profiles.

## References

1. Hachaj, T., Ogiela, M.R.: Full body movements recognition - unsupervised learning approach with heuristic R-GDL method. *Digit. Signal Process.* **46**, 239–252 (2015)
2. Larboulette, C., Gibet, S.: A review of computable expressive descriptors of human motion. In: *Proceedings of the 2nd International Workshop on Movement and Computing, MOCO 2015*, pp. 21–28. ACM, New York (2015)
3. Morel, M.: *Multidimensional time-series averaging: application to automatic and generic evaluation of sport gestures*. Theses, Université Pierre & Marie Curie, November 2017. <https://hal.archives-ouvertes.fr/tel-01659753>

4. Nunes, J.F., Moreira, P.M.: Handbook of Research on Computational Simulation and Modeling in Engineering (2016)
5. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)
6. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
7. Struyf, A., Hubert, M., Rousseeuw, P.: Clustering in an object-oriented environment. *J. Stat. Softw.* **1**(4), 1–30 (1997)



# How to Help Teachers Adapt to Learners? Teachers' Perspective on a Competency and Error-Type Centered Dashboard

Iryna Nikolayeva<sup>1</sup>(✉), Bruno Martin<sup>1</sup>, Amel Yessad<sup>1</sup>, Françoise Chenevotot<sup>2</sup>, Julia Pilet<sup>2</sup>, Dominique Prévité<sup>1</sup>, Brigitte Grugeon-Allys<sup>2</sup>, and Vanda Luengo<sup>1</sup>

<sup>1</sup> Modèles et Outils en ingénierie des Connaissances pour l'Apprentissage Humain (MOCAH), Laboratoire d'Informatique de Paris 6 (LIP6), Sorbonne Université, Paris, France

[iryna.nikolayeva@lip6.fr](mailto:iryna.nikolayeva@lip6.fr)

<sup>2</sup> Laboratory of Didactics André Revuz (LDAR), Universities Artois, Cergy-Pontoise, Paris Diderot, Paris Est Créteil et Rouen, Paris, France

**Abstract.** The main research goal of this paper is to reveal what information helps teachers adapt to students within a dashboard, how they use it, and how to provide better support. In this research, we observe information acquisition by teachers through a created stand-alone dashboard and interviews. The dashboard presents not only exercise performances and competency level of acquisition, but also error-type information. The analysis of these observations uncovers a first conceptual model of teacher adaptation workflow, as well as additional suggestions to ease adaptation in a future version of such a dashboard.

**Keywords:** Dashboard · Adapt · Differentiate · Diagnose · Error type

## 1 Introduction

Throughout their career teachers need to diagnose difficulties of their students and then adapt the teaching to their needs, which is very cognitively demanding. One emerging means that helps teachers diagnose students' levels and tackle their difficulties are learning dashboards [6]. In a recent paper, Xhakaj and colleagues underline the lack of error information in existing dashboards [7]. We here created a dashboard including error-type information along with competencies and success rates. We then interviewed teachers and developed a first conceptual model of how teachers use this information.

## 2 Student Diagnostic Test

Previous studies typically associate Knowledge Components to exercises, and then determine whether they are mastered based on the validity of the

---

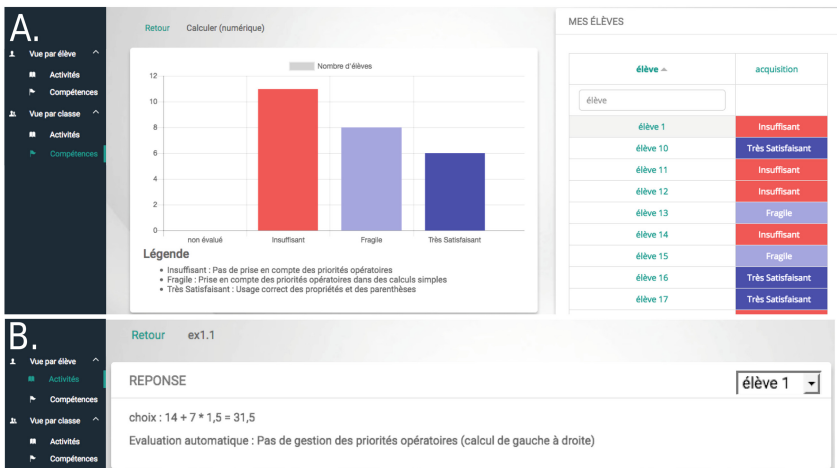
Supported by the French Ministry of Education.

response [1]. In our case, we use a knowledge diagnosis based on results of didactics of algebra for middle school students (12–16 years old) [2,4]. The diagnosis is a two-step process: first, for each exercise, the answer of each student is analysed in terms of validity and algebraic reasoning via error interpretation. Then, an individual diagnosis evaluates the coherency of responses to all exercises and builds the cognitive profile of the student in algebra.

## 3 Methods

### 3.1 Dashboard Design

We used user-centered design approach: From interviews with 19 teachers, we built personas and use scenarios. We first designed our prototype with the software JustInMind, then evaluated it on six other teachers using the “think aloud” method. We used as input the standardized xAPI format [3] with an xAPI profile specific to this diagnosis. The dashboard itself has four main views: class-level or student-level information, each subdivided in exercise and competency views. The exercise view contains the success rates per exercise, and links to answers of each student, along with an automatic diagnosis of a potential error. The competency views contain level of mastery of each competency (Fig. 1).



**Fig. 1.** A. Class view, detail for competency numerical calculus accompanied with student by student details. (Translation of the legend: Insufficient: Do not take into account priorities; Fragile: Take into account priorities in simple calculations; Very Satisfying: Correct usage of priorities and parentheses.) B. Student view, detail for exercise 1.1: Student’s 1 response and error type evaluation (Translation: No management of priorities of operations (calculus from left to right)).

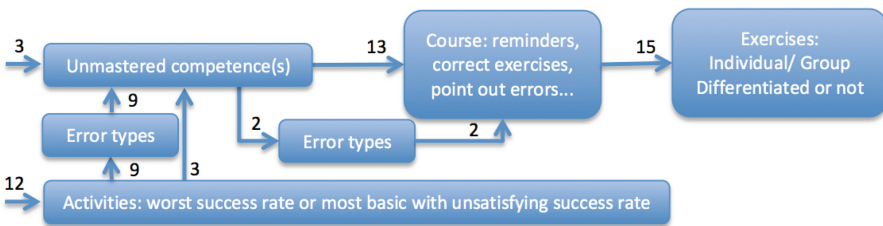


### 3.2 Interviews and Questionnaires

We questioned 15 mathematics middle school teachers, recruited within a training programme for differentiated teaching in Créteil academy, France [5]. Students took the diagnostic test in elementary algebra from September to November 2017. For each teacher, we then conducted from December to March 2018 a one-hour screen- and voice-recorded interview, accompanied by pre- and post-interview questionnaires. The pre-questionnaire aimed at recording general information. The post-questionnaire gathered the teacher’s opinion about the dashboard. During the interview, the researcher asked to plan the next session after the diagnostic test, and presented the dashboard. The teacher manipulated results of the diagnostic test of his/her class and prepared the plan. Finally the researcher questioned the teacher. We transcribed the interviews, and analyzed them. Common patterns of usage and critics were highlighted, and organised.

## 4 Results

All but two teachers claimed that they would gain in reactivity and adapt more to groups and individuals if a dashboard as this one was available. Figure 2 presents the model that sums up how teachers used the dashboard. Teachers always started by observing the class view, either by competencies or by exercises when competencies did not evoke precise meaning (they had only been introduced one year earlier by a national reform). In total, 11 out of the 15 teachers viewed errors. They claimed that this allowed them to better understand students’ reasonings and provide a more adapted remediation for the given competence. For instance, when the competency numerical calculation is not mastered, and the error, as in Fig. 1B, shows that the student does not take into account priorities, it clarifies what to work on within this competency.



**Fig. 2.** Model of the adaptation process via the dashboard: problem identification and resolution. The numbers correspond to the number of teachers who followed the corresponding edge.

To easier chose the best remediations for groups of students, teachers asked for histograms of most frequent error types and for time evolution of performances on different tests. Finally, the wordings of competencies required clarification. Teachers’ main practical constraint to adaptation being lack of time,

a database of associated exercises, classified by competencies and by levels, and associated step-by-step corrections was the most requested addition. In conclusion, our observations allowed us to confirm that error analysis, completed with exercise and competency views, was widely used for remediation. Improvements to the dashboard were suggested.

## 5 Conclusion and Discussion

In this paper, we present a dashboard that shows class and student-level summaries of results based on exercises, competencies and errors. We suggested a common model of data usage within the dashboard to diagnose difficulties and adapt the course, that teachers, we interviewed welcomed. We here confirm the suggestion of [5,7]: error types complement skill mastery levels to help teachers precisely diagnose and adapt. Our model of the dashboard usage for diagnosis and adaptation underlines the usefulness of automatic error analysis and shows how teachers use it, thereby expanding the general timeline proposed in [8].

## References

1. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction Based on Adaptive Learning Technologies. Handbook of Research on Learning and Instruction. Routledge, London (2016)
2. Chenevotot, F., Grugeon, B., Pilet, J., Prévité, D., Delozanne, E.: The diagnostic assessment Pépité and the question of its transfer at different school levels. In: The Ninth Congress of the European society for Research in Mathematics Education (CERME) (2015)
3. Corbett, A.: Cognitive computer tutors: solving the two-sigma problem. In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds.) UM 2001. LNCS (LNAI), vol. 2109, pp. 137–147. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44566-8\\_14](https://doi.org/10.1007/3-540-44566-8_14)
4. Grugeon-Allys, B., Chenevotot-Quentin, F., Pilet, J., Prévité, D.: Online automated assessment and student learning: The *PEPITE* project in elementary algebra. In: Ball, L., Drijvers, P., Ladel, S., Siller, H.-S., Tabach, M., Vale, C. (eds.) Uses of Technology in Primary and Secondary Mathematics Education. ICME-13, pp. 245–266. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-76575-4\\_13](https://doi.org/10.1007/978-3-319-76575-4_13)
5. Pilet, J.: Réguler l'enseignement en algèbre élémentaire par des parcours d'enseignement différencié. *Recherches En Didactique Des Mathématiques* **35**(3), 273–312 (2015)
6. Schwendimann, B.A., et al.: Understanding learning at a glance: a systematic literature review of learning dashboards. In: Proceedings of the 6th International Conference on Learning Analytics and Knowledge (LAK 2016), vol. 10, no. 1, pp. 148–157 (2016)
7. Xhakaj, F., Aleven, V., McLaren, B.M.: How teachers use data to help students learn: contextual inquiry for the design of a dashboard, pp. 340–354 (2016)
8. 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)



# MedSense: The Development of a Gamified Learning Platform for Undergraduate Medical Education

Justin Choon Hwee Ng<sup>1(✉)</sup>, Sarah Zhuling Tham<sup>1</sup>, Chin Rui Chew<sup>2</sup>,  
Amelia Jing Hua Lee<sup>2</sup>, and Sook Muay Tay<sup>1,3</sup>

<sup>1</sup> Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore  
justinnch@gmail.com

<sup>2</sup> School of Information Systems, Singapore Management University, Singapore, Singapore

<sup>3</sup> Department of Anaesthesiology and Perioperative Medicine, Singapore General Hospital,  
Singapore, Singapore

**Abstract.** Educational technology offers compelling educational possibilities for learners and educators: providing access to online databases, enabling simulation, and facilitating collaborative learning. Technology is rapidly being incorporated within traditional teaching methods to engage the millennial learners accustomed to technology-enhanced learning. Following the success of gamification strategies in medical education such as SonoGames in educating Emergency Medicine residents about the use of point-of-care ultrasound, we aimed to create a gamified learning platform for undergraduate medical education. MedSense is a collaborative gamified learning platform, where students can trial faculty-vetted case-based simulations and share interesting cases with their near-peers. The application is designed to benefit the learning of students and curriculum development by educators based on student performance. Amongst its key features are the case upload panel, game element, free response marking algorithm, recommendation panel and the analytics dashboard. MedSense is currently a prototype under development, which has been well-received by users. Given the success of gamification in other settings, we hope to reiterate the benefits of this education strategy with the development of MedSense.

**Keywords:** Gamified learning platform · Personalised feedback  
Collaborative learning

## 1 Pedagogical Background

The advancement of education technology tools offers to educators and students an extensive variety of educational possibilities. Current technology allows learners today to access a vast amount of knowledge in online databases and electronic textbooks, enables simulation of real-life scenarios to provide a safe environment for the practice of skills and facilitates collaborative learning using online sharing platforms. Technology is rapidly being incorporated within traditional teaching methods to engage the millennial learners accustomed to technology-enhanced learning [1].

Gamification refers to the concept of applying game design elements to traditionally non-game contexts [2]. Gamification strategies are increasingly being utilised in traineeship programs of various specialties, both surgical and medical. [3] For instance, SonoGames [4] is an annual event held during the Society of Academic Emergency Medicine meeting, which is a game-based event to educate residents about the use of point-of-care ultrasound and to boost their confidence in using such technology in clinical practice. Additionally, games designed for medical education have also been developed. Microbe Invader is one such example: a role-playing game designed to teach clinical microbiology. [5] With multiple success stories in the application of gamification strategies in medical education, we aimed to create a gamified learning platform for undergraduate medical education.

## 2 Technical Solution and Complexity

### 2.1 Application Overview

MedSense is a collaborative gamified learning platform where students can trial faculty-vetted cases and share interesting cases with their near-peers. It is hosted on Amazon Web Services and accessible via the domain <https://www.themedsense.com>.

MedSense is designed to assist students in making a smooth transition to clinical attachments by enabling them to practise clinical reasoning using case-based simulations. It also helps them to understand their personal strengths and weaknesses through analytic dashboards and encourages learning in a relaxed and fun environment. For educators, it provides valuable information on common weaknesses amongst students to guide course planning.

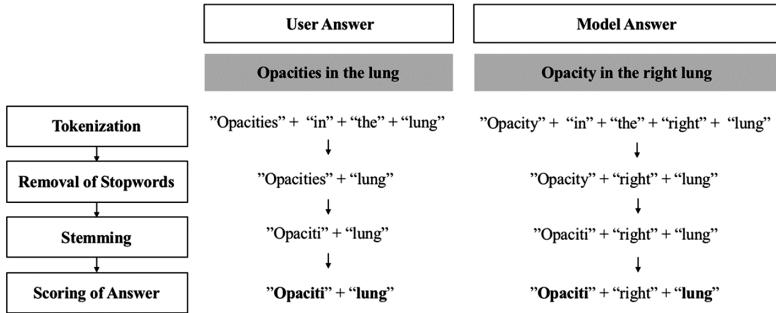
### 2.2 Application Features

**Case Upload.** MedSense features an intuitive case upload page, accessible by both students and educators. Cases uploaded by students will be vetted by faculty members to ensure factual accuracy. As a form of incentivisation, titles are conferred to all contributors and top contributors are featured on the contributor leader board.

**Gamification.** MedSense introduces an element of fun and friendly competition to learning. As students attempt cases, they gain experience (XP) and level up. The leveling system is based on the popular board game “Dungeons and Dragons” [6], which is a common reference in the development of many gaming platforms. Students level up by obtaining a set amount of XP and for each increasing level, the amount of XP required increases based on an exponential function. In this way, the game continues to remain challenging yet rewarding. Top players are featured on the player leader board.

Mechanisms have been instituted to prevent students from exploiting the game mechanics to gain levels unfairly. The total amount of XP gained is halved for each subsequent gameplay for the same case. In other words, if 4 XP were awarded for the first attempt, 2 XP will be awarded for the next and so on. Fundamentally, this prevents students from levelling up quickly by repeatedly playing the same case.

**Marking Algorithm.** MedSense differentiates itself from existing online platforms with the ability to perform automated marking of free response answers. The capability to do so is attributed to the use of Natural Language Processing (NLP)<sup>1</sup> techniques. The marking algorithm is a four-step process, as shown in Fig. 1.



**Fig. 1. MedSense Marking Algorithm.** The algorithm analyzes the answers in a four-stage process and generates a score. In this example, a score of 66% is awarded.

First, the student’s answer is broken down into individual words. Stop words, which are commonly used words like “and” and “the”, are identified and removed. Stemming is subsequently applied using the Porter’s Algorithm<sup>2</sup>, reducing the remaining words to their base forms. Finally, the remaining words are compared with the processed model answer. The score is derived by dividing the number of matched words by the total number of words in the processed model answer.

To test the reliability of the marking algorithm, the scores generated by the system are compared with the scores derived from manual scoring. On average, the accuracy of the system scoring is evaluated to be about 70%.

**Recommendation System.** Another feature of MedSense is its ability to provide recommendations individualized to the students’ needs on their respective homepages. The algorithm takes into account various factors in providing its recommendations.

First, it evaluates the performance of the student in the cases he or she has attempted in the various specialties and prioritizes cases in the specialty he or she has not performed well in. This is meant to encourage the student to attempt more cases in that particular specialty so that he or she can improve in his or her area of weakness. Recommendations are also based on the academic year the student is in: beginner cases for the pre-clinical students and advanced cases for the more senior. Finally, MedSense recommends cases by popularity based on the frequency of attempts by all users. Cases which are more frequently played are presumed to be more useful for users and are hence prioritized in the list of recommended cases.

<sup>1</sup> NLP is a form of artificial intelligence, enabling computers to analyze, understand and derive meaning from human language in order to later organize into structured knowledge [7].

<sup>2</sup> Porter’s stemming algorithm is first described by Porter et al. in 1979 as a process for removing common suffixes from words in English [8].

**Analytics Dashboard.** The analytics dashboard provides students with a broad overview of their performance in the cases they have attempted thus far. The dashboard provides useful infographics, breaking down students' scores by question and offering a comparison with the global average score of all other students who have attempted the same case. Alternatively, students may view their performance scores by specialty and subspecialty, enabling students to better understand their weaknesses and to spend more time and effort in improving their knowledge in these areas.

### 3 Case Demonstration

The case demonstration features a case titled "A Good Samaritan". We invite the audience to trial the case to experience timed case-based simulation. Special attention should be paid to (a) the types of question featured in the case – multiple choice questions, extended matching questions and free response questions, with or without image attachments; (b) the immediate feedback and score provided by the marking algorithm; and (c) the summary of their performance at the end of the case and in their respective analytic dashboard. Thereafter, making use of real-time information gathered through the attempts made by the audience, we will demonstrate how students and educators may utilize the analytic dashboard to enhance learning.

### 4 Conclusion

MedSense is a prototype of a multi-phase project. Moving forward, we intend to enhance the element of simulation by designing a role-playing platform, where users will play as medical students, and later doctors, to tackle various admissions to the hospital. Their decisions will lead to different outcomes, ranging from a debilitated patient who may not make it to a satisfied and well patient on his or her way to discharge. Through simulation, we hope to equip medical students with the ability to make critical decisions in time-sensitive settings and better prepare them for clinical practice.

In summary, given the success of gamification in other educational settings, we developed a gamified learning platform for undergraduate medical education. We hope to reiterate the benefits of this education strategy.

**Acknowledgements.** We like to extend our heartfelt gratitude to the MedSense team for their time and effort in developing the application. This would not have been possible without them.

### References

1. AAMC Institute for Improving Medical Education: Effective Use of Educational Technology in Medical Education: Colloquium on Educational Technology: Recommendations and Guidelines for Medical Educators. <https://members.aamc.org/eweb/upload/Effective%20Use%20of%20Educational.pdf>. Accessed 29 Apr 2018
2. Kapp, K.M.: *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Pfeiffer, San Francisco (2012)

3. Rutledge, C., et al.: Gamification in action: theoretical and practical considerations for medical educators. *Acad. Med.* **93**, 1014–1020 (2018)
4. Liteplo, A.S., Carmody, K., Fields, M.J., Liu, R.B., Lewiss, R.E.: SonoGames: effect of an innovative competitive game on the education, perception, and use of point-of-care ultrasound. *J. Ultrasound Med.* (2018)
5. Li, T.: Microbe Invader. <https://www.microbeinvader.com>. Accessed 24 June 2018
6. Howtomakeanrpg: How to Make an RPG: Levels. <http://howtomakeanrpg.com/a/how-to-make-an-rpg-levels.html>. Accessed 29 Apr 2018
7. Algorithmia: Introduction to Natural Language Processing (NLP). <https://blog.algorithmia.com/introduction-natural-language-processing-nlp/>. Accessed 29 Apr 2018
8. Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)



# edCrumble: Designing for Learning with Data Analytics

Laia Albó  and Davinia Hernández-Leo 

ICT Department, Universitat Pompeu Fabra, Barcelona, Spain  
{laia.albo, davinia.hernandez-leo}@upf.edu

**Abstract.** This demonstration introduces ILDE2/edCrumble, an online learning design platform that allows teachers the creation of learning designs (LDs) with the support of data analytics. ILDE2/edCrumble is built on top of the LdShake platform, which provides social features enabling the sharing and co-edition of LDs. The tool provides an innovative visual representation of LDs combining face-to-face and online learning in different places (in-class and out-of-class) and times (synchronous and asynchronous). Decision making during the LD process is supported by two types of analytics: resulting from the design of the activities sequenced in a timeline (LD analytics); and aggregated meta-data extracted from several grouped LDs (community analytics). Preliminary results conducted as part of an iterative design-based research process, show that the tool is being perceived as easy to use and useful. During the demo we will show the use case of how LD and community analytics can help balancing the workload and design between different courses which are part of a whole curriculum.

**Keywords:** Authoring tool · Learning design · Data analytics  
Communities of educators · Visualization · Pedagogical planner · edCrumble  
ILDE2 · LdShake

## 1 Introduction

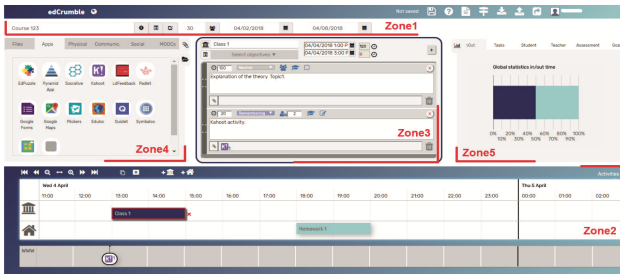
For some time now, Learning Design (LD) tools have been conceived to support teachers in the process of documenting their teaching practices, making their learning design ideas explicit and sharable [1–4]. The LD process often implies taking decisions about the selection of the most appropriate pedagogical model, the definition of the flow of tasks, the specification of roles as well as the choice of the most suitable resources and educational tools that can support the tasks defined, all to lead to potentially effective learning considering the needs of the educational context. However, despite existing proposed representations of pedagogical practice are varied, some are too specific for particular pedagogies and general approaches are not sufficiently accessible for teachers that do not have the required technical skills [5]. More intuitive visual representations of LD are needed [1, 2]. Moreover, with the spread of ICTs more complex educational scenarios are arising –combining face-to-face and online teaching in different places (in-class and out-of-class) and times (synchronous and asynchronous) [6]. [7] distinguishes two types of LD tools: “tools for visualizing designs” (which can be used to visualize



and represent LDs) and the “pedagogical planners” (which can guide and support practitioners in making informed learning design decisions). In this paper, we present a LD tool that aims fitting in both categories bringing together the advantages of both types of tools. ILDE2/edCrumble can be considered a pedagogical planner which provides an innovative visual representation of the LDs characterized by data analytics with the aim of facilitating the planning, visualization, understanding and reuse of complex LDs. Specifically, the decision-making during the LD process is supported by two types of analytics [8]: resulting from the design of the activities sequenced in a timeline (LD analytics); and aggregated meta-data extracted from several grouped LDs created by multiple teachers within a community, e.g. a school (community analytics).

## 2 Technological Background

edCrumble is a web-based running LD editor prototype developed in JavaScript and HTML5. It is mainly composed of five zones (see Fig. 1), described as follows.



**Fig. 1.** edCrumble screenshot with the zones indicated in red (<https://ilde2.upf.edu/edcrumble/>) (Color figure online)

**Zone1:** It allows users to provide general information about the LD. The title, number of students and the start and end dates of the LD. It has three buttons to specify: (a) the LD description, the educational level and topic; (2) the list of learning objectives; and (3) the evaluation. **Zone2:** It allows users to create in-class and out-of-class activities and place them in a timeline limited by the dates introduced in zone1. The timeline has two main layers by default (in and out-of-class), where the activities are visualized sequentially depending on their schedule and type. **Zone3:** It allows users to edit the activities. Once an activity is selected, the user can set up the corresponding learning objectives and add the tasks that compose it. Indicating and editing for each task: the time allocated, the cognitive process level associated (according to the Blooms’ taxonomy [9]), the students type of work (individual, in groups or the whole class), the teacher’s presence (teacher available face-to-face, online or not present), and the evaluation mode (graded task, not graded or task for auto-evaluation). The user can also write a description of the task to be done by the students with indicators for teachers and add the associated learning resources. **Zone4:** It allows users to select the resources for the activities. Resources are divided on different categories (placed in different tabs):

*Files, Apps, Physical, Communication, Social* and *MOOCs*. The user can drag and drop a resource to the task of an activity and edit its characteristics: title, description, target (teacher or student resource), host-medium type (miscellanea, LMS, local storage, MOOC platform, web, physical artifact, cloud storage) and host-medium name. Moreover, it is possible to specify an URL for the resource and/or upload a file. After adding a resource in an activity, a visualization of an icon associated to this resource appears automatically in the timeline, placed in a new layer depending on the host-medium type (see Fig. 1 where a resource added in the second activity's task in zone3 appears in a host-medium layer -in grey- into the timeline in zone2, aligned with the corresponding activity). **Zone5:** It allows users to consult LD analytics extracted from the meta-data of the produced LD itself. Design analytics are divided on different categories (placed in different tabs): in-class/out-of-class time analytics, tasks 'cognitive process, student type of work, teacher presence, tasks' evaluation mode. In each category it is possible to have 3 different visualizations: global time statistics, statistics depending on the activities 'type (in or out-of-class) and depending on the learning objectives. Last, a button on the **Zone2** allows users to have another view of the timeline hiding the time intervals between the activities and activating the analytics per activity (controlled by a legend composed by buttons corresponding to the different LD analytics' categories). Resulting in a completed interactive visual representation of the LD (see Fig. 2).



**Fig. 2.** Visual representation of a LD composed by 2 in-class and 1 out-of-class activities and 3 resources placed on 3 host-medium layers. Screenshot from the activities' analytics view.

edCrumble has been integrated as an authoring tool within the Integrated LD Environment (ILDE2) [4]. The integration of edCrumble into ILDE2 allows practitioners to co-edit, share, remix and comment their designs and others' designs within a community of teaching -ILDE2 is built on top of the LdShake platform that provides social network features [10]. Moreover, it facilitates teacher's access their designs for future design improvements during the iterative processes of the LD and teacher inquiry cycles (as LdShake acts as a repository of LDs). Once teachers have implemented their LDs, they can upload their evaluations to the edCrumble editor, helping others understand their impact and facilitating the adaptation and reusability of their LDs (for instance, describing the main challenges found or uploading links to the resulting learning analytics). The tool allows generating LD analytics aggregated from all the LDs placed in a folder, named as community analytics -supporting teachers' decision making during the LD process not only at their individual level but also allowing the possibility of considering the colleagues' LDs analytics in their community. The tool also offers the possibility of activating pedagogical guidelines (e.g. flipped classroom) during the design process as well as generating a LD summary including: (1) a printable syllabus

with all the analytics generated; and (2) an interactive visualization to be embedded or shared with the colleagues but also with the students to help them organize their courses.

### 3 Use Case, Preliminary Results and Future Work

In the demo we will show the use case of how LD and community analytics extracted from ILDE2/edCrumble can help balancing the out-of-class workload between different courses which are part of a whole curriculum and support the necessary reflection process for specifically improving the LD quality of the activities within a community of educators. Despite the final evaluations of ILDE2/edCrumble are part of an ongoing cycle of a design-based research process, preliminary results from initial evaluation workshops with stakeholders indicate that the tool is being perceived as easy to use and useful. But also, the need for further work has been identified in the line of providing more flexibility during the activities' creation process (e.g. allowing users to import their activities from existing calendars or creating grouped activities which follow a certain time pattern).

**Acknowledgements.** This work has been partially funded by RecerCaixa (CoT project) and the Spanish Ministry of Economy and Competitiveness under MDM-2015-0502, TIN2014-53199-C3-3-R, TIN2017-85179-C3-3-R.

### References

1. Conole, G., Wills, S.: Representing learning designs - making design explicit and shareable. *Educ. Media Int.* **50**(1), 24–38 (2013)
2. Agostinho, S.: The use of a visual learning design representation to support the design process of teaching in higher education. *Australas. J. Educ. Technol.* **27**(6), 961–978 (2011)
3. Laurillard, D., et al.: A constructionist learning environment for teachers to model learning designs. *J. Comput. Assist. Learn.* **29**(1), 15–30 (2013)
4. Hernández-Leo, D., Asensio-Pérez, J.I., Derntl, M., Pozzi, F., Chacon-Perez, J., Prieto, L.P., Persico, D.: An integrated environment for learning design. *Front. ICT* **5**, 9 (2018)
5. Pozzi, F., Asensio-Pérez, J.I., Persico, D.: The case for multiple representations in the learning design life cycle. In: Gros, B., Kinshuk, M.M. (eds.) *The Future of Ubiquitous Learning*, pp. 171–196. Springer, Heidelberg (2016). [https://doi.org/10.1007/978-3-662-47724-3\\_10](https://doi.org/10.1007/978-3-662-47724-3_10)
6. Norberg, A., Stöckel, M.B., Antti, M.: Time shifting and agile time boxes in course design. *Int. Rev. Res. Open Distrib. Learn.* **18**(6) (2017)
7. Conole, G.: *Designing for Learning in an Open World*, vol. 4. Springer, New York (2012). <https://doi.org/10.1007/978-1-4419-8517-0>
8. Hernández-Leo, D., Martínez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J.A., Rodríguez-Triana, M.J.: Analytics for learning design: a layered framework and tools. *Br. J. Educ. Technol.* (Accepted)
9. Krathwohl, D.R.: A revision of bloom's taxonomy: an overview. *Theory Pract.* **41**(4), 212–218 (2002)
10. Hernández-Leo, D., et al.: LdShake: learning design solutions sharing and co-edition. *Comput. Educ.* **57**(4), 2249–2260 (2011)



# Ensuring Novelty and Transparency in Learning Resource-Recommendation Based on Deep Learning Techniques

Wael Alkhatib<sup>(✉)</sup>, Eid Araache, Christoph Rensing, and Steffen Schnitzer

Fachgebiet Multimedia Kommunikation, Technische Universität Darmstadt,  
S3/20, Rundeturmstr. 10, 64283 Darmstadt, Germany  
{wael.alkhatib, christoph.rensing, steffen.schnitzer}@kom.tu-darmstadt.de

**Abstract.** In this paper, we present an innovative approach for learning resources recommendation. The approach takes into account users' short and long-term interests while ensuring transparency in explaining why a resource is recommended. Our approach relies on Deep Semantic Similarity Model (DSSM) to implicitly measure the semantic similarity between the user interest and the available resources for a recommendation. By taking into consideration the user previous activities, knowledge and current interest, the system reflects the user's history as queries of keywords. The experimental results proved the system usefulness based on a conducted survey.

**Keywords:** Recommender system · Word embeddings · Deep learning

## 1 Introduction

Due to the ever increasing amount of learning resources or learning activities available, it becomes more and more difficult to find suitable items to satisfy a particular learning need. Recommender systems, in general, aim to reduce this burden of information overload by predicting items of interest to a user. Also in the TeL area, recommendation systems has been the subject of research and development for around 10 years. The requirements for recommendations in the TeL area differ from the requirements for product recommendations. Learners differ in their characteristics and have different needs. A recommendation system must take this into account. A key aspect is the individual knowledge and skills that usually change during the learning process. Also the changing interests of a learner must be taken into account. In addition each learner has his own characteristics such as capabilities, learning speed and learning style. Pedagogical theories also suggest that learners should be confronted and challenged with unexpected content from time to time. To stimulate critical thinking and counter biases, recommendations should differ from those that a learner already knows [BS12]. Overall, recommendation systems for TeL should therefore recommend novel diverse and serendipitous learning resources [BS12].

In our work we present a new method for recommending learning resources, taking into account several of the aforementioned requirements. We use Deep Learning to implicitly embed the text semantic in the recommendation process and combine it with natural language processing techniques and external knowledge bases. We propose a query-based recommender system, which reflects the recommendation as results of a query of the user preferences taking into consideration his previous activities, knowledge and current interest.

## 2 Methodology

### 2.1 Ontology Generation

To build a comprehensive ontology we combine different existing lexicial databases, namely *WordNet* [Mil95], *YAGO* and *ConceptNet* [SH12] to build a first ontology. We expand this first ontology by creating additional relations, since the lexical databases are static and have small coverage of concepts for particular domains. We crawl new semantic relations from *Wikipedia* using lexico syntactic pattern-based approach, specifically, we use the six Hearst [Hea92] patterns to detect taxonomic relations.

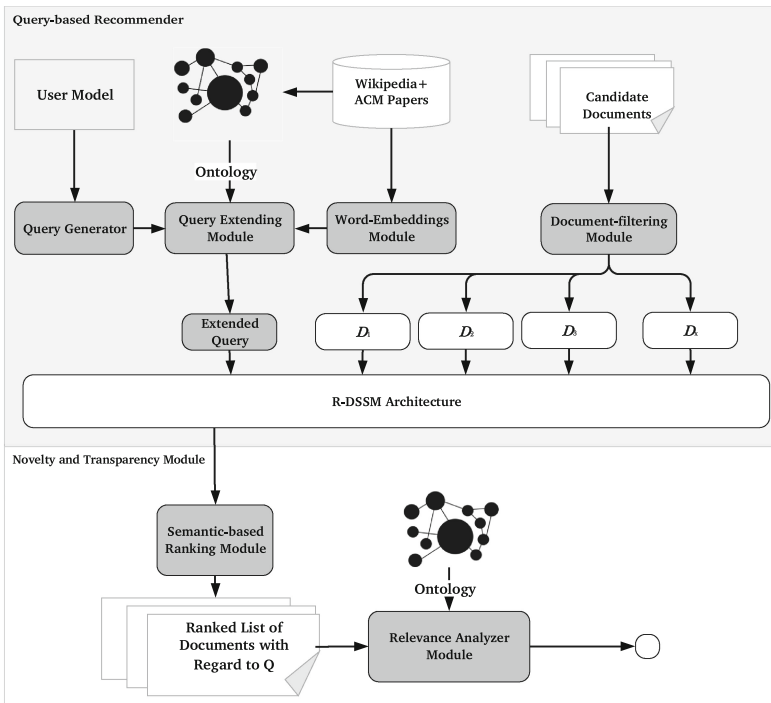


Fig. 1. Block diagram of the proposed learning next step recommender system

## 2.2 User Model Creation

The user model should provide implicit and explicit information about the user interest, progress, and preferences. For building the user model: First, the noun-phrases are extracted from the documents the user mastered so far using a linguistic filter. The interests of a user will be identified by applying Term Frequency/Inverse Document Frequency (TF-IDF) on these phrases. Terms with high TF-IDF weight indicate specialized concepts, contrary terms with low TF-IDF would represent the general interest of the user.

## 2.3 Query-Based Recommender System R-DSSM

The proposed recommender system transfers the recommendation process to the information retrieval space. Firstly, using the *Query Generator*, the user interest and preferences are reflected as a set of concepts by the *User Model*. The *Query Generator* generates a set of queries in such a way that each query partially reflects the user interest in some direction. The queries cannot be randomly generated. Rather, the semantic relations between the concepts are taken into consideration. Secondly, we extend the query using *Query Extending Module* by searching for concepts with semantic relationships to the keywords of the original query. A document that covers a more specific topic or concept can be found by replacing a concept in the query with its hyponym. Documents related to more general concepts can be found by replacing a concept with its hypernym.

Then, the *Document-filtering Module* represents documents as a set of related noun phrases. We use the word embeddings to measure the relatedness between the noun phrases. This reduced representation of the available documents will be passed to the word-hashing module. In the last step, the R-DSSM model is used to obtain a ranked list of the candidate documents for each query. The input for the model is the word-hashing of the query and the representations of the documents. The output of the DSSM is a vector with 120 features that represents the semantics of the input. To measure the similarity, a cosine-similarity layer is attached to the top of the DSSM. Based on this similarity the resources are ranked.

## 2.4 Novelty and Transparency Module

The output of the query-based recommender system is a ranked list of documents. Multiple queries are used and thus we have multiple lists of ranked documents. These will be combined based on a majority vote over the different queries. All the documents recommend by one query only will be filtered out. The remaining list is re-ranked based on matching the main keywords in these documents against the set of terms from our basic queries using word embeddings and a hard cosine similarity threshold of 0.9. Finally, the *Relevance Analyzer* module matches the representative concepts of the recommended documents with the user history and highlights the semantic relatedness between the concepts in the user history and the new documents based on the named relation in our ontology.

### 3 Evaluation

Since we had no suitable learning objects dataset available, a set of 100,000 ACM papers, published in different ACM conferences [SK15], was used. A survey was conducted in order to evaluate whether our approach is able to recommend novel resources and whether the transparency about the recommendation can be created. Two groups of people were involved in the evaluation. The first group consists of experts in machine learning while the second group represents developers with good background in computer science.

Five virtual users were created with different interests. For each user, we define the history by one paper in the corresponding domain. From this paper we extracted a set of concepts that represent his interests. The recommendations that provided by the R-DSSM corresponding to each query are analyzed regarding three aspects of the retrieved documents, mainly what are the more general, more specific and related concepts to the user query. After generating the recommendations for each user and extracting the more-general, more-specific and related concepts. The results of the system were given to all evaluators. They have been asked to rate how much the query and the recommended papers are related to each other. We used a scale from 1 to 5 for rating each recommended paper with regard to the query with 1 as a perfect match and 5 as no relation. We used the Normalized Cumulative Gain (NCG) to measure the system usefulness. The system usefulness from the experts perspective equals 0.61%, and according to the second group equals 0.73%.

### 4 Conclusion

In this paper we have proposed a query-based recommender system. The system relies on DSSM to implicitly measure the semantic relatedness between a query representing the user preferences and the resources available for recommendation. The conducted survey proved the effectiveness of using the proposed method for providing novel recommendations.

### References

- [BS12] Buder, J., Schwind, C.: Learning with personalized recommender systems: a psychological view. *Comput. Hum. Behav.* **28**(1), 207–216 (2012)
- [Hea92] Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th Conference on Computational Linguistics-Volume 2*, pp. 539–545. Association for Computational Linguistics (1992)
- [Mil95] Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
- [SH12] Speer, R., Havasi, C.: Representing general relational knowledge in ConceptNet 5. In: *LREC*, pp. 3679–3686 (2012)
- [SK15] Sugiyama, K., Kan, M.-Y.: A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *Int. J. Digit. Libr.* **16**(2), 91–109 (2015)



# Exploring Math Achievement Through Gamified Virtual Reality

Espen Stranger-Johannessen<sup>(✉)</sup>

Inland Norway University of Applied Sciences, Hamar, Norway  
espen.strangerjohannessen@inn.no

**Abstract.** The immersive nature of virtual reality head-mounted displays (HMDs) offers new ways of leveraging digital games and gamification for learning, particularly by increasing motivation. This study investigates the effects of a quasi-experiment in Norway where students (currently  $N = 79$ , but ongoing) in grade 5 practiced multiplication using HMD for six weeks as part of their regular math classes, compared to a control group (currently  $N = 37$ ). The exercises were buying items from shops, and the gamification consisted of users collecting credits for correct answers and attempts, which they could use to “develop” dragons from eggs to adult dragons. The control group had regular math instruction. The preliminary results suggest an increased score on the post-test for boys who used HMDs ( $N = 40$ ), but the number of boys in the control group ( $N = 16$ ) was too low for the results to be significant at the .05 level.

**Keywords:** Virtual reality · Mathematics education · Gamification  
Head-mounted display · Gender

## 1 Introduction

Low achievement in mathematics is a challenge in many countries, including Norway, and low motivation is particularly noted in boys [1]. While digital games for learning have been used to improve educational achievement for many years, virtual reality (VR) is a novelty in classrooms, with the potential of harnessing the motivational effects of immersive technology and gamification. There are different kinds of VR, and a major distinction is between non-immersive and immersive VR. The former typically involves manipulating a 3D environment on computer screens using a keyboard and mouse. Immersive VR entails complete visual immersion, either in the form of the Cave Automatic Virtual Environments (CAVE), or head-mounted displays (HMDs). In CAVE, a virtual reality is projected on all four walls and the floor of a room while the user wears 3D glasses and can move around the room freely. The need for a designated room and multiple projectors makes CAVE rather expensive and less flexible than HMDs. This study presents the design and preliminary findings of a study where grade five students used HMDs during six weeks to learn multiplication, and provides an account of future steps as well as implications if the preliminary findings are supported when the research is complete.



## 2 Previous Research and Conceptual Framework

While there is an extensive body of research on educational games, which has mostly identified improved learning outcomes [2], research on learning through using HMD is very limited since the technology has only been readily available for a few years. One study [3] found only 21 documents reporting experimental studies with HMDs, six of which measured cognitive or affective learning outcome, none in mathematics. Their call for studies in authentic settings is addressed in this paper.

Several models have been proposed for conceptualizing how digital games affect learning outcomes. Learning outcomes are often divided into the cognitive, affective, and psychomotor domains [4]. The author proposes a model where these domains are considered effects of VR, each with a primary learning outcome (line arrows), but subsidiary outcomes (dashed arrows) are also possible (see Fig. 1). In this model, gamification and the VR affordances of presence and flow primarily have an affective effect, making users more motivated and willing to spend more time on task [5], which is thought to lead to better retention of content. The potential of VR for natural semantics [6] allows for novel cognitive processing, which is thought to result in better comprehension. Finally, psychomotor practice in VR, such as surgical procedures, may yield better psychomotor skills.

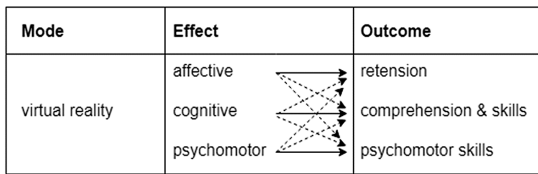


Fig. 1. Conceptual model of the effects of virtual reality.

The VR application in this experiment only attempts to address the affective effect of VR, which is hypothesized to lead to better retention (learning of the multiplication table), and subsidiarily increased comprehension and skills through increased interest in and effort in mathematics.

## 3 Research Design and VR Application

### 3.1 Research Design

This study investigates the effects of a quasi-experiment where students (currently N = 79, but ongoing) in grade 5 (age 10–11) practiced multiplication using HMD for six weeks as part of their regular math classes, compared to a control group (currently N = 37). The math test was administered by the teacher twice: before the students used the HMD and afterwards. The exercises were buying items from shops, and the gamification consisted of users collecting credits (visualized as diamonds) for correct answers and

attempts, which they could use to “develop” dragons from eggs to baby dragons to adult dragons. The control group had regular math instruction.

The HMD were distributed to two schools in different municipalities in Norway by the company VR Education. The invited schools were part of a pilot phase where VR Education tested hardware and software, and invited the author’s institution to carry out research in conjunction with the pilot. Two sister companies of VR Education developed the software used in the HMD. Colleagues of the author developed the math test with 55 questions, including 12 questions on multiplication. The questions were chosen based on the grade 5 curriculum and grade 5 textbooks, and vetted by a grade 5 math teacher. The questions differ in their degrees of difficulty, ranging from simple to more challenging questions in the categories addition, subtraction, number sequence, multiplication, and division.

### 3.2 VR Application

When putting on the HMD, the student enters the main scene, which has two buildings and a hill with a dragon on top of it, each of which represents a new scene the student can enter. Inside the buildings – a grocery store and sports equipment store – the students see a numeric keypad, a price list with pictures of items for sale and their price, ranging from one to ten. Pictures of items found in the price list appear one by one, and when the last item has been displayed the number representing the items appears as a blue digit and the user may enter the answer (the product of the price and the number of items displayed; see Fig. 2). “Typing” the digits is done by looking at a digit on the keypad so that a purple dot in the centre of the screen rests on the digit, then tapping a button on the side of the HMD. A correct answer yields one diamond. Students can leave the store by tapping another button on the HMD, which brings them back to the main scene. From here they may enter a store or the dragon scene.

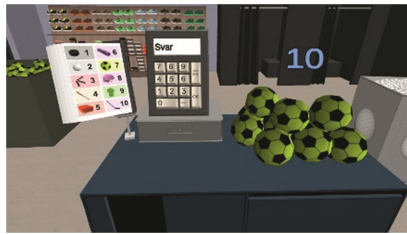


Fig. 2. Math activity in the sports equipment store.

## 4 Preliminary Results

To assess the effect of using HMDs for six weeks the effect size (Cohen’s  $d$ ) broken down by gender was calculated, and an analysis of variance (ANOVA) was carried out to measure the significance. The preliminary results suggest an increased score on the post-test, most notably for boys who used HMD ( $N = 40$ ), but the number of boys in the

control group (N = 16) is too low for the results to be significant at the .05 level (see Table 1). As the study expands in the next months to include more students and interview teachers, the effect of HMDs will be studied in more detail. One teacher noted increased interest in mathematics, which supports the assumption of an affective effect.

**Table 1.** Mean difference pre-test/post-test, effect size, and significance

Gender	Experiment				Control			
	N	Score	Cohen's <i>d</i>	Sig.	N	Score	Cohen's <i>d</i>	Sig.
Boys	40	2.63	0.16	.000	16	0.88	0.17	.393
Girls	38	2.24	0.21	.001	21	1.52	0.22	.070
Total	79	2.35	0.37	.000	37	1.24	0.20	.051

## 5 Conclusion and Future Steps





This study on the effects of VR HMDs on mathematics achievement, set in authentic classroom settings, which ensures a high ecological validity of the findings, is the first study of its kind. These preliminary findings are inconclusive as the sample size, particularly the number of boys in the control group, is too small for the results to be significant. Gender is a key topic in research on digital games and VR, and some points to the importance of boys' relatively higher prior experience and interest, although the research is inconclusive [7]. One teacher in this study reported that girls experienced cybersickness, which merits further exploration. A limitation of the VR application was that it only contained multiplication exercises, but if future research supports the tendency in these data of an overall learning gain, the assumption of an affective effect of VR is further strengthened, as this application does not offer natural semantics or other novel conceptual ways of teaching mathematics.

## References

1. Sax, L.: *Boys Adrift: The Five Factors Driving the Growing Epidemix of Unmotivated Boys and Underachieving Young Men*. Basic Books, New York (2007)
2. Boyle, E.A., Hainey, T., Connolly, T.M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., Pereira, J.: An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Comput. Educ.* **94**, 178–192 (2016)
3. Jensen, L., Konradsen, F.: A review of the use of virtual reality head-mounted displays in education and training. *Educ. Inf. Technol.* **23**, 1–15 (2017)
4. Bloom, B.S.: *Taxonomy of Educational Objectives*. McKay, New York (1956)
5. Alhalabi, W.S.: Virtual reality systems enhance students' achievements in engineering education. *Behav. Inf. Technol.* **35**(11), 919–925 (2016)
6. Mikropoulos, T.A., Natsis, A.: Educational virtual environments: A ten-year review of empirical research (1999–2009). *Comput. Educ.* **56**(3), 769–780 (2011)
7. Papastergiou, M.: Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* **52**(1), 1–12 (2009)



# Observational Scaffolding for Learning Analytics: A Methodological Proposal

Jairo Rodríguez-Medina<sup>1</sup> , María Jesús Rodríguez-Triana<sup>2,3</sup> , Maka Eradze<sup>2</sup> ,  
and Sara García-Sastre<sup>4</sup> 

<sup>1</sup> Center for Transdisciplinary Research in Education, University of Valladolid, Valladolid, Spain  
jairo.rodriguez.medina@uva.es

<sup>2</sup> School of Digital Technologies, Tallinn University, Tallinn, Estonia

<sup>3</sup> School of Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

<sup>4</sup> GSIC-EMIC Research Group, University of Valladolid, Valladolid, Spain

**Abstract.** Temporal analysis of learning data is attracting the interest of researchers, and a growing body of Learning Analytics (LA) research applies lag sequential analysis. However, lack of methodological frameworks that guide the data gathering and analysis poses multiple conceptual, methodological, analytical and technical challenges. While observation as a technique has been already used in LA, systematic observation methods and designs have not been applied so far, and parameters often used in the observational domain (such as order and duration) are still under-researched. In this paper we propose a methodological framework, and illustrate its potential by applying it in the analysis of a Knowledge Forum dataset. Results show the potential of the proposed method to uncover behavioral patterns prospectively (lag +1 to lag +5) or retrospectively (lag -1 to lag -5), and to reduce this information through *polar coordinate analysis*. Moreover, as illustrated in this paper, observational methods offer a rigorous framework for LA datasets, enabling the replicability, validity and reliability of the results.

**Keywords:** Learning Analytics · Observational methodology  
Temporal analytics · Lag sequential analysis · Polar coordinate analysis

## 1 Introduction

Observational methods have been used in education for decades. While humans have traditionally mediated observations, current Learning Analytics (LA) solutions provide automatic means to assist the data collection and analysis process. We could consider LA as “modern” data gathering and analysis technique that support the observational process, either by reducing the workload (thanks to the automation of the process) or enriching the datasets with data coming from digital spaces. Thus, the combination of both human and computer-mediated observations could offer a complementary view of educational contexts [1, 2], which is needed when teaching and learning process happen across spaces [3]. Even though LA offer new insights in the educational domain, current

solutions often lack of systematic methodological frameworks, compromising the replicability, validity and reliability of the results [4]. In this paper, we hypothesize that systematic observational methods [5, 6] could contribute to alleviating these challenges.

As Ochoa et al. [8] mention, three methodological challenges threaten the development of LA: (1) the difficulty of rigorously assessing the research results; (2) the studies are rarely comparable; and (3) sub-optimal methodologies and tools are often applied when better alternatives exist. These issues only recently have come to the foreground of LA research challenges, and evaluation frameworks have been proposed to assess the users' subjective impression of using an LA system [9] and to provide a more diagnostic view of the performance of such systems [10]. Moreover, a few cases involve several data sources for triangulation, and even more rarely, data from physical spaces [3, 11]. Eradze et al. [2] analyze the difficulties of integrating observational records into multi-modal datasets, highlighting the need for a systematic procedure that defines the nature of the data and the unit of analysis, so that the observational design and the parameters to be registered are adjusted accordingly. This need could be addressed through indirect observation, a concept recently coined in the area of observational methods and considers *studying both verbal behavior and textual material, whether in the form of transcripts or original material produced by the participants in a study* [12]. The approach already applied to both conventional and new forms of communication (WhatsApp, Twitter, blog posts messages) [13] involves the analysis of data generated in physical or digital settings.

## 2 Proof of Concept and Discussion

In this “proof of concept”, to assess the added value of applying a systematic observational approach, we have chosen a publicly available dataset previously analyzed by other authors [7]. The dataset contains information about 1101 notes in 50 threads supported by Knowledge Forum (<http://www.knowledgetforum.com>). Since we work with a predefined dataset, we will tackle the same research question posed by Chen et al. [7] in their data analysis i.e.: “What are the underlying behavioral patterns that could distinguish productive knowledge-building dialogues – dialogues with apparent attempts to advance collective knowledge?”. In order to answer the question, we will apply lag sequential analysis to better understand the temporal relations and patterns, while in a slightly different manner. Our proposal for data transformation is novel as it is supported by observational methods devoted to analyze participants' behavior in an authentic setting using ad hoc observation instruments [5]. This scientific procedure is suited for the analysis of social interaction and its temporal evolution [5, 15].

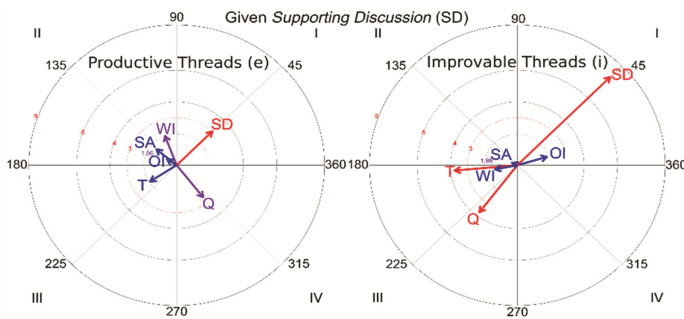
## 3 Data Analysis

*Lag sequential analysis* was used to investigate sequential relationships between discrete behaviors (events) and interactive states. Additionally, we apply *polar coordinate analysis* [16, 14], this technique allows for data reduction by using the Zsum statistic ( $Z_{sum} = \Sigma z / \sqrt{n}$ ), where Z represents the independent values obtained from the adjusted

residuals found for the respective lags of  $-5$  to  $-1$  and  $1$  to  $5$ , with  $n$  as the number of lags. To carry out this analysis we used SDIS-GSEQ software package v. 5.1 [6] and HOISAN v. 1.6 [17].

## 4 Results

To illustrate the potential of polar coordinate analysis for the data reduction, we show an example focused on the *Supporting Discussion* (SD) behavioural category. As Fig. 1 depicts, in non-productive (or improvable) there was a significant mutual inhibition between posts coded as Supporting Discussion (SD) and Questioning (Q) (see Quadrant III, radius = 3.74,  $p < .01$ ), while in the productive threads Supporting Discussion (SD) significantly activated Questioning (Q) (see Quadrant IV, radius = 2.57,  $p < .05$ ). Similarly, Supporting Discussion (SD) posts had a significant inhibition on Theorizing (T) in the non-productive threads while no impact was detected on productive threads. Finally, Working with Information (WI) posts significantly activated the focal behaviour (SD) in the productive threads while in the non-productive threads a non-significant inhibition was detected on the focal behaviour (SD).



**Fig. 1.** Polar coordinate analysis results for Supporting Discussion (SD) as the focal behavior. Significant relationships between focal and conditional behaviors marked in red ( $p < .01$ ) and purple ( $p < .05$ ) colors (Color figure online)

## 5 Discussion

This paper proposes the application of systematic observational methods [5, 6] as a way to alleviate the methodological and analytical challenges of the Learning Analytic community identified before [8]. The example also provides a proof of concept for the informative potential that polar coordinate analysis may have for data reduction in the field of LA. The application of a rigorous observational design allowed us to uncover behavioral patterns prospectively (lag  $+1$  to lag  $+5$ ) or retrospectively (lag  $-1$  to lag  $5$ ), and to reduce this information through polar coordinate analysis [16]. Thus, this technique may have a remarkable potential in order to interpret the analysis of big datasets, which is common in LA. Moreover, aligned with the open-source movement and the

existence of public datasets in LA, the application of open-source software widely adopted by the community of observational methods (e.g., SDIS-GSEQ, HOISAN or THEME) could contribute to address the assessment and comparison challenge reported by Ochoa et al. [8].

## References

1. Macfadyen, L.P., Dawson, S.: Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *J. Educ. Technol. Soc.* **15**, 149–163 (2012)
2. Eradze, M., Rodríguez-Triana, M.J., Laanpere, M.: Semantically annotated lesson observation data in learning analytics datasets: a reference model. *Interact. Des. Archit. J.* **33**, 75–91 (2017)
3. Rodríguez-Triana, M.J., Vozniuk, A., Holzer, A., Gillet, D., Prieto, L.P., Boroujeni, M.S., Schwendimann, B.A.: Monitoring, awareness and reflection in blended technology enhanced learning: a systematic review. *Int. J. Technol. Enhanc. Learn.* **9**, 126–150 (2017)
4. Blikstein, P., Worsley, M.: Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *J. Learn. Anal.* **3**, 220–238 (2016)
5. Anguera, M.T.: Observational methods (General). In: Fernández-Ballesteros, R. (ed.) *Encyclopedia of Psychological Assessment*, vol. 2, pp. 632–637. Sage, London (2003)
6. Bakeman, R., Quera, V.: *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press, Cambridge (2011)
7. Chen, B., Resendes, M., Chai, C.S., Hong, H.Y.: Two tales of time: uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interact. Learn. Environ.* **25**, 162–175 (2017)
8. Ochoa, X., Hershkovitz, A., Wise, A., Knight, S.: Towards a convergent development of learning analytics. *J. Learn. Anal.* **4**, 1–6 (2017)
9. Scheffel, M., Drachler, H., Toisoul, C., Ternier, S., Specht, M.: The proof of the pudding: examining validity and reliability of the evaluation framework for learning analytics. In: Lavoué, É., Drachler, H., Verbert, K., Broisn, J., Pérez-Sanagustín, M. (eds.) *EC-TEL 2017. LNCS*, vol. 10474, pp. 194–208. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_15](https://doi.org/10.1007/978-3-319-66610-5_15)
10. Rodríguez-Triana, M.J., Prieto, L.P., Martínez-Monés, A., Asensio-Pérez, J.I., Dimitriadis, Y.: The teacher in the loop: customizing multimodal learning analytics for blended learning. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 417–426. ACM (2018)
11. Schwendimann, B.A., Rodríguez-Triana, M.J., Vozniuk, A., Prieto, L.P., Boroujeni, M.S., Holzer, A., Gillet, D., Dillenbourg, P.: Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Trans. Learn. Technol.* **10**, 30–41 (2017)
12. Anguera, M.T., Portell, M., Chacón-Moscoso, S.: Indirect observation in everyday contexts: concepts and methodological guidelines within a mixed methods framework. *Front. Psychol.* **9**, 1–20 (2018). <https://doi.org/10.18608/jla.2017.43.1>
13. Radzikowski, J., Stefanidis, A., Jacobsen, K.H., Croitoru, A., Crooks, A., Delamater, P.L.: The measles vaccination narrative in Twitter: a quantitative analysis. *JMIR Pub. Heal. Surveill.* **2** (2016)
14. Sackett, G.P.: Lag sequential analysis as a data reduction technique in social interaction research. *Except. Infant. Psychosoc.* **4**, 300–340 (1980)
15. Bakeman, R., Gottman, J.M.: *Observing Interaction*. Cambridge University Press, Cambridge (1997)

16. Anguera, M.T.: From prospective patterns in behavior to joint analysis with a retrospective perspective. In: Colloque sur invitation "Méthodologie d'analyse des interactions sociales" Université de la Sorbonne, Paris (1997)
17. Mendo, A.H., López, J.A.L., Paulis, J.C., Sánchez, V.M., Brincones, J.L.P.: Hoisan 1.2: Programa informático para uso en metodología observacional. Cuad. Psicol. del Deport **12**, 55–78 (2012)





# Cohesion-Centered Analysis of Sociograms for Online Communities and Courses Using *ReaderBench*

Mihai Dascalu<sup>1,2(✉)</sup>, Maria-Dorinela Sirbu<sup>1</sup>, Gabriel Gutu-Robu<sup>1</sup>,  
Stefan Ruseti<sup>1</sup>, Scott A. Crossley<sup>3</sup>, and Stefan Trausan-Matu<sup>1,2</sup>

<sup>1</sup> University Politehnica of Bucharest,  
Splaiul Independenței 313, 60042 Bucharest, Romania  
{mihai.dascalu, gabriel.gutu, stefan.ruseti,  
stefan.trausan}@cs.pub.ro, maria.sirbu@cti.pub.ro

<sup>2</sup> Academy of Romanian Scientists,  
Splaiul Independenței 54, 050094 Bucharest, Romania

<sup>3</sup> Department of Applied Linguistics/ESL,  
Georgia State University, Atlanta 30303, USA  
scrossley@gsu.edu

**Abstract.** Computer Supported Collaborative Learning (CSCL) environments facilitated by technology have become a viable learning alternative from which valuable data can be extracted and used for advanced analyses centered on evaluating participants' involvement and their interactions. Such automated assessments are implemented within the *ReaderBench* framework, a Natural Language Processing platform that contains multiple advanced text analysis functionalities. The *ReaderBench* framework is based on Cohesion Network Analysis from which different sociograms, relying on semantic similarity, are generated in order to reflect interactions between participants. In this paper, we briefly describe the enforced mechanisms used to compare two Math communities, namely an online knowledge building community and an online course.

**Keywords:** Cohesion Network Analysis · Sociograms · Text cohesion  
Natural Language Processing · *ReaderBench* framework

## 1 Introduction

Teachers and tutors have a limited amount of time to manually assess and grade student output. Moreover, monitoring and scoring student activities using indicators reflective of their performance in terms of participation or collaboration with peers is a cumbersome process. Hence, there is necessity for automated analyses, which led to the development of the Cohesion Network Analysis (CNA) approach and its integration within the *ReaderBench* framework available online at <http://readerbench.com>. *ReaderBench* [1, 2] is a fully functional open-source framework centered on discourse analysis that consists of various Natural Language Processing (NLP) techniques designed to support students and teachers in their educational activities. This paper

presents a brief overview of Computer Supported Collaborative Learning (CSCL) experiments centered on online communities and performed within *ReaderBench*.

## 2 Performed Experiments

Two experiments in different CSCL environments were conducted. These experiments focused on Online Knowledge Building Communities (OKBC) and online courses. CNA transcends Social Network Analysis (SNA) by taking into account discourse quality reflected in semantic cohesion. CNA models interactions between participants and provides a scoring mechanism within collaborative conversations by combining NLP techniques with SNA. In *ReaderBench*, CNA is used to compute cohesion indices that are based on the discourse structure and which reflect participation and collaboration throughout the conversation [2]. Moreover, CNA is tightly coupled with dialogism and polyphony which define the theoretical framing of CSCL [2]. Moreover, CNA closely resembles SNA by relying on equivalent indices to quantify participation within the generated sociograms [1, 2]. Afterwards, hierarchical clustering is used to extract the community's socio-cognitive structure based on two CNA indices derived from the sociogram: in-degree (reflective of collaboration in terms of inbound messages) and out-degree (highlighting active participation in the community).

Two types of views are used to model the interaction between participants, namely a *Force-Clustered Graph* and a *Hierarchical Edge Bundling* visualization [3]. The views are generated using the d3.js library (<https://d3js.org>). The *Force-Clustered Graph* view shows the interactions between participants based on a graph in which the nodes represent participants who are clustered by considering the inter-exchanged messages between them. The size of nodes represents the average score of in-degree and out-degree values from the overall sociogram and it is directly proportional with the participant's score. The clustered participants were colored as follows in descending order of average in-degree CNA scores: central members are colored with blue, active members with green, and peripheral members with orange. The *Hierarchical Edge Bundling* view presents the interactions between participants in a branching manner. The participants are organized into their corresponding cluster: active, central or peripheral. The same colors like in the previous view were used.

The first experiment evaluated the involvement of participants in online blog communities, their interactions and evolution throughout the discussion threads [3]. The analysis was performed on a corpus of 85 conversations from 78 members, cumulating 250 contributions extracted from the online Math community <http://mathequalslove.blogspot.com>. Figure 1 has been blurred in order to anonymize the names of the participants and it depicts the sociogram of the OKBC in which the blog owner is the main person within the community.

The second experiment was used to predict students' completion rates in the context of an online Math course [4]. Based on the generated cohesion graph and a longitudinal analysis, CNA indices were computed for each of the 157 students (from 250 students, only 157 made posts on the forum). The method showed that students who are active on forums are more likely to finish the course, while the number of days spent on the forum and the consistency of posts are predictors for math success.

As the number of participants is considerably larger and the force directed view becomes too cluttered, we opted to include a force-clustered view besides the hierarchical edge bundling view (see Fig. 2). In this case, we can observe a higher presence of central and active members in contrast to the online community in which the discourse is centered around the blog owner who sustains the community (see Fig. 1), while the interactions from other participants are quite limited.

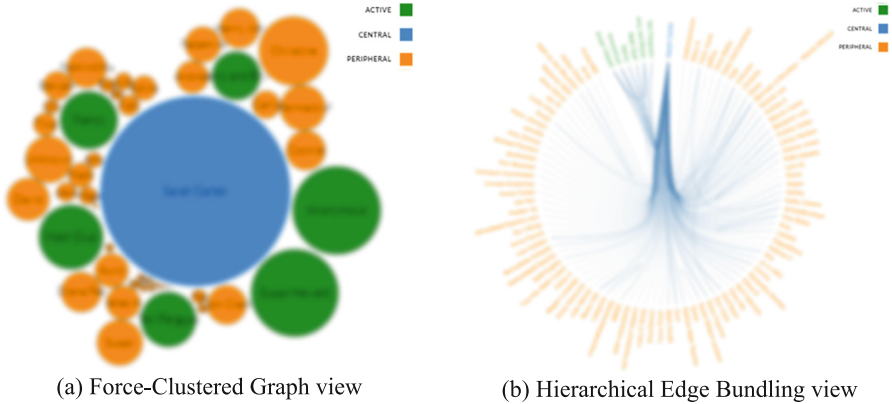


Fig. 1. Sociograms corresponding to the OKBC.

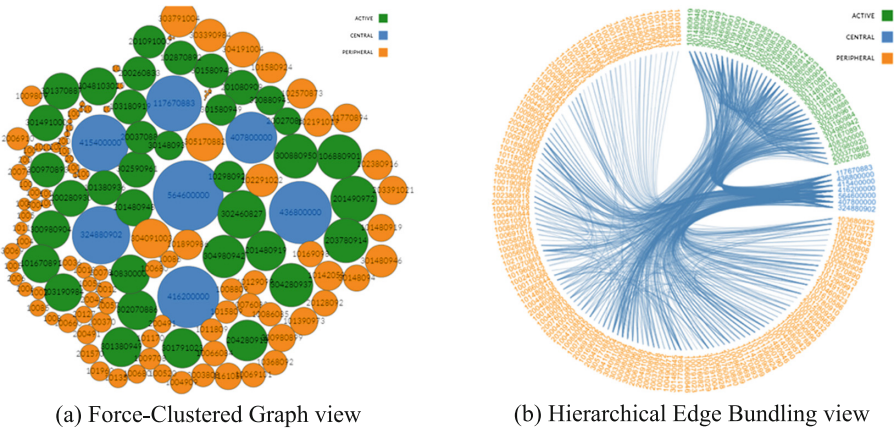


Fig. 2. Sociograms corresponding to the online course.

In addition to the global views, weekly timeframes are also generated in the performed longitudinal analysis in order to represent interactions and connections between participants in an interactive and intuitive manner. While participation in the online community exhibits little fluctuations between adjacent weeks, we can observe specific traits within the online course. Figure 3 shows participants' activity in the first, 9th

(mid-semester), and the last week of the online course. Some interaction patterns can be observed using the newly introduced Force-Clustered Graph views, as follows: (a) the peripheral members play a more important role in the community; (b) a decrease in involvement can be observed towards the end of the course; (c) the discussions are not dominated by a single member. These findings are aligned with the observations from our previous study [5].

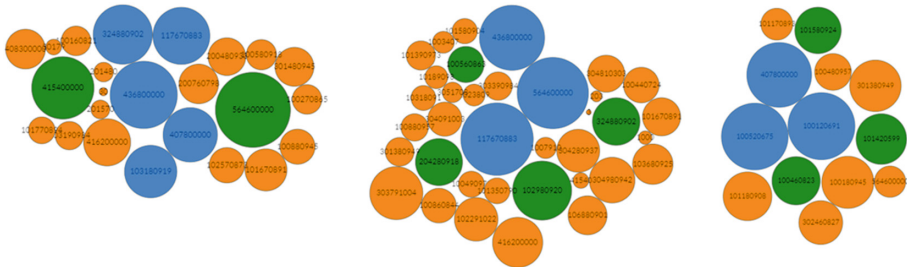


Fig. 3. Sociograms within Math Course - Snapshot views of weeks 1, 9 and 18.

### 3 Discussion and Conclusion

These experiments demonstrate the capability of the *ReaderBench* framework to analyze different online CSCL environments and to perform in-depth, cohesion-centered analyses of interaction patterns. The generated sociograms provide valuable insights with regards to different interactions patterns and can be used in follow-up experiments to provide personalized feedback to learners in order to actively engage them.

**Acknowledgments.** This research was partially supported by the 644187 EC H2020 RAGE project, the FP7 2008-212578 LTfLL project and the National Science Foundation (DRL-1418378). We would like to thank Nicolae Nistor, Tiffany Barnes, Collin Lynch and Catalina Durbala for their help in conducting the previous experiments.

### References

1. Dascalu, M., Trausan-Matu, S., McNamara, D.S., Dessus, P.: ReaderBench – automated evaluation of collaboration based on cohesion and dialogism. *Int. J. Comput. Support. Collab. Learn.* **10**(4), 395–423 (2015)
2. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Allen, L.K.: Cohesion network analysis of CSCL participation. *Behav. Res. Methods* **50**(2), 604–619 (2018)
3. Sirbu, M.-D., Panaite, M., Secui, A., Dascalu, M., Nistor, N., Trausan-Matu, S.: ReaderBench: building comprehensive sociograms of online communities. In: 9th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2017), Timisoara, Romania. IEEE (2017)

4. Crossley, S.A., Dascalu, M., Baker, M., McNamara, D.S., Trausan-Matu, S.: Predicting success in massive open online courses (MOOC) using cohesion network analysis. In: 12th International Conference on Computer-Supported Collaborative Learning (CSCL 2017), Philadelphia, PA, pp. 103–110. ISLS (2017)
5. Sirbu, M.-D., et al.: Exploring online course sociograms using cohesion network analysis. In: Penstein Rosé, C., et al. (eds.) AIED 2018, Part II. LNCS (LNAI), vol. 10948, pp. 337–342. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_63](https://doi.org/10.1007/978-3-319-93846-2_63)



# A Digital Ecosystem for Digital Competences: The CRISS Project Demo

Manolis Mavrikis<sup>1</sup>(✉), Lourdes Guardia<sup>2</sup>, Mutlu Cukurova<sup>1</sup>,  
and Marcelo Maina<sup>2</sup>  
on behalf of the CRISS Consortium

<sup>1</sup> UCL Knowledge Lab, Institute of Education,  
University College London, London, UK  
m.mavrikis@ucl.ac.uk

<sup>2</sup> Universitat Oberta de Catalunya/Open University of Catalonia, Barcelona, Spain  
{lguardia,mmaina}@uoc.edu

**Abstract.** CRISS is a flexible and scalable cloud-based digital learning ecosystem that has the potential to allow the guided acquisition, evaluation and certification of digital competences. This demonstration will highlight some of the key activities under development, their underlying pedagogy and how the platform's features support the acquisition, assessment and certification of digital competences.

**Keywords:** Digital competencies · e-portfolio · Certification

## 1 Introduction

There is an increasing demand for digital skills to be able to function effectively in modern societies. Moreover, with an estimated 90% of jobs requiring digital skills in the near future, it is essential that education and training systems provide individuals with the required competences [1]. However, the definition, identification, support and evaluation of digital skills have been proven to be a real challenge for existing educational systems. Unfortunately, in many schools and classrooms, 21st century ICT tools and skills are still used and taught as add-ons to “business-as-usual” type of teaching. We fail to initiate teaching and assessment of those important skills that are at the same time transforming the way we work, learn and interact [2]. This demonstration presents the underlying theoretical framework, pedagogy and the cloud-based CRISS platform designed for the acquisition and certification of digital competencies.

## 2 Pedagogical Background

### 2.1 Digital Competences Framework

Digital competencies are comparable to the literacy and numeracy of the past. They are equally needed and equally essential for people to function effectively in

modern societies. As social interaction and work is ever more dependent on technology, being digitally competent is a requirement and a right for every citizen [3]. Moreover, digital competence is one of the 8 key competences for lifelong learning identified by the European Union. According to Digital Skills Working Group (2010), digital competence is a set of knowledge, skills and attitudes (including abilities, strategies, values & awareness) that are required when using ICT and digital media to perform tasks [4].

The majority of digital competence frameworks are based on skills development and on the ability to use a specific set of tools and/or applications. As the above definition highlights, skills are only part of the learning domains that are included in Digital Competence; and the ability to use specific tools or applications is just one of the several competence areas that need to be developed by users in order to function in a digital environment [5,6].

The CRISS project, uses DigiComp 2.0 [5] and other frameworks across Europe to define 12 competences grouped in 5 digital skill areas, (1) Digital Citizenship, (2) Communication and collaboration, (3) Search and Manage Information, (4) Digital content creation, and (5) Digital Problem solving.

## 2.2 Integration Pedagogy

The CRISS evaluation system builds on a pedagogy that allows assessing students' digital competences embedded within disciplinary or interdisciplinary problem situations rather than testing them as individual skills out of context (see [7] for more details). As such, the project team is developing mostly cross-disciplinary scenarios that expect project- or problem-based learning but at times are even more open ended as in inquiry-based learning.

## 3 Technical Background

### 3.1 CRISS Platform

CRISS is an adaptive and flexible cloud-based ecosystem to offer new learning experiences around digital competences. The CRISS Core platform includes several modules and submodules as described below (see Fig. 1).

### 3.2 The ICT Manager Tool

This is the largest Module of the CRISS Core Platform and includes several Tools, Modules and Submodules:

- The Administration Module module is designed for user creation, management of roles and permissions, credentials etc.
- ICT Planning Tool. This is the module where teachers can create, in an individual or collaborative way, their lesson plan (calendar, dates of assessments, in order to provide the students with the necessary Scenarios, activities and tasks for Digital Competences).



**Fig. 1.** The navigation menu of the CRISS platform with key modules and submodules

- Scenarios' Creation Tool. This tool allows CRISS partners to create the Scenarios, Activities and Tasks for the acquisition of Digital Competences. It is connected to the Certification Module in order to allow teacher to align their Scenarios with the criteria established in the CRISS Methodological Framework for the Acquisition and Certification of Digital Competences.
- Evaluation and Assessment Tool. This module provides all the necessary features for the assessment and evaluation of Scenarios, tasks and activities, and, for the assessment and certification of Digital Competences.

### 3.3 ePortfolio

The ePortfolio is the 'working' core of the CRISS platform, where students and teachers can perform all their actions, follow the work, get access to all the other modules and see the results of evaluations and the progress in the acquisition of DC and certification. The core of the CRISS platform is a web-based ePortfolio environment where students and teachers can perform all their actions, follow the work, get access to all the other modules and see the results of evaluations and the progress in the acquisition of digital competences and their certification. See Fig. 1 for an example of content shared through the ePortfolio.

### 3.4 Learning Analytics and the ICT Dynamic Profile

Learning analytics contribute to the assessment of learning processes including students' skill development and evaluation. Currently, digital skills is one of those educational areas to which learning analytics has yet to contribute significantly. In CRISS this takes several forms. At the time of this writing, the ICT Dynamic profile offers the teachers and students the visualisation of the data related to the results of evaluation and assessments, the level of acquisition of digital competences and the certification of digital competences, including badges.



### 3.5 CRISS Certification Modules

A set of modules store and provide the criteria, rules and indicators for the acquisition, assessment, evaluation and certification of Digital Competences. This provides the ICT Manager Tool with the criteria and feedback to create Scenarios, activities, tasks, assessments and evaluations according to the CRISS framework. It also enables the teachers to evaluate the student's performance throughout the learning procedure based on the aforementioned criteria and indicators. Lastly, it recollects the information from the evaluation and assessments of the tasks and activities and compares it to the certification framework.

## 4 Use Case

CRISS is going to be available to students across Europe in the next academic year. The platform is introducing new approaches for the support and evaluation of student digital competencies. made possible through innovate assessment and certification techniques and an adaptive learning solution combined with robust pedagogical methodologies. The demonstration will make available scenarios across the five areas of digital competencies but gives particular attention on Digital problem solving as a case of computational thinking.

**Acknowledgements.** The CRISS core platform, including Eportfolio and ICT Dynamic profile, ICT Manager tool, Scenarios and activities creation tool, ICT Evaluation tool, Certification Module? Submodule A, Create Evidence tool and Portability authoring tool, have been conceived and developed by MyDocumenta. Other technical partners in the project include Education4Sight, Diginext and it is coordinated by Exus (see <https://www.crissh2020.eu/partners/> for more details). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No: 732489.

## References

1. EU: ICT for work: digital skills in the workplace: final report, September 2017
2. Pettersson, F.: On the issues of digital competence in educational contexts - a review of literature. *Educ. Inf. Technol.* **23**(3), 1005–1021 (2018)
3. OECD: Understanding the digital divide. OECD: Paris (2001). [www.oecd.org/sti/1888451.pdf](http://www.oecd.org/sti/1888451.pdf)
4. Ferrari, A.: Digital Competence in Practice: An Analysis of Frameworks (2012)
5. Punie, Y., Breko, B., Ferrari, A.: DIGCOMP: a framework for developing and understanding digital competence in Europe. No. 38, 3–17 (April 2014)
6. Guitert, M., Romeu, T., Baztn, P.: Conceptual framework on digital Competences in Primary and Secondary Schools in Europe. In: Proceedings of the 10th Annual International Conference of Education, Research and Innovation ICERI 2017. IATED, Seville, Spain, pp. 5081–5090, November 2017
7. Guardia, L., Maina, M., Julia, A.: Digital competence assessment system: supporting teachers with the CRISS platform. In: Proceedings of the 28th Central European Conference on Information and Intelligent Systems (CECIIS), Varazdin, Croatia, September 2017



# Towards Generation of Ambiguous Situations in Virtual Environments for Training

Azzeddine Benabbou<sup>(✉)</sup>, Domitile Lourdeaux<sup>(✉)</sup>,  
and Dominique Lenne<sup>(✉)</sup>

Sorbonne universités, Université de technologie de Compiègne,  
CNRS UMR 7253 Heudiasyc, 57 avenue de Landshut,  
60203 Compiègne Cedex, France

{azzeddine.benabbou, domitile.lourdeaux,  
dominique.lenne}@hds.utc.fr

**Abstract.** Ambiguous situations are referred to as situations that are open to more than one interpretation. Our objective is to train individuals to handle this kind of situations using Virtual Environments for Training (VET). However, producing a large panel of ambiguous situations adapted to the learner requires serious authoring efforts. To address this issue, we propose to generate these situations automatically without having to write them beforehand.

**Keywords:** Ambiguity · Scenario generation · Virtual environments  
Training

## 1 Introduction

Critical situations can be defined as complex and dynamic situations, often unexpected and difficult to anticipate. These situations are characterized by several dimensions such as ambiguity [1]. This latter refers to situations where the state of the world is subject to different interpretations. In complex domains, especially when the safety of the self and the others is concerned (e.g. healthcare, driving), individuals need to be trained to identify, handle and anticipate such situations. Failing to handle these situations can lead sometimes to disastrous consequences. This is why training in genuine conditions is not always possible. Therefore, using VET can help in that matter. Our pedagogical objective here, is to complete the initial training of the learners. We suppose that they have already acquired the needed technical skills, and we aim to train them to make use of their non-technical skills (e.g. communication, situation awareness). The purpose, in particular, is to train them to reduce the ambiguity in such situations in order to be able to make the most appropriate decisions. To provide such training, we need to confront the learners to various ambiguous situations. However, the complex nature of the domains stands against the possibility of writing all the possible situations beforehand. Especially if we want to have a control on the simulation and present “explicable” situations that enable us to debrief with the learners afterwards. One way to address this issue is to generate these situations automatically.

This paper is organized as follows: In Sect. 2, we present a quick review of the literature on ambiguity. Then, in Sect. 3, we list some related work. Finally, in Sect. 4, we detail the ambiguity generation process.

## 2 Ambiguity

The Merriam-Webster dictionary defines ambiguity as “*a word or expression that can be understood in two or more possible ways*”. This might be the most common definition that comes to people’s mind when ambiguity is referred to. In his famous paper in 1961, Ellsberg [2] defined ambiguity as “*a quality depending on the amount, type, reliability and ‘unanimity’ of information, and giving rise to one’s degree of confidence in an estimate of relative likelihoods*”. This description considers the information as the core item of ambiguity. It is also the case of several conceptions of ambiguity in statistics, economics and risk assessment. Camerer and Weber [3] for example, defines ambiguity as “*uncertainty about probability, created by missing information that is relevant and could be known*”. Ambiguity is also referred to as epistemic uncertainty [4]. This latter comes from the lack of information. Thus, unlike the aleatory uncertainty, it is reducible. Blockley [5] defines ambiguity as a mix of Fuzziness and Incompleteness that corresponds to the epistemic dimensions of his space. While these previous conceptions of ambiguity focus on information, Gaver et al. [6] distinguish three broad classes of ambiguity: (1) *Ambiguity of information* that finds its source in the artefact itself, (2) *Ambiguity of context* that finds its source in the sociocultural discourses that are used to interpret it and (3) *Ambiguity of relationship* that find its source in the interpretative and evaluative stance of the individual. In this paper, we focus on the generation of the first class of ambiguity. It emerges mainly from the remediable lack of relevant information and/or the poor quality of the available information. In healthcare for instance, an example for such ambiguity would be a doctor facing a situation where the medical record of the patient says that s/he is not allergic to a given substance, but manifested anyway an allergic reaction when this substance was administrated.

## 3 Related Work

Cottone et al. [7] conducted a study using a virtual city to investigate how people cope with ambiguity using different means of communication (face-to-face, chat and phone). The participants’ goal was to meet at a specific place of their choice. The virtual city was purposely designed to contain similar places to create ambiguity. Mantovani et al. [8] investigated the suitability of virtual environments for safety training, in particular in capturing ambiguity. The participant’s goal was to find a way out from a virtual library using two types of emergency signs. A particular group faced an ambiguous situation where there was a red ribbon blocking the exit. The participants did not know if they should respect it in this emergency context and find another way out, or they should pass over it. In military field, Raybourn et al. [9] created a multiplayer game to train Special Forces Team Leaders to cope with “uncertain” scenarios such as

ambiguous situations. In all these systems, ambiguous situations are written beforehand. As far as complex domains are concerned, especially when the learner has to be confronted to a large panel of situations, this approach is doomed to fail. Our ambition is to design an original system that generates automatically ambiguous situations. To our best knowledge, there is no system in the literature that adopts such approach.

## 4 Ambiguity Generation

Gaver et al. presented several tactics to create ambiguity of information. In this section, we detail four ways, inspired by these tactics, and illustrate them with examples.

**Using incomplete representation to emphasize uncertainty** by hiding relevant information that is crucial for determining which action to take. Let  $A = \{a_1, \dots, a_n\}$  be the possible actions and  $\{P_1, \dots, P_n\}$  be respectively the sets of their preconditions such as their intersection is not empty. We define the function  $f(A)$  which input is a set of actions  $A$ . The function output is the symmetric difference of the preconditions of the actions. This corresponds to the relevant information that needs to be hidden.

$$f(A) = (P_1 \cup P_2 \cup \dots \cup P_n) \setminus (P_1 \cap P_2 \cap \dots \cap P_n)$$

For example, let us consider the following actions:  $a_1 = \text{“Pass at green light”}$  and  $a_2 = \text{“Stop at red light”}$  with the following preconditions:  $P_1 = \{\text{“TrafficSign is Light”}, \text{“Light hasColor Green”}\}$  and  $P_2 = \{\text{“TrafficSign is Light”}, \text{“Light hasColor Red”}\}$ . The common precondition of the two actions is that there must be a traffic light. The relevant information here is the color of the light. As long as this information is unknown, this situation can be interpreted, at least, in two ways: either the light is green, therefore the action *“Pass at green light”* is relevant, or the light is red, therefore the action *“Stop at red light”* is relevant. Thus, according to the output of this function, the view to the traffic light must be obstructed in the simulation.

**Using fuzzy information to emphasize uncertainty** by casting impreciseness or vagueness in information. We define the fuzzifier function  $f(A_i, \varepsilon)$  which inputs are an assertion  $A_i$  and a threshold  $\varepsilon \in [0, 1]$  that represents the degree of fuzziness to go below it. The output of the function is a set of assertions  $A$  that correspond to a world state that needs to be reached. For example, the main character is followed by a Car. To make this information fuzzy we can provoke a fog:

$$f(\text{“LearnerCar BehindObject Car”}, 0.5) = \{\text{“Weather Is Foggy”}\}$$

**Casting doubt on sources to provoke independent assessment** by adjusting the world state in order to reduce the credibility of the sources. In a fuzzy representation, each source of information has a degree of credibility. We define the function  $f(s, \varepsilon)$  which inputs are the source  $s$  (e.g. object, character) and a threshold  $\varepsilon \in [0, 1]$  that

represents the degree of credibility to go below it. The output of the function is a set of assertions  $A$  that correspond to a world state that needs to be reached. For example, the main character wants to ask a pedestrian for directions. One way to reduce this source's credibility would be to make this pedestrian drunk:

$$f(\text{"Pedestrian", 0.2}) = \{\text{"Pedestrian is Drunk true"}\}$$

**Exposing inconsistencies to create a space of interpretation** by providing information that is conflictual with the learner's mental model. We define the function  $f(m)$  which input is the mental model of the learner  $m$  (set of assertions). The output is the set of assertions that are contradictory with  $m$ . For example, if the traffic light is red, one way to create a conflictual situation is to turn on the green light too.

$$f(\{\text{"TrafficLight Color Red"}\}) = \{\text{"TrafficLight Color Green"}\}$$

## 5 Discussion and Evaluation

This conceptual proposition is in need of evaluation. Firstly, we need to evaluate that the generated situations are truly ambiguous. To achieve that, we propose to confront individuals to both generated and scripted (written beforehand) situations. The comparison between how these two types of situations are perceived by the individuals will give us an indication about how successful is the system in generating ambiguity. Secondly, we need to investigate how confronting individuals to ambiguous situations improves their non-technical skills. To do so, we propose to study how they reduce the ambiguity before and after confronting them to a large panel of ambiguous scenarios.

## References

1. Burkhardt, J.-M., et al.: Simulation and virtual reality-based learning of non-technical skills in driving: critical situations, diagnostic and adaptation. *IFAC-PapersOnLine* **49**(32), 66–71 (2016)
2. Ellsberg, D.: Risk, ambiguity, and the savage axioms. *Q. J. Econ.* **75**(4), 643 (1961)
3. Camerer, C., Weber, M.: Recent developments in modelling preferences: uncertainty and ambiguity. *J. Risk Uncertain.* **5**, 325–370 (1992)
4. Hoffman, F.O., Hammonds, J.S.: Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Anal.* **14**(5), 707–712 (1994)
5. Blockley, D.: Analysing uncertainties: towards comparing Bayesian and interval probabilities'. *Mech. Syst. Signal Process.* **37**(1–2), 30–42 (2013)
6. Gaver, W.W., Beaver, J., Benford, S.: Ambiguity as a resource for design. In: *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2003*, p. 233, January 2003
7. Cottone, P., et al.: 'solving' ambiguity in the virtual space: communication strategies in a collaborative virtual environment. *Cogn. Technol. Work* **11**(2), 151–163 (2009)

8. Mantovani, G., Gamberini, L., Martinelli, M., Varotto, D.: Exploring the suitability of virtual environments for safety training: signals, norms and ambiguity in a simulated emergency escape. *Cogn. Technol. Work* **3**(1), 33–41 (2001)
9. Raybourn, E.M.: Adaptive thinking & leadership simulation game training for special forces officers. In: *Interservice/Industry Training, Simulation, and Education Conference* (2005)



# Digging for Gold: Motivating Users to Explore Alternative Search Interfaces

Angela Fessel<sup>1</sup>(✉), Alfred Wertner<sup>1</sup>, and Viktoria Pammer-Schindler<sup>2</sup>

<sup>1</sup> Know-Center GmbH, Inffeldgasse 13, 8010 Graz, Austria  
{afessl, awertner}@know-center.at

<sup>2</sup> Institute for Interactive Systems and Data Science,  
Graz University of Technology, Graz, Austria  
vpammer@know-center.at

**Abstract.** In this demonstration paper, we describe a prototype that visualizes usage of different search interfaces on a single search platform with the goal to motivate users to explore alternative search interfaces. The underlying rationale is, that by now the one-line-input to search engines is so standard, that we can assume users' search behavior to be operationalized. This means, that users may be reluctant to explore alternatives even though these may be suited better to their context of use/search task.

**Keywords:** Search user interface · Search behavior · Reflective learning  
Reflection guidance

## 1 Introduction

Searching is a key activity in knowledge work, however, people still experience problems in finding the information they are looking for [2]. Very often, people use the same search behavior independent of the information they are looking for or how successful they are. Users do not tend to use other functionality, even though this may be more efficient, depending on the query. This can be explained with the fact that on the one hand searching by now is already operationalized by many search engine users as things a user does often is internally operationalized [8]. On the other hand, motivating users to leave their comfort zone and adopting new search strategies could require a significant investment of time and effort, which is not easy to achieve [9].

Overall, we aim to design technology that stimulates users to explore alternative search interfaces, to avoid being stuck in a local optimum due to operationalization.

## 2 Background and Related Work

**Search Expertise:** In literature, one can find many papers trying to characterize the differences between search experts and novices, and characteristics to predict search performance [2]. These characteristics consist for example beside the number of hours spent per week on the web to search also of performing searches as part of the job for several years. Bateman et al. [2] identifies behavioral differences like the number of

search terms used, the usage of advanced operators or the time to complete a task. Moraveji et al. [10] defined measures of search performance relevant for search expertise like for example the domain expertise, the knowledge of the search engine's feature, general literacy and knowledge of search resources. In addition, learning how to search is learning in a continuously changing environment as technology evolves rapidly [4] and the expertise of a search expert "continues to develop" [12]. Aula and Northman [1] highlight that although experts and novices, if differentiated amongst using experience, do exhibit different search behavior on the same search interface, but that relationship to search performance is more unclear. The authors then go on to model differences between successful and less successful searchers (decided on based on task completion speed, and correctness of task results), but the model can only partially be used to guide searching. One possibility on how to educate a user to search more efficiently is by providing hints that raise users' awareness of available features reported in [11]. To sum up: the above studies tend to agree that search experience leads to better search performance, is a continuous learning process, and hinting at available features may increase search performance.

**Reflection Guidance:** By reflective learning we understand to re-evaluate past (search) behavior or (search) experiences in order to change future (search) behavior [3]. In literature, there exist different types of technologies like prompts or visuals [5] that aim to actively guide reflective learning. Reflective prompts, which we understand as interventions that consist of small text messages or questions trying to motivate a user to reflect are seen as very promising approach to stimulate reflection [6]. Also visualizations of relevant data can foster reflection like in Malacria et al. [9] who uses a visualization called "skillometer" for improving the usage of keyboard shortcuts combined with reflective questions motivating to improve the own behavior. Most literature on search behavior assumes standard text-based search interfaces, and we can safely assume that the one-line-text input field used by Google or Bing is the de facto standard search interface for near to all Internet users. Leaning on activity theory [see e.g., 8 for a recent authoritative book], we can therefore assume that we need to understand the concrete search behavior of users, in terms of using a given search user interface with its different features, as operations. These are "routine processes [...]" of which "people are typically not aware [...]". Over the course of learning and frequent execution, a conscious action may transform into a routine operation" (ibid, p62ff). Following this understanding, we see the role of reflection guidance in "deautomatizing" (ibid, p63) the operationalized search behavior, and bringing search behavior up to the level of conscious evaluation and thereby making it an object of active reflection and learning for users.

### 3 Demo: Reflective Widget in a Search Platform with Alternative Interfaces to an Underlying Search Engine

Based on the above described literature, we have designed and implemented a widget on a search platform. This platform itself provides besides a typical one-line-input and an advanced search, also different visualization for exploring the search results like a



concept graph, a keyword ranking visualization (uRank), a tag cloud and top concepts and top resources visualizations. An implemented activity tracking tool, tracks all activities a user is performing on the platform. The implemented widget consists of two parts, the search behavior visualization and the reflection guidance. The search behavior visualization (see Fig. 1, point 1) mirrors back the feature usage of a user on a search platform. This widget component is strongly influenced by the work of Malacria et al. [9], and has in addition been discussed with knowledge work professionals in focus groups [7]. The goal of this visualization is to raise the user’s awareness of the own feature usage and of other existing features. The reflection guidance (see Fig. 1, point 2) part presents reflection prompts adapted to the user’s search history, e.g. how long the user is already on the platform and the search activities conducted. For example, while a new user gets a sentence starter that makes her aware of a new, not used functionality on the platform, an experienced user receives a questions on how a feature might have influenced the own search performance. This part of the widget aims to breakdown operationalized behavior to the level of conscious action again, thereby opening up the possibility for re-shaping own behavior.

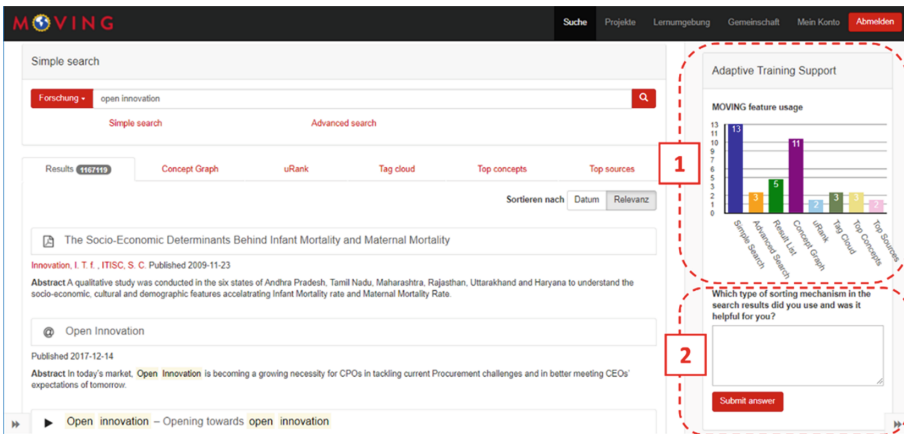


Fig. 1. Learning-how-to-search widget in the search platform

## 4 Research Questions and Outlook

The research questions that we aim to address in future follow-up work directly are:

- RQ1: Does the visualization of own feature usage stimulate users to explore unknown/little known features?
- RQ2: Do the reflection prompts generate insights relevant to searching
- RQ3: Do users change their search behavior given reflection prompts?

As a next step, we aim to set up a controlled experiment that answers RQ1 and RQ2. To answer RQ3, we aim to set up a quasi-experimental multi-week field study.

**Acknowledgement.** The project “MOVING - TraininG towards a society of data-saVvy inforMation prOfessionals to enable open leadership iNnovation” is funded under the Horizon 2020 of the European Commission (project number 693092). The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

## References

1. Aula, A., Nordhausen, K.: Modeling successful performance in web searching. *J. Am. Soc. Inf. Sci. Technol.* **57**, 1678–1693 (2006)
2. Bateman, S., Teevan, J., White, R.W.: The search dashboard: how reflection and comparison impact search behavior. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1785–1794. ACM, May 2012
3. Boud, D., Keogh, R., Walker, D.: Promoting reflection in learning: a model. In: *Reflection: Turning Experience into Learning*, pp. 18–40. Routledge Farmer, New York (1985)
4. Edwards, S.L., Bruce, C.S.: Reflective internet searching: an action research model. *Learn. Organ.* **9**(4), 180–188 (2002)
5. Fessler, A., Blunk, O., Prilla, M., Pammer, V.: The known universe of reflection guidance: a literature review. *Int. J. Technol. Enhanced Learn.* **9**(2–3), 103–125 (2016)
6. Fessler, A., Wesiak, G., Rivera-Pelayo, V., Feyertag, S., Pammer, V.: In-app reflection guidance: lessons learned across four field trials at the workplace. *IEEE Trans. Learn. Technol.* **10**(4), 488–501 (2017)
7. Fessler, A., Pammer, V., Wiese, M., Thalman, S.: Improving search strategies of auditors – a focus group on reflection interventions. In: *Proceedings of the 7th Workshop on Awareness and Reflection in Technology Enhanced Learning co-located with the 12th European Conference on Technology Enhanced Learning* (2017)
8. Kaptelinin, V., Nardi, B.A.: *Acting with Technology. Activity Theory and Interaction Design*. MIT Press, Cambridge (2009)
9. Malacria, S., Scarr, J., Cockburn, A., Gutwin, C., Grossman, T.: Skillometers: reflective widgets that motivate and help users to improve performance. In: *UIST 2013 Conference Proceedings: ACM Symposium on User Interface Software & Technology*, pp. 321–330 (2013)
10. Moraveji, N., Morris, M.R., Morris, D., Czerwinski, M., Riche, N.: ClassSearch: facilitating the development of web search skills through social learning. In: *Proceedings of CHI*, pp. 1797–1806 (2011)
11. Moraveji, N., Russell, D., Bien, J., Mease, D.: Measuring improvement in user search performance resulting from optimal search tips. In: *Proceedings of SIGIR*, pp. 355–363 (2011)
12. Tucker, V.M.: The expert searcher’s experience of information. In: Bruce, C., Davis, K., Hughes, H., Partridge, H., Stoodley, I. (eds.) *Information Experience: Approaches to Theory & Practice*, pp. 239–255. Emerald Group Publishing, Bingley (2014)



# The Role of Ubiquitous Computing and the Internet of Things for Developing 21<sup>st</sup> Century Skills Among Learners: Experts' Views

Olga Viberg<sup>(✉)</sup> and Anna Mavroudi

KTH Royal Institute of Technology, Stockholm, Sweden  
{oviberg, amav}@kth.se

**Abstract.** This explorative study aims to understand the role of ubiquitous computing and the IoT for developing and practicing learners' 21st century skills. Data was collected from ten expert interviews. Based on the conventional content analysis, our results suggest that the integration and use of such technologies in learning settings can enable the development of the learners' 21st century skills. Also, our findings identified several success factors and challenges that have to be considered when developing and practicing the identified skills. The paper is of interest to practitioners, researchers and to educational policymakers, since our study's results can guide them in planning effective learning interventions that exploit ubiquitous computing and the IoT with the aim to cultivate 21st century skills among learners.

**Keywords:** 21<sup>st</sup> century skills · Ubiquitous computing · Internet of Things

## 1 Introduction

As the integration of such cutting-edge technologies as ubiquitous computing and the Internet of Things (IoT) into people's lives continues to increase, educational modules on these technologies for the development of basic and advanced skills follow the same trend [1]. Provided that these technologies can extend the scope of what is possible in teaching and learning [1], there is a need to better understand if these technologies can be used to prepare tomorrow's citizens by cultivating their 21st century skills and if so, how. In the literature, there is an increasing body of research investigating the affordances of these technologies with respect to the cultivation of learners' 21st century skills. Examples include an intervention on microcontrollers and the IoT which helped students solved complex problems and improved their algorithmic thinking [3]. Kong et al. [4] discuss problem-solving skills cultivation using a smart ubiquitous learning system exploiting IoT to simulate authentic activities and detect learning behaviours.

## 2 Method

The data was collected from ten semi-structured interviews with experts, two females and eight males, from Greece, Netherlands, UK, Cyprus and Japan. The participants

have ascribed the status of experts in their posts and have (i) specialist professional or technical knowledge, (ii) knowledge of organisational procedures and processes, and (iii) interpretive knowledge about their field [5]. They have worked with ubiquitous computing and/or IoT technologies and used them as means to cultivate learners' 21st century skills; they are faculty members and/or researchers, and/or educational policy-makers in technology-enhanced learning (TEL). The interviews were based on an interview protocol, which was pilot tested and validated, and were conducted in an open manner [5]. Each interview lasted for about 30 min. All the interviews were conducted online by two researchers separately (five interviews each) and recorded, with the informed consent of the respondents. They were later transcribed and validated by the respondents. The data were coded by the two researchers independently and analysed via conventional content analysis, avoiding preconceived categories and allowing categories and names for categories to derive from the data [6].

### 3 Findings

Context: the majority of the projects mentioned by the experts applied to formal education. In particular, the primary or secondary education (school) context was often discussed (8 respondents), followed by university settings (5 respondents). In the school setting, examples of projects include educational robotics, or national students' competitions aiming to familiarise students with basic research procedures, herein with a focus on STEM. Another project revolved around the use of ubiquitous computing and the IoT as educational means per se, entailing the application of learning scenarios for these topics taking place in the classroom. This project was part of a European funded project focusing on using just technologies in STEM school education. In the university education context, one respondent (R3) was involved in professional development workshops for academic; two respondents mentioned measuring via biomedical data sensor technology, to help university students perform better when they take exams (R6) or to give better presentations (R10); two interviewees were involved in non-formal settings and one was involved in vocational training projects of athletes in semi-formal settings. An example of non-formal learning projects included projects on migrants' language learning.

Cultivating 21st Century Skills: communication was discussed often (7 respondents), followed by collaboration (6 respondents), critical thinking (5 respondents), and creativity (4 respondents). Respondent 10, for instance, emphasised the cultivation of communication skills, including presentation skills: "presenting is a subpart of the 21st century communication part [...]. In one project we focused on the [...] presentation and communication skills and especially non-verbal behaviour, as this is also you can more easily analyse with sensor technology". Collaboration was often strengthened via group work combined with project-based and/or game-based learning; for example, Respondent 4 said: "benefits of the technology enabled multiplayer real-time collaboration and interaction regardless of distance between players and that students were in general familiar and comfortable with digital gaming." With respect to critical thinking, it was frequently associated with students' investigations. As stated by Respondent 1: "they learn the skills of managing a learning science investigation, so the skills of being a scientist, i.e., the

skills that include critical thinking, problem solving, self-regulation.” Creativity was linked to novel and effective solutions: “They took some initiatives that I did not initially expect” (R2). The development and the assessment of learners’ creativity was underlined to be a challenging task, as highlighted by the Respondent 8: “I believe that it is difficult both to define and to develop creativity, it is not clear how to approach that.”

Apart from the 4C’s, the participants highlighted the development of skills as problem-solving (6 respondents), e.g., “[the students] definitely enhanced their problem solving because they simply got rapid feedback and self-regulation” (R3) and computational thinking (4 respondents). The latter was often related to STEM projects. Four participants discussed the development of self-regulation while using ubiquitous computing or IoT technologies (R 1, 2, 3 &9). For instance, the goal of two projects carried out by Respondent 2 was to “create autonomous learners”. Lastly, a project involved more generic skills as social and ethical skills: “we’ve tried to focus not only in problem, or cognitive skills, but also more generic social and ethical skills” (R1).

Success factors involve (i) support, (ii) technology affordances, (iii) personal interest and motivation, and (iv) game-based learning features. Regarding support, the respondents emphasized teacher development, i.e., that teachers need continuous support, both in terms of how to use the technology, but also in how to work with the novel learning scenarios (R1&5). The driving force is “the change of the ways [tutors] normally use for teaching” (R5). For the researchers leading and participating in such initiatives, it is critical to get the leadership support (R2). Learner support and scaffolding are also central: “the investigations and the activities are needed to be scaffolded, not just in a classroom, but also outdoors, so learners need to be guided [...] in a clear way” (R1). Learners’ personal interest and motivation involve using the novel technologies, the new learning scenarios, or a combination of both; for example, students’ personal interest for robotics allied with ubiquitous computing in a national student competition (R2). Also, the importance of personal meaning for the ubiquitous technology-supported investigations was stressed: “If interest is meaningful then the learners will be engaged [...], so instead of doing an investigation that a teacher tells you, conduct an investigation that has meaning for you, either because it is about yourself, or a local community or something that personally interests you” (R1). The personal motivation aspect, both on the part of the tutors and the students, was also mentioned by the Greek experts (R 5&6). The maturity of technology, its usability and robustness is another recurrently mentioned success factor. Characteristic statements include: “it is about innovative but mature enough technology” (R5), “the technology was seamless [...], we could make sure that it was robust and usable” (R1), “how precise can the system diagnose, record and analyze user behaviour” (R10). The fact that learners, in several projects (R 4, 6, 7), saw the suggested technology-supported learning scenarios as a game contributed to the projects’ success in cultivating students’ 21st century skills; as stated by Respondent 7: “they saw it as a game, but they were actually learning.” Others designed learning scenarios in a game-based form agreeing that it can be a success factor (R4).

Challenges pertain to scalability, sustainability, and technology. Four respondents explicitly highlighted scalability stressing the need for, e.g., adequate “financial support and appropriate organization/institutional/legal framework” (R5), referring to challenges in organising sustainable communities (R1), sustained commitments (R1&2),

partners' availability and sustainable technology solutions: "Sustainability of the solution that was produced as well became more apparent as a problem" (R9). Time and equipment, e.g., more educational robots (R7), are also prerequisite needed to scale up such projects. Finally, half of the respondents underlined technology limitations, such as technology not being functional, robust and mature enough (R 1&5).

## 4 Discussion and Conclusions

The discussed projects were successful in cultivating learners' 21st century skills at a small scale. To scale up and sustain such projects a suggestion would be to consider *teacher development and support*. The importance of supporting teachers in the process of developing students' 21<sup>st</sup> century skills, in a combination with the use of new ICT, has been similarly accentuated in recent research [2]. Teachers need to be continuously supported on how to use novel technologies and how to work with the innovative learning scenarios. Similarly, researchers need to come closer to the teachers' needs and practices. One way of achieving this is to include teachers as co-designers of new learning scenarios. The development of *sustainable learning communities* is another important prerequisite for scalability. This is a challenging task, which requires planning, implementing, and promoting sustainability goals that benefit the individuals, the community and the educational policy. Finally, the *maturity of the used technology*, its *usefulness and robustness* is a significant requirement that should not be underestimated. The contribution of this paper pertains to the identification of contexts, enabling factors and challenges for the cultivation of the 21<sup>st</sup> century skills in relation to the exploitation of the ubiquitous computing and the IoT technologies, hence potentially informing future educational policies and strategies, as well as curriculum designers.

## References

1. Delaney, K., O'Keeffe, M., Fragou, O.: A design framework for interdisciplinary communities of practice towards STEM learning in 2nd level education. In: Auer, M.E., Guralnick, D., Simonics, I. (eds.) ICL 2017. AISC, vol. 715, pp. 739–750. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-73210-7\\_86](https://doi.org/10.1007/978-3-319-73210-7_86)
2. Valtonen, T., et al.: Insights into finnish pre-service teachers' twenty-first century skills. *Educ. Inf. Technol.* **22**, 2055–2069 (2017)
3. Magdalinou, K., Papadakis, S.: The use of educational scenarios using state-of-the-art it technologies such as ubiquitous computing, mobile computing and the internet of things as an incentive to choose a scientific career. In: Auer, M.E., Tsiatsos, T. (eds.) IMCL 2017. AISC, vol. 725, pp. 915–923. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-75175-7\\_89](https://doi.org/10.1007/978-3-319-75175-7_89)
4. Kong, X., Chen, G., Huang, G., Luo, H.: Ubiquitous auction learning system with TELD (Teaching by Examples and Learning by Doing) approach: a quasi-experimental study. *Comput. Educ.* **111**, 144–157 (2017)
5. Littig, B., Pöchhacker, F.: Socio-translational collaboration in qualitative inquiry: the case of expert interviews. *Qual. Inq.* **20**(9), 1085–1095 (2014)
6. Hsieh, H.-F., Shannon, S.: Three approaches to qualitative content analysis. *Qual. Health Res.* **15**(19), 1277–1288 (2005)



# An Exploratory Study on Student Engagement with Adaptive Notifications in Programming Courses

David Azcona<sup>1</sup>, I-Han Hsiao<sup>2</sup>, and Alan Smeaton<sup>1</sup>

<sup>1</sup> Dublin City University, Glasnevin, Dublin 9, Ireland  
{David.Azcona, Alan.Smeaton}@insight-centre.org

<sup>2</sup> Arizona State University, Tempe 85281, USA  
Sharon.Hsiao@asu.edu

**Abstract.** This paper presents a study on students' engagement and personalized weekly performance notifications. Students were offered to voluntarily opt-in to receive customized notifications regarding their predicted course performances and recommended resources. In addition, the predicted at-risk students were also recommended with code solutions from higher performers in the class. Data was collected from Computer Science programming courses. Students' engagement with the notifications and resources were tracked and have been found to be an indicator of their differential improvement between their exams.

**Keywords:** Computer science education · Learning analytics  
Engagement · Predictive model · Personalized notification

## 1 Introduction

In this work, we explore how predictive analytics models work in distinguishing students struggling with programming courses. We implemented multimodal models for each course that aggregates sources of student data: student characteristics, prior academic history, students' programming laboratory work, and logged interactions between students' offline and online resources. Classification models are built by developing features and extracting patterns of success on these courses, then trained with two years of ground truth data and cross-validated, and finally predictions are generated every week with incoming student data. A report containing whether each student is likely to pass or fail the next formal assessment and their associated confidence is sent to the lecturers for each course. During the second part of the semester, typically after the first laboratory computer-based examination, students are free to opt-in to receive weekly personalized notifications. These notifications are sent via email and contain information regarding their predicted performance, based on the student data modalities gathered such as their progress with laboratory sheets; programming code solutions, from predicted top-ranked students within the same class; and

university resources to reach out for help if needed, such as Student Support, the course's lecturer or our system. The accuracy of the predictions generated is crucial as students will receive a customized message regarding their predicted performance and code recommendations for failed submissions from higher performers in the class if they are below a performance threshold. In our work, we measure the engagement with these customized notifications and how that could be an indicator of their performance. The research questions are stated as the following:

**RQ1:** How accurately are predictive models able to classify students in programming modules for new cohorts of students?

**RQ2:** What are the effects for students that engage with customized performance and programming feedback notifications?

## 2 Literature Review: Adaptive Feedback in Learning

Feedback is one of the most effective methods in enhancing student's learning [4]. There is an abundance of factors that affect educational achievement. Some factors are more influential than others. For instance, feedback types and formats, timing of providing feedback, etc. Studies have reported that positive feedback is not always positive for students' growth and achievement; "critical" rather than "confirmatory" is the most beneficial for learning regardless of whether feedback was chosen or assigned [3]; content feedback achieves significantly better learning effects than progress feedback, where the former refers to the qualitative information about the domain content and its accuracy, and the latter describes the quantitative assessment of the student's advancement through the material being covered. Several of the different feedback factors were explored on the intersections with the learner's variables (i.e. skills, affects) and reported to support personalized learning. For instance, cognitive feedback was found to make a significant difference in the outcomes of both student learning gains in an intelligent dialogue tutor; student's affects were being adapted to improve motivational outcome (self-efficacy); using student characteristics as tutoring feedback strategies to optimize students' learning in adaptive educational systems. While a large body of empirical studies investigate the feedback impacts in the context of learning, less is focused on researching adaptive notification as feedback in programming courses.

## 3 Research Methodology

Programming modules in our institution are being delivered through a Virtual Learning Environment that allows students to access the material online and verify their computer-based programming work. The student programming digital footprint gathered is then leveraged using Artificial Intelligence techniques and combining them with other student data modalities to identify students having issues [1] and adapt their learning on this discipline [2]. At the middle



of the semester, a feature is enabled for students to opt-in or opt-out of weekly personalized notifications. These include a performance message based on the predictions being run on the incoming class and trained with historical student cohorts’ data; recommended material and laboratory sheets resources to review based on their progress; programming code solutions from top-ranked students in the class and additional support resources. A gain index is developed to measure the student’s improvement between two examinations, see Eq. 1, and normalized to output values between  $-1$  and  $1$  on Eq. 2:

$$gi(e1, e2) = \frac{(e2 - e1)}{e1} \tag{1}$$

$$normgi(e1, e2) = \begin{cases} 1 & e1 = 0 \\ 1 & gi(e1, e2) > 1 \\ gi(e1, e2) & otherwise \end{cases} \tag{2}$$

## 4 Results

We will now analyse the results obtained by running predictions and sending adaptive feedback on new cohorts of students in 2017/2018 and what this means for the research questions proposed for one of the courses: Shell Scripting for first-year students.

### 4.1 RQ1: Predictive Modelling

Table 1 shows an increasing accuracy and F1 metrics from the first assessment to the last. That is, as more data is collected around student engagement, we are better able to distinguish students struggling with the material and, thus, giving them more accurate performance notifications.

### 4.2 RQ2: Normalized Gain for Different Groups

Table 2 shows the groups analysed: opt-ins vs. opt-outs and engaged-with-the-notifications vs. not-engaged. Students that opted-in in week 7 showed a greater

**Table 1.** Exam weeks, model’s at-risk prediction rates, passing rates, prediction results and correlations between the prediction confidence and the actual results

Exam week	Predicted at-risk	Passing rate	Accuracy	F1	Precision	Recall	Correlation coefficient
W7	51.32%	67.11%	65.79%	70.45%	83.78%	60.78%	43%**
W12	40.79%	72.37%	84.21%	88%	97.78%	80%	65%**

\*\*  $p - value < 0.01$

normalized gain compared with students that opted-out between pairs of examinations. Students that engaged with the notifications by clicking on any of the resources (material or laboratory sheets), which were not many, also showed a greater normalized gain compared to students that did not.

**Table 2.** Normalized gain improvement between student groups created

First exam week	Second exam week	Group (Number)	Mean (Std.Dev.) Exam-1	Mean (Std.Dev.) Exam-2	Mean (Std.Dev.) <i>normgi</i>
W7	W12	Opt-IN (45)	68.33% (34.32%)	77.33% (29.69%)	+27.41% (55.92%)
		Opt-OUT (5)	90% (20%)	92% (16%)	+4% (8%)
W7	W12	Engaged (4)	50% (30.62%)	90% (10%)	+70% (51.96%)
		Did-not-engage (67)	62.31% (38.72%)	64.18% (37.78%)	+20.70% (61.06%)

## 5 Conclusion and Future Work

Engaging with personalized notifications is proven to have a positive effect on the defined normalized gain index between two different examinations. However, this improvement has not yet been found to be significant. In the near future, we are exploring how students engage with the programming code solutions from higher performers and how it affects their programming design learning.

## References

1. Azcona, D., Hsiao, I.-H., Smeaton, A.F.: PredictCS: personalizing programming learning by leveraging learning analytics. In: Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK 2018), pp. 462–468 (2018)
2. Azcona, D., Smeaton, A.F.: Targeting at-risk students using engagement and effort predictors in an introductory computer programming course. In: Lavoué, É., Drachler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 361–366. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66610-5\\_27](https://doi.org/10.1007/978-3-319-66610-5_27)
3. Cutumisu, M., Schwartz, D.L.: Choosing versus receiving feedback: the impact of feedback valence on learning in an assessment game. In: EDM, pp. 341–346 (2016)
4. Hattie, J., Timperley, H.: The power of feedback. *Rev. Educ. Res.* **77**(1), 81–112 (2007)



# Instrumentation of Classrooms Using Synchronous Speech Transcription

Vincent Bettenfeld<sup>1</sup>(✉), Salima Mdhaffar<sup>2</sup>, Christophe Choquet<sup>1</sup>,  
and Claudine Piau-Toffolon<sup>1</sup>

<sup>1</sup> LIUM-EIAH, 53 rue des Drs Calmette et Guérin, 53000 Laval, France  
{vincent.bettenfeld,christophe.choquet,  
claudine.piau-toffolon}@univ-lemans.fr

<sup>2</sup> LIUM-LST, Avenue Laennec, 72085 Le Mans Cedex 9, France  
salima.mdhaffar@univ-lemans.fr

**Abstract.** This paper presents the ongoing conception of a set of tools, based on automatic speech recognition for capturing communications in educational context in real time and to offer results to learners instantly. This concept is integrated in an environment to support learners in blended contexts. Its goal is to help students stay focus on the teacher's explanations and offer them greater possibilities of interactions.

**Keywords:** Learning web environment · Transcription aided learning  
User needs assessment

## 1 Project Overview

PASTEL (Performing Automatic Speech Transcription for Enhanced Learning) is a research project driven by LIUM, LS2N and Orange Labs, aiming to explore the potential of synchronous speech transcription and its applications in teaching situations. Speech transcription allows human actors to access the textual version of a sentence a few seconds after it was pronounced, and browse the whole text as they wish. This tool can help students solving comprehension problems caused by hearing, or allows them to use down time to read again a more complex section. Different researches in the literature have demonstrated the advantages of synchronous speech transcription for online courses. For example, Ho et al. [2] argued that synchronous speech transcription helps non-native English students to better understand lectures that are delivered in English. [4] mentioned that synchronous speech transcription can help students with cognitive or physical disabilities, online students (if the quality of audio communication is insufficient), or non-native speakers. To our knowledge, usage of synchronous transcriptions is very limited in pedagogical situations. As part of the project, we will additionally test tools based on other technologies, such as real-time material recommendation and thematic segmentation.

In this project, we intent to explore how the product of real-time speech transcription can help learners and teachers. We also plan to research which user needs can be satisfied, particularly in terms of information. These needs could be satisfied by content derived

from the transcription, or by using this transcription as a support of communication. Finally, this project will be the opportunity to study how to index and browsing a great quantity of data growing and being formatted in real-time.

## 2 Environment Presentation

We chose to develop our toolset as a Moodle plugin so that it can be integrated to existing platforms. Moodle is an open-source learning management system on which teachers can create online courses and enroll their students [3]. Students enrolled in a class can access our tool as they would access other activities or resources, and teachers can create a virtual classroom in the same way they would add a page to their course. Functionalities described in the following sections have been implemented in answer to a prior analysis of the practices and needs of students and teachers described in [1].

**Interface for Students<sup>1</sup>.** The material shown during lecture can vary depending on time, and the student may want to finish reading a slide before the teacher moves on to the next one, according to student practice. The environment displays the slide currently projected in class, and users have access to other slides already projected. On the left, students can watch the teacher’s video stream in real-time. Under the video feed, they can enter their questions or needs in a text field and submit it by pressing enter. The text is then sent to the teacher, and analyzed by the resource recommendation system.

This system offers a set of links on the right part of the screen, evolving in real time, redirecting to external resources. Each of these resources has a title and a preview limited to a single line of text extracted from the beginning of the document. The first five links suggested are relevant to the topic currently discussed, and changed throughout the lecture. More links can be suggested to answer the demand expressed by the students. Each of these resources can be evaluated in regard of its usefulness by the students as soon as they discover it, by either clicking on a “useful” or “irrelevant” button.

The transcription display area is located in the center of the screen and is updated synchronously. It is divided in several paragraphs, each of them being associated to the slide projected at the moment they were transcribed. Each time the professor projects a new slide, a new paragraph is created and the transcription of the current speech is added to it. Each of these paragraphs can be selected, which causes the note-taking panel to open. Besides writing down their notes, students can notify a need for further information to the teacher by clicking on a dedicated button. The material recommendation system also takes in consideration this alert, and analyzes the concepts being explained or cited in the paragraph to provide relevant information.

To prevent the students from hiding every tool whenever they need to zoom on the slide displayed, a system of panels was implemented. Using this system, students are able to select the tools they wish to keep on screen. The slideshow is displayed as big as possible depending on the remaining space. Ultimately it can be displayed at full size if all tools are hidden, in order to guarantee readability. Students can hide and show

---

<sup>1</sup> Interfaces screenshots can be accessed at <https://umbox.univ-lemans.fr/index.php/s/XdR3fDFoTr2DePY>.

every tool panel at any moment according to their needs. To support the use case of students reading the transcription comfortably for extended periods of time, it is possible to expand the transcription area.

**Interface for Teachers.** A portion of the screen gives visual feedback of the camera, displaying the same video stream as the students receive. The teacher can accordingly place themselves in the camera's field of view. In addition to this page designed for the teacher, a second page displays the slides destined to be projected to the audience. The teacher can navigate the slideshow using the mouse or the left and right arrow keys.

The right part of the screen is dedicated to a text feed showing open-ended questions asked by the audience in real time. The most recent question is stacked on top of the list, and displayed anonymously (even though the sender can be identified by searching the Moodle plugin database, to prevent eventual counter-productive behavior) as soon as it is typed and sent.

At the bottom of the screen is presented a set of indicators, two of them shown as bar graphs. The first one displays the proportion of alerts sent by students estimating the lecture's pace is too quick. If this bar grows too large, the professor can slow down their explanation, to ensure most of the audience can keep up. The second bar displays the proportion of students looking at material corresponding to a past moment. If this bar fills up, the professor may have lost his audience at a previous point, and can react by adapting their speech accordingly. The last indicator is a table detailing the three slides on which the greater number of student expressed a difficulty. If the teacher wishes to know precisely where the problem lay, they could ask student to type it, or look at which slide they asked for more explanation.

The resources recommended to students are displayed in real time, in a table. The display is more concise as the professor does not need to rate, or get a preview of the resources they selected. Only the titles appear, allowing them to recommend a one particular resource to students, or to showcase it on the projected view.

### 3 Experiment

The prototype was tested with students and a professor in actual class context in order to evaluate usability, to list the benefits of using the system, to be able to analyze usage, and consequently to improve tools supporting the least satisfying tasks. One of the main objectives was to check that the amount of information provided to testers is not a source of cognitive overload. Providing text in real time during a learning activity can trigger such overload, but conversely saving lecture content for later use can relieve students' memory, as it was studied with podcasts [5].

Since the recommendation system is in development, material was selected and tagged by the team designing the system, and associated to a particular slide. However, the system did analyze questions asked by students during the session and automatically considered their demand for more information through the interface. Resources fetched were collected from the web in real time.

The prototype was tested on March 22<sup>nd</sup> 2018, at Laval during an information and communication lecture. Beforehand, the teacher and students received a presentation of

the toolset in order to facilitate later use. During the experiment, 13 students were located in the same room as the teacher and 16 others were located in a remote classroom. During the one-hour lecture, students were given access to the platform and their activity was monitored. Afterwards, the teacher and students were interviewed.

Students were satisfied with the quality of the translation. They reported a frustration in regards of real-time communications as their open-ended questions were not correctly transmitted to the teacher, and the teacher's pedagogical scenario did not include time dedicated to review the audience's questions. Students were confused by the number of tools available simultaneously. They are themselves aware that a comprehensive interface could be useful to users familiar with the system. Yet, as users still discovering the toolset, they are not comfortable enough to both concentrate on the lecture and use every functionality offered. They still expressed the need for more flexibility, particularly the possibility of changing the position of tools on screen. Different students had different opinions concerning which element should be displayed at a large size at the core of the interface. The teacher was not accustomed to monitoring the indicators on his computer screen in real time, and quickly abandoned this behavior to adopt a more usual behavior. His evaluation of the students' activity was hence limited to the audience physically present in the classroom.

Given the high number of class configurations and personal preferences towards practices, we plan to further develop the environment by adding new tools, and a more flexible interface. Other situations than synchronous lecture are also considered.

**Acknowledgment.** The current work is supported by the ANR project PASTEL <ANR-16-CE38-0007>. The authors would like to thank the others participants of the project for their collaboration.

## References

1. Créatin-Pirolli, R., Pirolli, F., Bettenfeld, V.: Analyse des Besoins - Document de synthèse, livrable du projet PASTEL (2017)
2. Ho, I., Kiyohara, H., Sugimoto, A., Yana, K.: Enhancing global and synchronous distance learning and teaching by using instant transcript and translation. In: 2005 International Conference on Cyberworlds (CW 2005), Singapore, pp. 373–377, November 2005
3. Moodle Homepage. <https://moodle.org/>. Accessed 28 Jan 2018
4. Shadiev, R., Hwang, W.Y., Chen, N.S., Yueh-Min, H.: Review of speech-to-text recognition technology for enhancing learning. *J. Educ. Technol. Soc.* **17**(4), 65 (2014)
5. Traphagan, T., Kucsera, J.V., Kishi, K.: Impact of class lecture webcasting on attendance and learning. *Educ. Tech. Res. Dev.* **58**, 19 (2010). <https://doi.org/10.1007/s11423-009-9128-7>



# A Programming Language Independent Platform for Algorithm Learning

Bruno Burke<sup>1</sup>  , Peter Weßeler<sup>2</sup>, and Jürgen te Vrugt<sup>2</sup>

<sup>1</sup> Wandelwerk Quality Management Unit,  
Münster University of Applied Sciences, Steinfurt, Germany  
[burke@fh-muenster.de](mailto:burke@fh-muenster.de)

<sup>2</sup> Department Electrical Engineering and Computer Science,  
Münster University of Applied Sciences, Steinfurt, Germany  
{[peterwesseler](mailto:peterwesseler@fh-muenster.de), [vrugt](mailto:vrugt@fh-muenster.de)}@fh-muenster.de

**Abstract.** Teaching People to program is a crucial requirement for our society to deal with the complexity of 21st-century challenges. In many teaching systems, the student is required to use a particular programming language or development environment. This paper presents an intelligent tutoring system to support blended learning scenarios, where the students can choose their programming language and development environment. For that, the system provides an interface where the students request test data and submit results to unit test their algorithms. The submitted results are analyzed by a machine learning system that detects common errors and provides adaptive feedback to the student. With this system, we are focusing on teaching algorithms rather than specific programming language semantics. The technical evaluation tested with the implementation of Mean and Median algorithm shows that the system can distinguish between error cases with an error rate under 20%. A first survey, with a small group of students, shows that the system helps them detect common errors and arrive at a correct/valid solution. We are in the process of testing the system with a larger group of students for gathering statistically reliable data.

**Keywords:** Language-independent programming  
Tutoring system · Algorithm learning

## 1 Introduction

One of the key skills in computer science, or even engineering and technical studies, is the ability to implement algorithms and other logical constructs in programming languages. In technology enhanced learning there are various teaching tools, where the students perform programming exercises and receive feedback. In order to provide appropriate feedback, these tools limit the choice of programming language so that only a particular set of programming languages can be used in a given development environment [2, pp. 276–277]. The type of feedback

varies from a simple correct vs. wrong rating to the evaluation of the stylistic rules of the source code [2, p. 59].

In this paper we introduce an adaptive tutoring system which aims at deriving logical algorithmic errors from test data rather than the implementation. This approach allows the prototype to operate independent from the actual implementation by the student, thus being usable by a variety of common programming languages.

## 2 Pedagogical Background

In order to teach advanced programming concepts, Universities of Applied Sciences often have practical assignments, with face-to-face meetings, in addition to the lectures. Within advanced courses, this is not about learning the syntax of certain programming languages, but about complex programming tasks, including software systems consisting of several algorithms that are dependent on each other. The understanding of learning is based on the principles of Constructivism, so that the tasks should be handled as independently and practically as possible [4, p. 66]. Therefore students should decide for themselves which programming language they use and work on the task in a real development environment. The tutoring system should serve as a learning companion and should not disturb the natural development process by enforcing a specific workflow. Instead of correct vs. wrong assessment, we provide an estimation, e.g. a probability, that there may be certain error cases. Since these are only hints, this should motivate the student to reflect on their own solution. In this way, we would like to encourage the implementation of a blended learning scenario [6, p. 125] in which the student can solve the exercises in advance, using the tutoring system, and use the presence session to discuss further topics instead of troubleshooting with the supervisor.

## 3 Technical Background

The system consists of two components: A web frontend that displays task descriptions and recent feedbacks, and a backend, consisting of multiple test server, that provides test data and evaluates user's solutions. The test servers are accessible via a Web Interface and can, therefore, be used by a large variety of common programming languages that support HTTP Requests. This independence from the programming language is additionally supported by the use of a machine learning system within the test server. Depending on the context of the task (e.g. the calculation of the median for a given series of numbers), we first implement reference algorithms both for the correct solution and for incorrect solutions identified by the supervisor. We then generate test data for the given context of the task. These are used to evaluate the difference between the output of the incorrect implementation of the algorithm and the output of the correctly implemented algorithm.



The system then generates different features from these deviations, such as the relative and absolute error. These features are then used in the form of feature vectors as training data to develop an error class model using standard classification algorithms, e.g. Decision Trees, of Machine Learning (see [3, pp. 82–103]).

In the concrete exercise situation the students can work on an exercise, let their implementation process a given test data and send the result back to the test server. The system then calculates the deviation between the user's result and the result of the correct algorithm implementation and classifies the error case using the error case model. How well the system can distinguish between solutions depends heavily on the amount of result data of the task. For this purpose, interim results could also be recorded in a further development step and included in the classification process. Previously unknown incorrect solutions identified by the supervisor are manually implemented based on the student's feedback in presence sessions.

An additional component is a recommendation service where the teacher can set specific messages and links to learning resources, based on the classification results.

## 4 Use Case

Our use case shows how the system can support the implementation of an algorithm without giving restrictions on the programming language or development environment. The demo scenario is a Student who is taking a class about data processing. The student participates in an assignment of a practical lesson where they have to choose a specific exercise from the front end Website and use any programming environment to connect to the Testserver, implement a first algorithm and test it by retrieving test data from the back end, process the test data and send back the result data. On the front end feedback about the solution is given. In case of an error, the solution must be changed until the feedback indicates that the solution is correct. To adapt the learning content continuously, the prototype enables the supervisor to enrich the set of tasks by adding current or changed content. This ensures that the student always solves the subject-specific, current tasks.

Another advantage is that the teachers can use the incorrect solutions indicated by the system to identify possible difficulties of the students in the learning process and thus prepare themselves better for their lessons. That aims at making teaching more efficient for teachers and students to improve the teaching experience and learning outcomes or overall learning success.

## 5 Results and Outcomes Achieved

The system currently has three exercises with which it has already been tested both technically and by a small group of five students. The technical evaluation

**Table 1.** Classifier evaluation for three exercises.

Exercise	Classes	Data	Error-Rate	Cohen's kappa
Arithmetic mean	4	4000	0.0733	0.9023
Median	6	6000	0.1990	0.7612
Moving average	5	5000	0.0000	1.0000

was done using 1000 entries of generated test data for each error case and evaluate a C4.5 decision tree classifier [7] with a 10-fold cross-validation [1, pp. 372–375].

Table 1 shows that the error cases of the arithmetic mean classify correctly in 92.6% of the cases and that the Cohen's kappa [5] reaches a value of 0.9. The median error cases classify correctly in 80,1% of the cases, and a Cohen's kappa of 0.76 is reached. The moving average error cases are 100% correctly classified. That shows our System identifies the correct error cases on average with a precision of 90.9%. That means the students are highly likely to receive the accurate feedback on the actual error case (Table 1).

The small group of students who tested the prototype rated it as helpful. It turned out that the system helped those students who thought they had a correct solution, by pointing out a potential mistake. The additional effort of using the API connection was considered acceptable. We are in the process of testing the system in the courses “Pattern Recognition and Machine Learning” and “Automatic Speech Processing”, where the students make sure that they have properly implemented the basic algorithms and can fall back on them for following tasks like Feature Extraction required in both subjects. We test the system with a larger group of students and exercises with different levels of difficulty for gathering statistically reliable data.

## References

1. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press Inc., New York (1995)
2. Bott, O.J., Fricke, P., Priss, U., Striewe, M.: Automatisierte Bewertung in der Programmierausbildung. Waxmann Verlag, Münster (2017)
3. Flach, P.: Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press, Cambridge (2012)
4. Helmke, A.: Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts. Kallmeyer, Seelze-Velber (2009)
5. Vieira, S.M., Kaymak, U., Sousa, J.M.C.: Cohen's kappa coefficient as a performance measure for feature selection. In: International Conference on Fuzzy Systems, pp. 1–8, July 2010. <https://doi.org/10.1109/FUZZY.2010.5584447>
6. Waldherr, F., Walter, C.: Didaktisch und praktisch: Ideen und Methoden für die Hochschullehre, 2nd edn. Schäffer-Poeschel, Stuttgart (2014)
7. Wu, X., et al.: Top 10 algorithms in data mining, January 2008. <https://doi.org/10.1007/s10115-007-0114-2>



# Using Thematic Analysis to Understand Students' Learning of Soft Skills from Videos

Björn Sjöden<sup>1</sup>✉, Vania Dimitrova<sup>2</sup>, and Antonija Mitrovic<sup>3</sup>

<sup>1</sup> Halmstad University, Halmstad, Sweden  
bjorn.sjoden@hh.se

<sup>2</sup> University of Leeds, Leeds, UK  
v.g.dimitrova@leeds.ac.uk

<sup>3</sup> University of Canterbury, Christchurch, New Zealand  
tanja.mitrovic@canterbury.ac.nz

**Abstract.** Learning from visual and social media makes a complex area of study and a vital part of understanding the development of 21st century skills. The Active Video Watching (AVW) platform was developed in order to scaffold students' active learning of soft skills from videos, by encouraging users to engage with the content (e.g. marking important aspects and writing comments). Previous studies of AVW used learning analytics to identify student comments which can be used in "intelligent nudges" for triggering reflection among others who watch the same video. Here, we describe the methodology and reasoning for conducting a qualitative thematic analysis of such comments, with respect to learning presentation skills. Our aim is to uncover additional learning opportunities from the data and how they might be explained within a broader theoretical framework of observational learning. As a basis for discussion, we present a preliminary thematic map of the results and how students' remarks on good/bad examples in the videos relate to the types of knowledge they gain from it. We suggest several resulting topics for future study.

**Keywords:** Video-based learning · Soft skill learning · Thematic analysis  
Interactive systems · Intelligent support

## 1 Introduction

Over the past century, educational technology has developed from offering procedural support, such as the calculator and word-processor, to also facilitating navigation, selection and assessment of content, such as search engines and virtual assistants. This raises questions as to the role of human abilities and learning in relation to AI and machine learning. In digital schools, teachers and researchers alike need to ask when and to what extent human judgment is needed of data that are automatically collected and fed back to learners.

Here, we exemplify and discuss a method for understanding learners' knowledge needs when using technological support for learning soft skills, such as communication and doing presentations. The starting point is a web-based platform called Active Video Watching (AVW) which supports student' learning from videos (e.g. on YouTube).

Students are encouraged to actively elaborate the content, by highlighting important aspects and writing comments while they watch the video. Previous studies of the AVW with actual students, informed the development of AI-based support for intelligent “nudging”, that is, signposting and prompts which trigger reflective learning or induces opinion at key points in the videos. At this time, two versions of the AVW platform have been tested with 821 students in which more than 3,000 comments have been collected and analysed for use as nudges in the system [1, 2].

The present study adds to previous learning analytics of students' comments by exploring a qualitative analytic method known as thematic analysis [3]. Thematic analysis is used for identifying meaningful patterns (themes) within a qualitative dataset that capture something important about the data in relation to the research questions. We were interested in not only *what* students deem as important while watching the videos but also *how* and *why* specific content is emphasized and contextualized with respect to learning the target skills. Watching a real person demonstrating the relevant skills and knowledge makes a fruitful basis for observational (or vicarious) learning, which points to the relevance of social learning theory for gaining deeper insights into the findings. The outcome of this extended analysis is important for understanding students' attentional and meaning-making processes, and the basis for developing metacognitive skills and self-regulation.

Our aim was thus to uncover and characterize learning opportunities from user data that previous learning analytics have not captured, and which might be explained within a broader theory of observational learning [4]. We posed two explorative research questions, (i) What can thematic analysis bring for distinguishing relevant learning opportunities from students' comments to the videos?, (ii) What knowledge did students gain from good/bad examples in the videos?

## 2 Method

### 2.1 Dataset

The analyzed dataset comprised 335 free-response comments to four TED-like example videos of research presentations (about 3–8 min. long) used for learning presentation skills. The comments were written by 33 university students who participated in a previous study using the AVW platform [1]. Comments varied in length from 4 (e.g. “good”) to 366 characters. The dataset included data on when a comment was posted, to which of the four videos, by whom, and which aspect, if any, the student associated with the comment (*Speech, Delivery, Visual aids, Structure or No Aspect Selected*).

### 2.2 Coding and Thematic Analysis

In order to capture the full diversity of comments and identify meaningful themes across the whole dataset, we combined a data-driven (bottom up) approach with a theoretical interest-driven (top down) approach. That is, we already had some analytic preconceptions, in the sense that the system provided some mentioned predefined aspects. Also, our focus was guided by our interest in observational learning, which motivated our attention to comments about “good” (desired) behavior versus “bad” (undesired)

behavior. However, the comments were in free response format and not necessarily associated with any particular aspect (e.g. a comment labeled *Speech* by one student might be labeled *Delivery* by another). The data were thus coded inclusively, accounting for the full variety of free responses. Comments were coded using a coding scheme, by two of the authors as well as an independent rater. The different codes were then sorted into potential themes which were further reviewed and refined.

First, each comment was coded for two variables: (i) whether it represented a “good”, “neutral” or “bad” example for learning presentation skills and (ii) whether it related to the qualities of the factual “content” of the presentation (e.g. interesting points made, quality of visuals) or demonstrated “skills” of the presenter (e.g. use of body language, approach to the audience). These codes were motivated by (i) capturing both desirable and undesirable aspects (in effect, positive and negative feedback) for learning the target skill/knowledge, and (ii) applying the common distinction between declarative and procedural knowledge.

Second, the (minor) differences in coding between the raters were discussed and consensus was reached (e.g. a comment on “good visual aids” was interpreted as referring primarily to content, not skill, whereas “good use of visuals” was interpreted as commenting primarily on a skill, not content).

Third, one of the authors (who also rated the comments but had not been involved in previous studies) searched and reviewed all comments for recurring themes, relating back to the research questions, using the guidelines by Braun and Clarke [3]. The thematic analysis comprises six phases, starting with familiarizing oneself with the data (Phase 1), generating initial codes (Phase 2), searching for themes (Phase 3), reviewing/refining themes (Phase 4), and defining/naming the themes (Phase 5). Lastly, compelling examples were selected in relation to the research questions (Phase 6). It is not self-evident what makes a “theme” but taking the viewer’s perspective, one could ask: What is there to learn from? How do I understand what there is to learn?

### 3 Preliminary Findings and Topics for Further Study

We addressed the two research questions by (i) constructing a thematic map, which shows the identified main themes and subthemes of students’ comments (see Fig. 1); (ii) examining the distribution of commented qualities (good, bad, neutral) over the two types of identified knowledge (content and skills) (see Table 1).

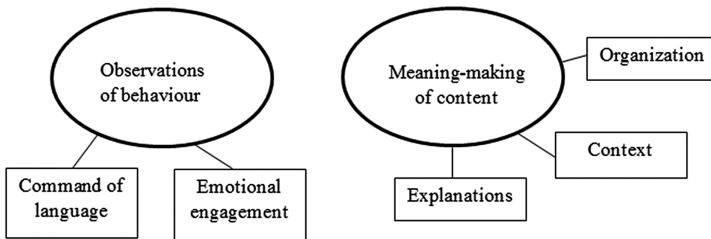


Fig. 1. Thematic map, showing the main themes and subthemes of comments to the videos.

**Table 1.** Number of commented examples in relation to types of knowledge.

Knowledge type	Example quality		
	Bad	Neutral	Good
Content	19	31	81
Skill	44	33	106

The analysis suggests two main themes of comments: explicit observations of the presenter's behavior (Fig. 1, left) and implicit meaning-making of the presented content (Fig. 1, right). Subthemes to these categories included *Command of language* (e.g. "Fantastic vocal variety", "Effective hand gestures", "Swaying is a bit distracting") and *Emotional engagement* (e.g. "use of humour", "feels a bit nervous", "boring", "cool"). As to meaning-making, there were general comments on *Organization* (e.g. "simple and well organized", "uncoordinated"), *Context* (e.g. "uses questions as transition", "consistency on the style of visual slides") and *Explanations* (e.g. "Clear diagrams that explain points well", "use of analogy").

Further (Table 1), skills were overall more frequently attended to than content (61% vs. 39% of comments) and "good" examples were overall attended to more often than "bad" examples. Nevertheless, the results suggest that students consciously attend to both positive and negative examples in order to learn both *what* to do (to reproduce) and *what not* to do (to avoid) and, hence, that the comments can inform both positive and negative verbal feedback in the system.

Topics for discussion and further study include how themes and their emotional valence relate to capturing students' attention in previously identified "high attention intervals" (HAI) of the videos [2], how the types of comments and themes relate to different student profiles, and practical educational implications, such as where human teachers best invest their efforts with respect to the support given by the AVW technology.

## References

1. Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., Mathews, M.: Supporting constructive video-based learning: requirements elicitation from exploratory studies. In: André, E., Baker, R., Hu, X., Rodrigo, M., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 224–237. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_19](https://doi.org/10.1007/978-3-319-61425-0_19)
2. Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., Weerasinghe, A.: Using learning analytics to devise interactive personalised nudges for active video watching. In: Proceedings of the 25th Conference on UMAP, pp. 22–31. ACM, New York (2017)
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006)
4. Bandura, A.: *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall Inc., Englewood Cliffs (1986)



# Formalizing CSCL Scripts with Logic and Constraints

Andreas Papasalouros<sup>(✉)</sup>

Department of Mathematics, University of the Aegean, 83200 Karlovassi, Greece  
andpapas@aegean.gr

**Abstract.** CSCL (Computer-Supported Collaborative Learning) scripts describe pedagogically effective practices for organizing the experiences of individuals when engaged in collaborative activities. This paper presents a new language named COSTLy for describing certain mechanisms of CSCL scripts, namely, group formation and task distribution. The expressiveness of the proposed language is demonstrated by describing jigsaw, a well-known CSCL script. The proposed descriptions are automatically transformed into constraint logic programs which can be executed so as to support group formation in actual scenario instances.

**Keywords:** Learning activity representation · Collaborative scripts  
Constraint logic programming

## 1 Introduction

CSCL (Computer-Supported Collaborative Learning) scripts [1] describe pedagogically effective practices for organizing the experiences of individuals when engaged in collaborative activities. They consist of proper specifications of educational scenarios defined by teachers or instructional designers. According to [3], these specifications comprise a number of components, that is, participants, activities, roles, groups and resources, as well as certain mechanisms, i.e. task distribution, group formation and sequencing.

The automation of group formation in collaborative scripts is currently an active field of research [1, 2, 4, 5]. In this paper we present a logic-based formalism for describing scripts. That is, we describe COSTLy, a new language that can express the logical conditions that define certain formations of groups. This language is expressive enough so that it can formally describe various CSCL scripts that exist in the literature. We also support the automatic translation of script descriptions in COSTLy into constraint logic programs. By running these programs with a constraint solver, we provide a reasoning mechanism for performing group formation in actual educational settings.

## 2 Towards a CSCL Script Definition Language

In the presentation of COSTLy, the proposed language, we use as an example the jigsaw script, which is widely used and studied in the relative literature. The script comprises two distinct phases, related to group formation:

- In the first *Expert Group* (EG) phase, participants are divided in expert groups so that, given a set of tasks, all participants in an expert group are assigned to the same task.
- In the second *Jigsaw Group* (JG) phase, there should be at least one student for each task within each jigsaw group.

The allocation in groups is specified in COSTLY as a partition definition. A partition is a collection of non-empty groups of participants such that every participant belongs to only one group and the union of all partition groups adds to the initial set of participants. The definition of a partition in a specific phase is expressed as a constraint. A constraint is a logical condition that is considered that holds during an assignment of a partition.

In the proposed formalism, a script definition is a set of phase descriptions. A partition of a set of participants can be defined in each phase of the script. Such a description for a partition  $P$  takes as parameters a set of participants,  $S$ , and an optional set of resources or *Tasks*. The number of groups in the partition or the size of each group must be defined. In a partition definition a logical predicate is provided that formally describes a constraint which must be satisfied by the partition for the particular phase. Thus, the first, expert phase of the jigsaw script is defined as follows:

*Example 1.* Expert phase definition.

```

1 phase EG: create-partition  $P$  for  $S, Tasks$  with  $|Tasks|$  groups:
2   forall  $T$  in  $Tasks$  exists!  $G$  in  $P$ 
3     forall  $St$  in  $G$ 
4       ASSERT( $performs(St, T)$ ).
```

where  $|Tasks|$  is the number or tasks. The predicate for a partition can be a simple binary predicate formula, such as  $reads(John, Chapter1)$  or a complex formula involving logical connectives (**and**, **or**, **not**) or, as in line 2 of Example 1, a quantifier expression: **forall** or **exists**. The latter provide universal and existential quantification over the elements of a set.

A special quantifier named **exists!** appears in line 2 of Example 1. It appears in the scope of, that is, it syntactically follows, a universal quantifier (**forall**). This form defines a relation among the elements of the two quantified sets so that the elements of the second are evenly allocated to the elements of the first. Thus, in the expression in line 2, the elements of set  $P$ , that is, the groups of the partition  $P$  are evenly associated with the tasks in set  $Tasks$ . As an example, in a specific setting with sets  $Tasks = \{a, b\}$  and  $P = \{g_1, g_2, g_3, g_4\}$ , the relation pairs  $\{(a, g_1), (a, g_3), (b, g_2), (b, g_4)\}$ , satisfy the condition in line 2.

The **ASSERT**, (meta-)predicate, used in logic programming, is a special form that asserts the truth of a certain fact, in the form of an atomic predicate that it takes as its argument. This is of particular importance since this predicate generates new relations during a certain phase of script. Thus, in Example 1, it is designated that each participant,  $St$  in group  $G$ , (line 3) is assigned to perform a certain task,  $T$  (line 4).



The description of the second phase expressed as follows:

*Example 2.* Jigsaw phase definition

```

1 phase JG: create-partition  $P_j$  for  $S, Tasks$  with  $|S|/|Tasks|$  groups :
2   forall  $G$  in  $P_j$ 
3     forall  $T$  in  $Tasks$  exists!  $St$  in  $G$ 
4        $EG.performs(St, T)$ .
```

In the jigsaw phase, the number of formed groups is the integer ratio of the number of participants, over the number of tasks,  $|S|/|Tasks|$ . The allocation of students to groups must satisfy the following constraint (lines 2–4 in Example 2): In each jigsaw group  $G$  in the new partition  $P_j$  (line 2), the students of the group must be evenly associated to each task (line 3) according the allocation of students to tasks in the previous, expert phase ( $EG.performs(St, T)$  in line 4).

Predicates such as *performs* above are not part of the language and they can be arbitrarily defined by the designer of a certain script, allowing for group formation based on arbitrary binary relations. However, there is also a number of special, built-in predicates and expressions in COSTLY such as `max`, `min` and `sum`. Also, there are defined mathematical expressions and functions, not presented here. Certain variants of a basic script can be defined by providing additional, extrinsic constraints [1], in the form of new logical conditions that are conjunctively connected with the constraint of the basic script. For example, we can alter the group formation in the expert phase so as to minimize knowledge diversity of the members among expert groups with a `min` predicate that minimizes the sum of all absolute differences between two members of each group.

### 3 Implementation and Preliminary Evaluation

As in previous works [1], the allocation of participants to groups is expressed as an integer constraint satisfaction problem. In the implementation presented here, a partition is represented as a table of binary elements where each element  $p_{ij}$  is 1 if participant  $i$  is allocated in group  $j$  and 0, otherwise. All  $M$  columns of this table have non zero sum, representing non-empty subsets of participants:  $\sum_{i=1}^N p_{ij} > 0, \forall j \in \{1, \dots, M\}$ . Furthermore, each participant is considered to participate in only one group:  $\sum_{j=1}^M p_{ij} = 1 \quad \forall i \in \{1, \dots, N\}$ .

The descriptions of the proposed language are automatically transformed into appropriate Prolog expressions containing constraints, which are executed by using the CLP(FD) constraint solver [6]. These constraints are imposed on the table  $p_{ij}$  that represents the formation of groups. The input to the transformation process is given in abstract syntax form, programmatically, by means of a Java API. The transformation is based in certain rules, some of them depicted in Table 1. The program for generating Prolog constraints was also implemented in Java. In this table, `pred'` is the generated Prolog clause head that corresponds of the initial `pred` predicate.

**Table 1.** Indicative constraint generation rules

Expression	Generated Prolog code
<code>forall A in Set: pred(Otherparams, A)</code>	<code>maplist(pred'(Otherparams,Set)</code>
<code>exists A in Set: pred(Otherparams,A)</code>	<code>member(A,Set),pred'(Otherparams,A)</code>

The generated Prolog code was run in CLP(FD) and has successfully generated group formations for the script described above in various settings of participants and tasks. Also, by using COSTLy, we have been able to define other scripts, apart from the one presented here.

## 4 Conclusions and Future Work

This paper proposed a formal representation and a reasoning mechanism, in the form of a formal language that is executable by being translated into constraint logic programs. While other works exist in the literature that support group formation by constraint satisfaction, this work proposes a new specialized language for representing CSCL scripts rather than using a generic formalism, only accessible by technical experts.

Currently we are working on a visual representation of COSTLy, so as to enable instructional designers to edit scripts in an intuitive manner. Also, we aim at providing support for roles inside groups, so as to further extend the expressiveness of the descriptions of CSCL scripts.

## References

1. Amarasinghe, I., Hernández-Leo, D., Jonsson, A.: Intelligent group formation in computer supported collaborative learning scripts. In: 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), pp. 201–203. IEEE (2017)
2. Balmaceda, J.M., Schiaffino, S.N., Pace, J.A.D.: Using constraint satisfaction to aid group formation in CSCL. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial* **17**(53), 35–45 (2014)
3. Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämmäläinen, R., Häkkinen, P., Fischer, F.: Specifying computer-supported collaboration scripts. *Int. J. Comput. Support. Collab. Learn.* **2**(2), 211–224 (2007)
4. Ounnas, A., Davis, H.C., Millard, D.E.: A framework for semantic group formation in education. *J. Educ. Technol. Soc.* **12**(4), 43 (2009)
5. Tacadao, G., Toledo, R.P.: Forming student groups with student preferences using constraint logic programming. In: Dichev, C., Agre, G. (eds.) *AIMSA 2016. LNCS (LNAI)*, vol. 9883, pp. 259–268. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-44748-3\\_25](https://doi.org/10.1007/978-3-319-44748-3_25)
6. Triska, M.: The finite domain constraint solver of SWI-Prolog. In: Schrijvers, T., Thiemann, P. (eds.) *FLOPS 2012. LNCS*, vol. 7294, pp. 307–316. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-29822-6\\_24](https://doi.org/10.1007/978-3-642-29822-6_24)



# Correction to: Lifelong Technology-Enhanced Learning

Viktoria Pammer-Schindler, Mar Pérez-Sanagustín,  
Hendrik Drachler, Raymond Elferink, and Maren Scheffel

**Correction to:**

**V. Pammer-Schindler et al. (Eds.):**

***Lifelong Technology-Enhanced Learning*, LNCS 11082,**

**<https://doi.org/10.1007/978-3-319-98572-5>**

In the original version of this paper, the affiliation of the third editor was not correct. This has now been rectified.

---

The updated version of the book can be found at  
<https://doi.org/10.1007/978-3-319-98572-5>

© Springer Nature Switzerland AG 2018

V. Pammer-Schindler et al. (Eds.): EC-TEL 2018, LNCS 11082, p. E1, 2018.

[https://doi.org/10.1007/978-3-319-98572-5\\_69](https://doi.org/10.1007/978-3-319-98572-5_69)

# Author Index

- Abolkasim, Entisar 561  
Alario-Hoyos, Carlos 355  
Albó, Laia 406, 605  
Aleven, Vincent 412  
Alexandron, Giora 74  
Alkhatib, Wael 609  
Araache, Eid 609  
Azcona, David 510, 644  
Azevedo, Otávio 231
- Barria-Pineda, Jordan 437  
Bemelmans, Roger 297, 311  
Benabbou, Azzeddine 631  
Bettenfeld, Vincent 648  
Biedermann, Daniel 579  
Boon, Peter 262  
Bowser, Amy S. 216  
Bozzon, Alessandro 467  
Brinkhuis, Matthieu 531  
Broos, Tom 399  
Brusilovsky, Peter 437, 524  
Buendía, Félix 566  
Bull, Susan 524  
Burke, Bruno 652
- Chau, Hung 437  
Chen, Guanliang 467  
Chenevotot, Françoise 596  
Chenevotot-Quentin, Françoise 262  
Chew, Chin Rui 600  
Choquet, Christophe 648  
Conlan, Owen 136  
Cornelisz, Ilja 311  
Couland, Quentin 591  
Crossley, Scott A. 622  
Cukurova, Mutlu 291, 627
- Dascalu, Mihai 427, 482, 622  
Davis, Dan 122  
De Laet, Tinne 399  
de Lange, Peter 172  
de Morais, Felipe 231  
Delgado-Kloos, Carlos 355
- DEMMANS EPP, Carrie 216  
Demmans Epp, Carrie 497  
Di Mitri, Daniele 45  
Dillenbourg, Pierre 370  
Dimitrova, Vania 561, 570, 656  
Drachsler, Hendrik 45, 187, 297, 311, 579  
Drijvers, Paul 262  
Dueire Lins, Rafael 245
- Eradze, Maka 617  
Excell, Peter S. 551
- Farrell, Tracie 172  
Ferreira, Rafael 245  
Fessler, Angela 636  
Fuglik, Viktor 166, 575
- García-Sastre, Sara 617  
Gašević, Dragan 245, 385, 556  
Gayoso-Cabada, Joaquín 566  
George, Sébastien 591  
Ghaffar, Faisal 452  
Giannakos, Michail 326  
Göschlberger, Bernhard 172  
Grugeon-Allys, Brigitte 262, 596  
Guardia, Lourdes 627  
Guerra, Julio 524  
Gutu-Robu, Gabriel 482, 622
- Håklev, Stian 370  
Hamon, Ludovic 591  
Hauff, Claudia 101, 122  
Heeren, Bastiaan 262  
Henderikx, Maartje 3  
Henríquez, Valeria 340  
Hernández-Leo, Davinia 406, 605  
Hofman, Abe 531  
Hori, Masumi 587  
Horne, Joe 216  
Houben, Geert-Jan 122, 467  
Hsiao, I-Han 510, 644
- Ioannou, Andri 537

- Jalal, Ghita 59  
 Jansen, Renée S. 116  
 Janssen, Jeroen 116  
 Jaques, Patricia A. 231  
 Jeuring, Johan 262
- Kalz, Marco 3  
 Kane, Irene 216  
 Kester, Liesbeth 116  
 Kirschner, Paul A. 277  
 Kita, Toshihiro 587  
 Klamma, Ralf 172  
 Knobbout, Justian 88  
 Koops, Jesse 262  
 Kovanović, Vitomir 245  
 Kreijns, Karel 3, 31, 277  
 Kuzilek, Jakub 166, 575
- Lachand, Valentin 59  
 Laforcade, Pierre 151  
 Laghouaouta, Youness 151  
 Langie, Greet 399  
 Lau, Lydia 561  
 Lee, Amelia Jing Hua 600  
 Lee, Sunbok 74  
 Lee-Cultura, Serena 326  
 Leidig, Sebastian 583  
 Lenne, Dominique 631  
 Limbu, Bibeg 45  
 Lins, Rodrigo 245  
 Lofi, Christoph 101, 467  
 Lourdeaux, Domitile 631  
 Luengo, Vanda 596
- Maina, Marcelo 627  
 Maldonado-Mahauad, Jorge 16, 355  
 Mangaroska, Katerina 326  
 Martin, Bruno 596  
 Masala, Mihai 482  
 Mavrikis, Manolis 627  
 Mavroudi, Anna 640  
 McNamara, Danielle S. 427  
 Mdhaffar, Salima 648  
 Meinel, Christoph 202  
 Mesbah, Sepideh 467  
 Michel, Christine 59  
 Mitrovic, Antonija 561, 656  
 Miyahara, Hiroki 587  
 Miyashita, Kensuke 587  
 Moreno-Marcos, Pedro Manuel 355
- Muñoz-Merino, Pedro J. 355  
 Müter, Laurens 531
- Nasyrov, Rinat R. 551  
 Neto, Valter 245  
 Ng, Justin Choon Hwee 600  
 Nikolaedou, Elena 537  
 Nikolayeva, Iryna 596
- Olsen, Jennifer K. 412  
 Ono, Seishi 587
- Pammer-Schindler, Viktoria 636  
 Papasalouros, Andreas 660  
 Paraschiv, Ionut Cristian 427  
 Pardo, Abelardo 385  
 Paredes, Yancy Vance 510  
 Peirce, Neil 452  
 Pérez-Álvarez, Ronald 16  
 Pérez-Sanagustín, Mar 16, 355  
 Piau-Toffolon, Claudine 648  
 Pilet, Julia 596  
 Piotrkowicz, Alicja 570  
 Praharaj, Sambit 187  
 Prévít, Dominique 596  
 Pritchard, David E. 74
- Rahmani Hanzaki, Mahdi 497  
 Rajagopal, Kamakshi 31  
 Rebedea, Traian 482  
 Rensing, Christoph 609  
 Robal, Tarmo 101  
 Roberts, Trudie E. 570  
 Rodríguez-Medina, Jairo 617  
 Rodríguez-Triana, María Jesús 617  
 Rohloff, Tobias 202  
 Rolim, Vitor 245  
 Roller, Wolfgang 583  
 Ruipérez-Valiente, José A. 74  
 Rummel, Nikol 412  
 Ruseti, Stefan 482, 622
- Saint, John 385  
 Sakashita, Shiu 587  
 Scheffel, Maren 187, 297, 311, 556  
 Scheihing, Eliana 340  
 Schmitz, Marcel 297, 311  
 Schneider, Jan 45, 579  
 Schnitzer, Steffen 609  
 Scolieri, Britney B. 216

- Serlie, Alec 452  
 Sharma, Kshitij 326, 370, 412  
 Sierra, José-Luis 566  
 Silva, Marta 340  
 Sirbu, Maria-Dorinela 622  
 Sjöden, Björn 656  
 Skocilas, Jan 575  
 Slotta, Jim 370  
 Smeaton, Alan 510, 644  
 Socratous, Chrysanthos 537  
 Sosnovsky, Sergey 262, 531  
 Specht, Marcus 187  
 Staikopoulos, Athanasios 136  
 Stranger-Johannessen, Espen 613  
 Streicher, Alexander 583
- Tabard, Aurélien 59  
 Tacoma, Sietske 262  
 Tay, Sook Muay 600  
 te Vrugt, Jürgen 652  
 Tham, Sarah Zhuling 600  
 Trætteberg, Halvard 326  
 Trausan-Matu, Stefan 427, 482, 622  
 Triglianios, Vasileios 122  
 Tsai, Yi-Shan 556
- Vaclavek, Jonas 166, 575  
 Valkenier, Marc 531  
 Valle Torre, Manuel 467  
 van der Stappen, Esther 88  
 van Halem, Nicolette 311  
 van Klaveren, Chris 311  
 van Leeuwen, Anouschka 116  
 van Limbeek, Evelien 297, 311  
 Van Soom, Carolien 399  
 van Walree, Ferdinand 262  
 van Wijk, Jorn 262  
 Verbert, Katrien 399  
 Viberg, Olga 640
- Weidlich, Joshua 31  
 Weinberger, Armin 262  
 Wertner, Alfred 636  
 Weßeler, Peter 652
- Yamaji, Kazutuna 587  
 Yessad, Amel 596  
 Yousuf, Bilal 136
- Zdrahal, Zdenek 166, 575  
 Zhao, Yue 101