

LNCS 13891

Claude Frasson  
Phivos Mylonas  
Christos Troussas (Eds.)

# Augmented Intelligence and Intelligent Tutoring Systems

19th International Conference, ITS 2023  
Corfu, Greece, June 2–5, 2023  
Proceedings



Springer

## Founding Editors

Gerhard Goos

Juris Hartmanis

## Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Claude Frasson · Phivos Mylonas ·  
Christos Troussas  
Editors

# Augmented Intelligence and Intelligent Tutoring Systems

19th International Conference, ITS 2023  
Corfu, Greece, June 2–5, 2023  
Proceedings



Springer

*Editors*

Claude Frasson   
University of Montreal  
Montreal, Canada

Phivos Mylonas   
University of West Attica  
Athens, Greece

Christos Troussas   
University of West Attica  
Athens, Greece

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-32882-4

ISBN 978-3-031-32883-1 (eBook)

<https://doi.org/10.1007/978-3-031-32883-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

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

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

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

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

# Preface

The 19th International Conference on Intelligent Tutoring Systems (ITS 2023) was held in Corfu, Greece, from June 2 to June 5, 2023.

This year, ITS evolved to a new concept of Artificial Intelligence which can be found in various disciplines and can serve to promote human education and well-being. This new concept is named: AUGMENTED INTELLIGENCE. Augmented Intelligence concerns any system, technique, architecture, method, or tools which can be used to enhance HUMAN intelligence, cognitive capacities, memory, or learning. Instead of trying to reproduce human behavior or recognize human traits using artificial intelligence techniques or learning analytics, Augmented Intelligence provides means to increase human cognitive potential.

The conference supported the following tracks in which the concepts, mentioned above, contribute to the increase of intelligence:

**T1: Augmented Intelligence in Tutoring Systems.** The goal of this track was to show how new techniques inspired by artificial intelligence and new methods in education can improve learning and teaching, and augment the capacity for knowledge acquisition. The track included the following topics: augmented learning strategies, distance education, learning analytics for tutoring systems, deep learning and machine learning for tutoring systems, online and distance learning, augmented learner model, emotion recognition, human machine interaction, case-based reasoning, cognitive modeling, open learning, authoring systems, cultural learning, and adaptive learning.

**T2: Augmented Intelligence in Healthcare Informatics.** The goal of this track was to show the progress of AI tools for increasing the propagation of healthcare techniques and their efficiency. Informatics provides means to improve the prediction, analysis, and treatment of disease and control of patients over their own care. The track included the following topics: AI and telemedicine, medical image processing, and virtual systems for healthcare

**T3: Augmented Intelligence in Games, Serious Games, and Virtual Reality.** The goal of this track was to show the progress of interactive games using augmented intelligent techniques. Intelligent games can adapt to the characteristics of the player and can be used to enhance learning, skills, memory, cognitive capacities, and strategic decisions. They can be used in various applications (education, healthcare, group management, decision support systems, and industrial control).

**T4: Neural Networks and Data Mining** The track included the following topics: supervised machine learning, genetic algorithms, Markovian regulation, smart sensor networks, determinate regulation, games and strategies, fuzzy systems, web information processing, applications of data mining in social sciences, data-driven reasoning, deep learning and statistical methods for data mining, big data mining, algorithms for data mining, ethical data analytics, data mining for recommendation, and uncertain reasoning.

**T5: Augmented Intelligence and the Metaverse** The track included the following topics: technology and creativity around the Metaverse, gaming and interactivity, mixed

reality and the virtual world, social and digital identity, extended reality, digital art, social communication, applications of the Metaverse in Health, global Metaverse, knowledge increase with the Metaverse.

**T6: Security, Privacy, and Ethics in Augmented Intelligence** The track included the following topics: commercial security, data privacy and security, web security, applied cryptography, authentication, identity management and biometrics, electronic payments, culture of ethics, business, and human rights, diversity and inclusion in teaching and learning, environmental ethics, machine learning and security, cloud computing and data outsourcing security, mobile payments, security in games, security of peer-to-peer networks, security metrics, sustainability, language-based security, security and privacy for the Internet of Things, and socio-technical security.

**T7: Applied Natural Language Processing** The track included the following topics: language modeling, domain ontologies, computational linguistics, cognitive semantics, text mining, translation, question answering, dialog systems, information retrieval, speech recognition and synthesis, discourse, machine translation, lexical semantics, and linguistic resources.

The call for scientific papers solicited works presenting substantive new research results in using advanced computer technologies and interdisciplinary research to enable, support, and enhance human learning.

The international Program Committee consisted of 57 leading members (27 senior and 30 regular) of the Intelligent Tutoring Systems community, assisted by additional external reviewers.

Research papers came from 26 countries and were reviewed by three reviewers through a double-blind process. ITS 2023 retained the strict and high standards in the review process that it has established during the previous years, and which have rendered it a top-flight, rather selective, and high-quality conference. This year, from 84 submissions, 40 papers were accepted as full, and 20 were accepted as short papers. We believe that the selected full papers describe some very significant research and the short papers some very interesting new ideas.

The management of the review process and the preparation of the proceedings was handled through EasyChair.

Finally, we had two outstanding invited speakers in the plenary sessions: Seiji Isotani (University of Sao Paolo, Brazil) a renowned figure in the field of Gamification, Learning Technologies and STEM Education, and Panos Kourouthanassis (Ionian University, Greece) who specializes in Augmented Reality and Pervasive Computing. Both are leaders in different specialized areas of the ITS field.

In addition to the contributors mentioned above, we would like to thank all the authors, the members of the Program Committees of all tracks, the external reviewers, and the Steering Committee members.

We would also like to acknowledge the Institute of Intelligent Systems (IIS), which acted as the conference organizer, particularly Kitty Panourgia and her excellent team, and the Ionian University, being the hosting institution.

April 2023

Claude Frasson  
Phivos Mylonas  
Christos Troussas

# **Organization**

## **Conference Committee**

### **Program Chairs**

Claude Frasson

University of Montreal, Canada

Phivos Mylonas

Ionian University, Greece

Christos Troussas

University of West Attica, Greece

### **Organization Chair**

Kitty Panourgia

Neoanalysis, Greece

The Conference was held under the auspices of the Institute of Intelligent Systems.



## **Program Committee**

### **Program Chairs**

Claude Frasson

University of Montreal, Canada

Phivos Mylonas

Ionian University, Greece

Christos Troussas

University of West Attica, Greece

### **Senior Program Committee**

Roger Azevedo

University of Central Florida, USA

Benedict du Boulay

University of Sussex, UK

Bert Bredeweg	University of Amsterdam, The Netherlands
Stefano A. Cerri	University of Montpellier, France
Maiga Chang	Athabasca University, Canada
Mihai Dascalu	Politehnica University of Bucharest, Romania
Michel Desmarais	Polytechnique Montréal, Canada
Spyros Doukakis	Ionian University, Greece
Claude Frasson	University of Montreal, Canada
Yugo Hayashi	Ritsumeikan University, Japan
Kinshuk	University of North Texas, USA
Amruth Kumar	Ramapo College of New Jersey, USA
Vive Kumar	Athabasca University, Canada
Susanne Lajoie	McGill University, Canada
Noboru Matsuda	North Carolina State University, USA
Alessandro Micarelli	Roma Tre University, Italy
Phivos Mylonas	Ionian University, Greece
Riichiro Mizoguchi	Japan Institute of Science and Technology, Japan
Roger Nkambou	UQAM University, Canada
Elvira Popescu	University of Craiova, Romania
Demetrios Sampson	University of Piraeus, Greece
Filippo Sciarrone	Universitas Mercatorum, Italy
Stefan Trausan-Matu	Politehnica University of Bucharest, Romania
Marco Temperini	Sapienza University of Rome, Italy
Christos Troussas	University of West Attica, Greece
Julita Vassileva	University of Saskatchewan, Canada
Panagiotis Vlamos	Ionian University, Greece

## Program Committee

Mohammed Abdel Razek	Al-Azhar University, Egypt
Galia Angelova	Bulgarian Academy of Sciences, Bulgaria
Renu Balyan	SUNY Old Westbury, USA
Hamdi Ben Abdessalem	University of Montreal, Canada
Emmanuel Blanchard	Edu Interactive, Canada
Evandro Costa	Federal University of Alagoas, Brazil
Chih-Yueh Chou	Yuan Ze University, Taiwan
Diego Derméval	Federal University of Alagoas, Brazil
Cyrille Desmoulins	University Joseph Fourier, France
Davide Fossati	Emory University, USA
Reva Freedman	Northern Illinois University, USA
Ashok Goel	School of Interactive Computing, Georgia, USA
Ella Haig	University of Portsmouth, UK

Seiji Isotani	University of Sao Paulo, Brazil
Mizue Kayama	Shinshu University, Japan
Charalampos Karagiannidis	University of Thessaly, Greece
Fuhua Oscar Lin	Athabasca University, Canada
Mirko Marras	University of Cagliari, Italy
Silvia Margarita Baldiris Navarro	University of Cartagena, Colombia
Tassos Mikropoulos	University of Ioannina, Greece
Wolfgang Mueller	University of Education Weingarten, Germany
Kuo-Liang Ou	National Hsin-Chu University of Education, Taiwan
Valéry Psyché	Université TELUQ, Canada
Ricardo Queirós	Agostinho Neto University, Portugal
Traian Rebedea	University Politehnica of Bucharest, Romania
Olga C. Santos	National Distance Education University (UNED), Spain
Lei Shi	Durham University, UK
Thepchai Supnithi	NECTEC, Thailand
Radu Vasiu	Politehnica University of Timisoara, Romania
Riina Vuorikari	Joint Research Centre - European Commission, Belgium

## Organization Committee

### Organization Chair

Kitty Panourgia	General Coordination / Proceedings / Program
-----------------	--

### Members

Aggelos Amarandos	Registration
Alexia Kakourou	Coordination on Site
Eliana Vassiliou	Conference Publicity / Website Management
Isaak Tselepis	Website Architect

## Steering Committee

### Chair

Claude Frasson

University of Montreal, Canada

### Members

Stefano A. Cerri	University of Montpellier and CNRS, France
Maiga Chang	Athabasca University, Canada
Isabel Fernandez-Castro	University of the Basque Country, Spain
Yugo Hayashi	Ritsumeikan University, Japan
Alan Lesgold	University of Pittsburgh, USA
Amruth Kumar	Ramapo College of New Jersey, USA
Alessandro Micarelli	Roma Tre University, Italy
Roger Nkambou	Université du Québec à Montréal, Canada
Giorgos Papadourakis	Hellenic Mediterranean University, Greece
Elliot Soloway	University of Michigan, USA
John Stamper	Carnegie Mellon University, USA
Daniel Suthers	University of Hawaii, USA
Stefan Trausen-Matu	University Politehnica of Bucharest, Romania
Christos Troussas	University of West Attica, Greece
Beverly Woolf	University of Massachusetts, USA

## Advisory Committee

### Members

Luigia Carlucci Aiello
Maria Grigoriadou
Demetrios G. Sampson

University of Rome, Italy
University of Athens, Greece
University of Piraeus, Greece

## **Invited Talks**

# Personalization of Gamification in Education: Where Augmented Intelligence Meets Playfulness

Seiji Isotani<sup>1,2</sup> 

<sup>1</sup> Graduate School of Education, Harvard University,  
Cambridge MA 02138, USA

<sup>2</sup> University of Sao Paulo, Sao Carlos SP 13566-590 Brazil  
[seiji\\_isotani@gse.harvard.edu](mailto:seiji_isotani@gse.harvard.edu)

Gamification, i.e., the use of game design elements in non-game contexts, has been successfully used to design better (virtual) learning environments aiming to increase positive experiences such as motivation, engagement, satisfaction, meaningful social interactions, self-efficacy, and flow. Recent advances highlight the importance of tailoring gamification design in education to maximize its benefits to students and reduce its potential negative effects. Nevertheless, the process of adapting and personalizing gamification for educational purposes requires knowledge about gamification design and how these elements influence students' flow state. This influence may differ according to the context (learning materials, pedagogical approaches, feedback, etc) and the individuals' characteristics (culture, age, user type, etc). Thus, in the past 10+ years, my research group has studied the factors that influence gamification design and use artificial intelligence techniques to augment human capabilities to better personalize gamification in educational contexts, particularly for STEM (Science, Technology, Engineering, and Mathematics) learning. To discuss the personalization of gamification in education, we will first discuss some of the positive and negative effects of gamification based on recent literature reviews. Then, a series of experiments carried out by my group will be presented to show how different contextual variables (e.g., culture, educational content, time on task, students' interests, etc.) affect the potential benefits of gamification in education, particularly in the context of Computer Science Education. Finally, using these findings, we will present a taxonomy and data-driven framework to assist teachers in making decisions to design personalized gamified experiences.

**Keywords:** Artificial Intelligence in Education · tailored gamification · educational technology.

# Designing Immersive Information Systems in the ERA of Artificial Intelligence

Panos Kourouthanassis

Department of Informatics, Ionian University, Greece

**Abstract.** Immersive information systems may shape end users' perceptions of reality. Immersive information systems employ such information and communication technologies as augmented reality (AR), virtual reality (VR), and gamification [2]. Scholars have identified the application potential of AR/ VR in several fields that range from medical applications [4], education [5], manufacturing [1], and retail [6]. The keynote speech will discuss the design challenges of immersive information systems from both engineering/technological and user adoption examination stances. Furthermore, the keynote speech will discuss how extant information systems design science research perspectives (e.g. information systems design theories) may be employed for capturing the design elements of immersive information systems by emphasizing on the design of AR/ mobile-AR applications [3] and, in general, AR/ VR augmented intelligence immersive systems.

## References

1. Baroroh, D.K., Chu, C.H., Wang, L. Systematic literature review on augmented reality in smart manufacturing: collaboration between human and computational intelligence. *J. Manuf. Syst.* **61**, 696–711 (2021)
2. Cavusoglu, H., Dennis, A.R., Parsons, J.: Editorial of special issue: immersive Systems, *J. Manag. Inf. Syst.* **36**(3), 680–682. (2019)
3. Kourouthanassis, P.E., Boletsis, C., Lekakos, G.: Demystifying the design of mobile augmented reality applications. *Multimedia Tools Appl.* **74**, 1045–1066 (2015)
4. Lungu, A.J., Swinkels, S., Claesen, L., Tu, P., Egger, J., Chen, X.: A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery. *Expert Rev. Med. Devices.* **18**(1), 47–62 (2021)
5. Videnovik, M., Trajkovik, V., Kiønig, L.V., Vold, T.: Increasing quality of learning experience using augmented reality educational games. *Multimedia Tools Appl.* **79**, (23861–23885) (2020)

6. Zimmermann, R., et al.: Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence. *J. Res. Interact. Mark.* **17**(2), 273–298 (2023)

# Contents

## Augmented Intelligence in Tutoring Systems

Language Proficiency Enhanced Knowledge Tracing .....	3
<i>Heeseok Jung, Jaesang Yoo, Yohaan Yoon, and Yeonju Jang</i>	
Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory .....	16
<i>Matteo Orsoni, Alexander Pögelt, Nghia Duong-Trung, Mariagrazia Benassi, Milos Kravcik, and Martin Grüttmüller</i>	
Automatic Execution of the Ontological Piloting Procedures .....	29
<i>Marc-Antoine Courtemanche, Ange Tato, and Roger Nkambou</i>	
Method and Tools to Supporting Math Learning in Inclusive Education of Blind Students .....	42
<i>Michał Maćkowski, Mateusz Kawulok, Piotr Brzoza, and Dominik Spinczyk</i>	
Understanding the Usages and Effects of a Recommendation System in a Non-formal Learning Context .....	54
<i>Anaëlle Badier, Mathieu Lefort, and Marie Lefevre</i>	
Detecting Mental Fatigue in Intelligent Tutoring Systems .....	66
<i>Alyssa Hajj Assaf, Hamdi Ben Abdessalem, and Claude Frasson</i>	
An Approach for Detecting Gaming the System Behavior in Programming Problem-Solving .....	75
<i>Hemilis Joyse Barbosa Rocha, Patricia Cabral de Azevedo Restelli Tedesco, Evandro de Barros Costa, and Julios Suruagi Rocha</i>	
An Approach to Generating Adaptive Feedback for Online Formative Assessment .....	88
<i>Fuhua Lin and Supun De Silva</i>	
How to Generate Early and Accurate Alerts of At-Risk of Failure Learners? ...	100
<i>Amal Ben Soussia, Azim Roussanaly, and Anne Boyer</i>	
Detecting Interlingual Errors: The Case of Prepositions .....	112
<i>Natawut Monaikul and Barbara Di Eugenio</i>	

Automated Extraction of Domain Models from Textbook Indexes for Developing Intelligent Tutoring Systems .....	124
<i>Rabin Banjade, Priti Oli, and Vasile Rus</i>	
Teaching Cardiovascular Physiology with a Minigame-Based ITS .....	137
<i>Reva Freedman, Lucas Edwards, Ian Sullivan, Annalise Brockmann,     Dean LaBarbera, and Virginia Naples</i>	
Expert-Centered Design Recommendations to Promote Engagement in Authoring Activities of Adaptive Learning Technologies .....	147
<i>Emmanuel G. Blanchard</i>	
Towards a Chatbot-Based Learning Object Recommendation: A Comparative Experiment .....	159
<i>Wijdane Kaiss, Khalifa Mansouri, and Franck Poirier</i>	
Development of a Conversational Agent for Tutoring Nursing Students to Interact with Patients .....	171
<i>Tomasz Sosnowski, Moh'd Abuazizeh, Thomas Kirste,     and Kristina Yordanova</i>	
Integrating an Ontological Reference Model of Piloting Procedures in ACT-R Cognitive Architecture to Simulate Piloting Tasks .....	183
<i>Guy Carlos Tamkodjou Tchio, Marc-Antoine Courtemanche,     Ange Adrienne Nyamen Tato, Roger Nkambou, and Valéry Psyché</i>	
Generating Pedagogical Questions to Help Students Learn .....	195
<i>Oleg Sychev and Marat Gumerov</i>	
Attention Assessment of Aircraft Pilots Using Eye Tracking .....	209
<i>Maryam Ghaderi, Amin Bonyad Khalaj, Hamdi Ben Abdessalem,     and Claude Frasson</i>	
KEMMRL: Knowledge Extraction Model for Morphologically Rich Languages .....	220
<i>Daniel Vasić, Branko Žitko, Ani Grubišić, and Angelina Gašpar</i>	
Synthesizing Didactic Explanatory Texts in Intelligent Tutoring Systems Based on the Information in Cognitive Maps .....	233
<i>Viktor Uglev and Oleg Sychev</i>	
Using the ITS Components in Improving the Q-Learning Policy for Instructional Sequencing .....	247
<i>Amel Yessad</i>	

Stargazing Live! Inspiring with Real Data in a Mobile Planetarium and Learning Through Conceptual Modelling .....	257
<i>Bert Bredeweg, Marco Kragten, Joanna Holt, Dennis Vaendel, Joris Hanse, and Steven Bloemen</i>	
A GPT-Based Vocabulary Tutor .....	270
<i>George Zografos and Lefteris Moussiades</i>	
ETHOSCHOOL: An Artificial Moral Agent Model for Collaborative Learning .....	281
<i>Paulo Roberto Córdova, Iderli Pereira de Souza Filho, Gilleannes Thorwald Araujo Guedes, and Rosa Maria Vicari</i>	
Toward a Smart Tool for Supporting Programming Lab Work .....	290
<i>Anis Bey and Ronan Champagnat</i>	
Towards Embodied Wearable Intelligent Tutoring Systems .....	298
<i>Injila Rasul, Francisco Castro, and Ivon Arroyo</i>	
iQUIZ!: A Collaborative Online Learning System that Promotes Growth Mindset Using Persuasive Feedback .....	307
<i>Mehnuma Tabassum Omar, Nafisul Kiron, and Julita Vassileva</i>	
Helping Teachers to Analyze Big Sets of Concept Maps .....	315
<i>Michele La Barbera, Filippo Sciarrone, and Marco Temperini</i>	
Learning Engagement and Peer Learning in MOOC: A Selective Systematic Review .....	324
<i>Fatma Miladi, Daniel Lemire, and Valéry Psyché</i>	
Personalized Study Guide: A Moodle Plug-in Generating Personal Learning Path for Students .....	333
<i>Ted Krahn, Rita Kuo, and Maiga Chang</i>	
Learning by Building Chatbot: A System Usability Study and Teachers' Views About the Educational Uses of Chatbots .....	342
<i>Stavros A. Nikou and Maiga Chang</i>	
Towards Integrating Learnersourcing, Microlearning and Gamification in Moodle .....	352
<i>Mariana Madalina Nastase and Elvira Popescu</i>	
Towards a Social Learning Analysis Using Learning Management System and Learning Experience to Predict Learners' Success .....	364
<i>Hiba Gharbaoui, Khalifa Mansouri, and Franck Poirier</i>	

Explanatory Didactic Dialogue in the Intelligent Tutoring Systems Based on the Cross-Cutting Approach .....	371
<i>Viktor Uglev</i>	
Estimation of Piloting Attention Level Based on the Correlation of Pupil Dilation and EEG .....	381
<i>Maryam Ghaderi, Hamdi Ben Abdessalem, Maxime Antoine, and Claude Frasson</i>	
Functional Programming of Intelligent Systems .....	391
<i>Vladymyr Meitus and Clara Simon de Blas</i>	
Model-Based Support for Collaborative Concept Mapping in Open-ended Domains .....	404
<i>Junya Morita, Masaji Kano, Shigen Shimojo, Yoshimasa Ohmoto, and Yugo Hayashi</i>	
<b>Augmented Intelligence in Healthcare Informatics</b>	
CPR Emergency Assistance Through Mixed Reality Communication .....	415
<i>Manuel Rebol, Alexander Steinmaurer, Florian Gamillscheg, Krzysztof Pietroszek, Christian Gütl, Claudia Ranniger, Colton Hood, Adam Rutenberg, and Neal Sikka</i>	
The Relative Importance of Cognitive and Behavioral Engagement to Task Performance in Self-regulated Learning with an Intelligent Tutoring System ...	430
<i>Xiaoshan Huang, Shan Li, and Susanne P. Lajoie</i>	
Emotional Impact of Cognitive Priming on Alzheimer's Disease .....	442
<i>Hamdi Ben Abdessalem and Claude Frasson</i>	
Combining XR and AI for Integrating the Best Pedagogical Approach to Providing Feedback in Surgical Medical Distance Education .....	452
<i>Yves S. Nkulu-Ily</i>	
Adversarial Learning for Improved Patient Representations .....	467
<i>Bharath Shankar and Carol Anne Hargreaves</i>	
<b>Augmented Intelligence in Games, Serious Games and Virtual Reality</b>	
User-Defined Hand Gesture Interface to Improve User Experience of Learning American Sign Language .....	479
<i>Jindi Wang, Ioannis Ivrissimtzis, Zhaoxing Li, Yunzhan Zhou, and Lei Shi</i>	

- Distraction Detection and Monitoring Using Eye Tracking in Virtual Reality ... 491**  
*Mahdi Zarour, Hamdi Ben Abdessalem, and Claude Frasson*

### **Neural Networks and Data Mining**

- Using Feature Interaction for Mining Learners' Hidden Information in MOOC Dropout Prediction ..... 507**  
*Tingfeng Pan, Guang Feng, Xin Liu, and Wenyan Wu*

- On an Integrated Assessment for the Students Within an Academic Consortium ..... 518**  
*Popescu Doru-Anastasiu, Cristea Daniela-Maria, and Bold Nicolae*

- An Approach to Automatic Flight Deviation Detection ..... 530**  
*Massimo Pietracupa, Hamdi Ben Abdessalem, and Claude Frasson*

- Automatic Learning of Piloting Behavior from Flight Data ..... 541**  
*Ange Tato, Roger Nkambou, and Gabrielle Tato*

- Towards Student Behaviour Simulation: A Decision Transformer Based Approach ..... 553**  
*Zhaoxing Li, Lei Shi, Yunzhan Zhou, and Jindi Wang*

### **Augmented Intelligence and Metaverse**

- Mixed Reality Agents as Language Learning Tutors ..... 565**  
*Benedikt Hensen and Danylo Bekhter*

- Metaverse and Virtual Environment to Improve Attention Deficit Hyperactivity Disorder (ADHD) Students' Learning ..... 576**  
*Azza Mohamed, Mai Zohiar, and Ibrahim Ismail*

### **Security, Privacy and Ethics in Augmented Intelligence**

- Culture of Ethics in Adopting Learning Analytics ..... 591**  
*Dimitrios Tzimas and Stavros Demetriadis*

- Promoting Ethical Uses in Artificial Intelligence Applied to Education ..... 604**  
*Ana Matias and Iraide Zipitria*

**Applied Natural Language Processing**

Improving Collaboration via Automated Intelligent Nudges .....	619
<i>Solomon Ubani and Rodney Nielsen</i>	
Preliminary Performance Assessment on Ask4Summary’s Reading Methods for Summary Generation .....	630
<i>Rita Kuo, Maria F. Iriarte, Di Zou, and Maiga Chang</i>	
Joint Topic Model with Selected Side Information for Inter-University Syllabus Analysis Focusing on the Variety of Instructional Approaches .....	638
<i>Kentaro Shiga and Naoyuki Morimoto</i>	
Plug & Play with Deep Neural Networks: Classifying Posts that Need Urgent Intervention in MOOCs .....	651
<i>Laila Alrajhi and Alexandra I. Cristea</i>	
Personalized Feedback Enhanced by Natural Language Processing in Intelligent Tutoring Systems .....	667
<i>Christos Troussas, Christos Papakostas, Akrivi Krouskas, Phivos Mylonas, and Cleo Sgouropoulou</i>	
SC-QUES: A Sentence Completion Question Dataset for English as a Second Language Learners .....	678
<i>Qiongqiong Liu, Yaying Huang, Zitao Liu, Shuyan Huang, Jiahao Chen, Xiangyu Zhao, Guimin Lin, Yuyu Zhou, and Weiqi Luo</i>	
Conversational Agents and Language Models that Learn from Human Dialogues to Support Design Thinking .....	691
<i>Stavros Demetriadis and Yannis Dimitriadis</i>	
<b>Author Index .....</b>	701

# **Augmented Intelligence in Tutoring Systems**



# Language Proficiency Enhanced Knowledge Tracing

Heeseok Jung<sup>1</sup>, Jaesang Yoo<sup>1</sup>, Yohaan Yoon<sup>1</sup>, and Yeonju Jang<sup>2(✉)</sup>

<sup>1</sup> Classting AI Research, Seoul, Republic of Korea

{poco2889, jsyoo, straightfor}@classting.com

<sup>2</sup> Creative Informatics and Computing Institute, Korea University, Seoul, Republic of Korea  
spring0425@korea.ac.kr

**Abstract.** With an increasing interest in personalized learning, active research is being conducted on knowledge tracing to predict the learner's knowledge state. Recently, studies have attempted to improve the performance of the knowledge tracing model by incorporating various types of side information. We propose a knowledge tracing method that utilizes the learner's language proficiency as side information. Language proficiency is a key component of comprehending a question's text and is known to be closely related to students' academic performance. In this study, language proficiency was defined with Elo rating score and time window features, and was used in the knowledge tracing task. The dataset used in this study contains 54,470 students and 7,619,040 interactions, which were collected from a real-world online-learning platform. We conducted a correlation analysis to determine whether the language proficiency information of students was related to their ability to solve math word problems. In addition, we examined the effect of incorporating the language proficiency information on the knowledge tracing models using various baseline models. The analysis revealed a high correlation between the length of word problems and students' language proficiency. Furthermore, in experiments with various baseline models, utilizing the language proficiency information improved the knowledge tracing model's performance. Finally, when language proficiency information was incorporated, the cold start problem of the knowledge tracing model was mitigated. The findings of this study can be used as a supplement for educational instruction.

**Keywords:** Knowledge Tracing · Student Modeling · Learning Analytics · Language Proficiency

## 1 Introduction

As interest in personalized online learning grows, the research in the field of knowledge tracing (KT) to model a learner's knowledge state also increases. Many studies have recently been conducted to improve the performance of the KT model by using various side information, such as response time [31, 33], the number of attempts [10, 43], question text [21, 36], and relationship between concepts [9, 23].

On the other hand, previous studies have found that the academic achievement in one subject is closely related to achievement in other subjects. For example, math grades are significantly correlated with science grades [16, 39], while English grades have a significant impact on math and science grades [3, 4]. As a result, if learning data from multiple subjects in which students participated can be collected, problem-solving information from one subject can be used to predict achievement in other subjects.

In particular, if we collect problem-solving performance data for students' first language subjects (e.g., Korean, English, French, etc.), we can estimate their language proficiency (LP) information. LP is an important factor when learners acquire domain knowledge, and it has been shown to have a significant impact on academic performance in a variety of subjects [6, 15, 30, 35]. Moreover, LP is an important factor in the learner's problem solving. This is owing to the fact that learners must be able to read and comprehend the problem to complete it correctly [6]. For example, in mathematics, students with low LP are known to struggle with long word problems [37]. Therefore, LP information extracted from students' problem-solving data of language subjects is likely to be useful for the KT task to predict their future performance in other subjects.

However, despite the fact that LP is an important factor in predicting students' academic performance, to our knowledge, no study in the KT field has used this information. In addition, studies that used academic achievement data in one subject to predict the student's knowledge state in another subject have not been sufficiently conducted.

In this study, we propose a KT method using LP information extracted from Korean problem solving data from students. First, we analyzed whether LP information was related to students' math word problem performance. In addition, through experiments with real-world datasets, we predicted the knowledge state using students' LP information and investigated whether using this information was effective in improving the performance of the KT model. Furthermore, we investigated whether using LP information can help mitigate the cold start problem.

As a result, the following research questions (RQs) were posed in this study, in an attempt to find answers.

**RQ1:** Is there a correlation between students' LP and their word problem performance?

**RQ2:** Is it possible to improve the performance of the KT models by incorporating LP information?

**RQ3:** Does utilizing LP information alleviate the cold start issue of the KT model?

The remainder of this paper is organized as follows. In Sect. 2, related works on deep learning-based knowledge tracing, students' LP and academic performance, and the Elo rating system are presented. Section 3 describes the methodology used in the study. Sections 4 and 5 present the results and a discussion, respectively. Finally, Sect. 6 presents the conclusion of this study.

## 2 Related Works

### 2.1 Deep Learning-Based KT

KT aims to estimate a learner's knowledge state and predict future performance based on past learning histories. Typically, the learning history is represented by

$X = \{(e_1, a_1), (e_2, a_2), \dots, (e_t, a_t)\}$ . Here,  $e_t$  denotes the exercise solved at time  $t$ , and  $a_t$  denotes the correctness of  $e_t$ .

As online education has grown in popularity, it has become possible to collect large amounts of learning activity data, and consequently, deep learning (DL)-based KT models are performing well. DL-based KT models include Recurrent Neural Network (RNN)-based models [29], Memory Augmented Neural Network (MANN)-based models [1, 17, 42], Attention Mechanism-based models [13, 19, 27, 33], and Graph Neural Network (GNN)-based models [26, 34, 41].

Various studies have attempted to estimate the knowledge state more precisely by utilizing side information in addition to exercise and response data. According to [24], the side information used in the KT models can be categorized as question side information, student side information, and knowledge component (KC) side information. Question side information includes question text content [21, 36] and question difficulty level [13, 23, 43], and so on. Student side information includes the number of attempts [10, 43] and students' prior knowledge [32], and so forth. As KC side information, relations between KCs, or questions [23] are used. In this study, the LP—which is student side information—was used for the KT task.

## 2.2 Students' LP and Academic Performance

When we learn something, we obtain a significant amount of information from text. Therefore, the LP is known to have a strong relationship with students' academic performance. In other words, students with low LP inevitably learn little from the text [6]. Previous studies have found that LP influences academic performance in various subjects, including mathematics, science, and social sciences [6, 15, 30, 35].

Especially in the case of mathematics, the ability to solve word problems is highly related to LP [37]. Word-problem solving differs from other types of mathematical competence; students are required to interpret text that explains a problem situation and deduce the number of sentences that represent that situation [12]. In particular, if LP is insufficient, calculation problems can be solved using key concept understanding, but not word problems.

Consequently, there is a possibility that LP plays an important role in modeling a learner's knowledge state. Therefore, in this study, LP was used as side information in the KT task to validate its effect.

## 2.3 Elo Rating System

The Elo Rating System (ERS) [11] was used to estimate the relative ability of a chess player. ERS is a dynamic paired comparison model that is mathematically closely related to the one parameter logistic (1PL) IRT model [7]. Players are given a provisional ability score  $\theta$  at first, which is incrementally updated according to the weighted difference in match outcome  $S$  and the expected match outcome  $E(S)$  (Eq. 1). Here,  $K$  is a constant specifying the sensitivity of the estimate in the last attempt. The expected match outcome  $E(S)$  is calculated based on the difference between the ability estimates of both players  $j$  and  $k$  preceding the match, and expresses the probability of winning as in Eq. 2.

$$\hat{\theta}_j = \theta_j + K(S_j - E(S_j)), \hat{\theta}_k = \theta_k + K(S_k - E(S_k)) \quad (1)$$

$$E(S_j) = \frac{1}{1 + 10^{\frac{(\theta_k - \theta_j)}{400}}}, E(S_k) = \frac{1}{1 + 10^{\frac{(\theta_j - \theta_k)}{400}}} \quad (2)$$

If we interpret a student's answer to a question as a match between the student and question, the ERS can be used to estimate learners' proficiency and question difficulty. As the rating score is updated by a heuristic rule, this is simple, robust, and effective for adaptive educational systems [28].

### 3 Methodology

#### 3.1 Language Proficiency-Enhanced Knowledge Tracing

The Elo rating score, calculated from the Korean problem-solving data, was used to estimate the students' LP. A time window feature was used to reflect the time at which the Elo rating score was updated, based on the method of Lindsey et al. [20]. The reason for considering the timing information of the Elo rating score update is that students' LP may change over time [5]. Because the students' LP in this study was measured using data from Korean problem solving, if the student had not recently participated in Korean problem solving, the student's Elo rating score would be fixed for a while. Therefore, by including a time window feature that contains information on the student's Korean problem-solving by period, the possibility of a change in the student's LP can be considered. The time window included information on the number of times the student attempted Korean problems and the number of correct answers by month, week, day, and hour. In addition, sentence length information from mathematical problems was used. Problem length was defined as the word count of each problem.

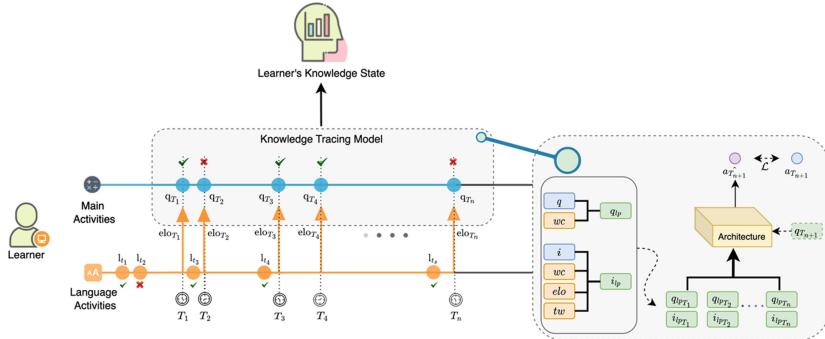
Figure 1 shows the structure of the LP-enhanced KT model proposed in this study.  $q_{T_1}, q_{T_2}, \dots, q_{T_n}$  indicate the questions of main activities, the math problem-solving event, whereas  $l_{t_1}, l_{t_2}, \dots, l_{t_s}$  indicate the questions of language activities, which is the Korean problem-solving event.  $T_1, T_2, \dots, T_n$  represent the timestamps of the main activities and  $t_1, t_2, \dots, t_s$  represent the timestamps of the language activities. The Elo rating score calculated from the results of language activities up to  $T_n$  is used for timestamp  $T_n$ .

Compatibility with existing models was considered when incorporating the LP information into the model. Because most KT models utilize question embedding and/or interaction embedding, LP information was incorporated into the model as in Eq. 3.

$$q_{lp} = \mathbf{W}_1^T[q \| wc], i_{lp} = \mathbf{W}_2^T[i \| wc \| elo \| tw] \quad (3)$$

Here,  $q$  refers to the question embedding. The interaction embedding of a tuple  $(q, a)$ , where  $q$  denotes the question and  $a$  denotes the response (or correctness), is denoted by  $i$ .  $\|$  denotes the concatenation.

Elo rating score, word count, and each features in time window are continuous values. Like [33], we used trainable matrices for each LP related features and get embedding vectors  $elo$ ,  $wc$ ,  $tw$ . The trainable parameters  $\mathbf{W}_1$  and  $\mathbf{W}_2$  project the concatenated vector to the size of the original question and interaction embedding, respectively.



**Fig. 1.** The illustration of language proficiency enhanced knowledge tracing.

Final prediction and loss function are as follows:

$$a_{\hat{T}_{n+1}} = \sigma(\text{Architecture}(q_{lpT_{1,n}}, i_{lpT_{1,n}})) \quad (4)$$

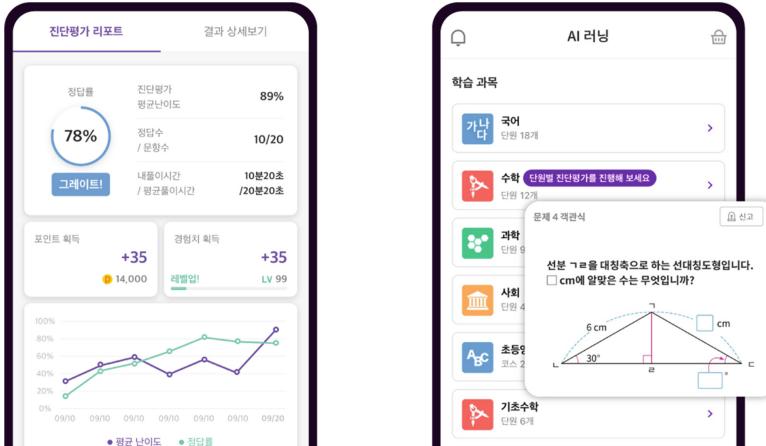
$$\mathcal{L} = BCE(a_{\hat{T}_{n+1}}, a_{T_{n+1}}) \quad (5)$$

### 3.2 Dataset

Data from Classting AI Learning<sup>1</sup> were used in this study. Classting AI Learning is an online learning platform for K-12 that provides contents in Korean, English, mathematics, social studies, and science. Classting AI Learning currently has approximately three million users and is accessible via the web, Android, and iOS applications. Figure 2 shows the user interface of Classting AI Learning.

In this study, data collected in 2021 on students' Korean and math problem-solving histories were used. First, LP information of students was extracted using Korean problem-solving data. The Korean problem-solving data included 71,870 students, 5,862 questions, 73 KCs, and 6,617,370 interactions. Using the learner's LP information, we attempted to trace the learner's knowledge in other subjects (math in this study). A total of 139,271 students, 31,449 questions, 77 KCs, and 11,930,740 interactions were included in the math problem-solving data. The data of students who solved less than five Korean and math problems were removed. As a result, data of 54,470 students, 31,418 questions, 77 KCs, and 7,619,040 interactions were used. Table 1 lists the detailed statistics of the dataset.

<sup>1</sup> <https://www.classting.com>.



**Fig. 2.** User interface of Classting AI Learning.

**Table 1.** Detailed statistics of the dataset.

Number of	Before Processing		After Processing
Students	Math	139,271	54,470
	Korean	71,870	
Questions	Math	31,449	31,418
	Korean	5,862	
KCs	Math	77	77
	Korean	73	
Interactions	Math	11,930,740	7,619,040
	Korean	6,617,370	

### 3.3 Baseline KT Models

In this study, the following baselines were used to verify the effectiveness of the proposed LP-enhanced KT model:

- DKT [29]: a representative RNN-based KT model that employs a single layer of LSTM [14] to estimate the learners' knowledge state [8]
- DKVMN [42]: a MANN-based model utilizing a static key memory to store latent KCs and a dynamic value memory to trace the mastery of corresponding KCs [22]
- SAKT [27]: the first KT model to incorporate an attention mechanism [2]

### 3.4 Experimental Settings

A 5-fold cross validation was performed in which folds are split based on learners, whereby 10% of training set was used as the validation set. The performance was measured using 100 recent sequences of each learner [19]. Each model's embedding size was set to 64; the learning rate was set to 0.001; the batch size was set to 256; and Adam optimizer [18] was used. The early stopping technique was used with 10-epoch patience according to validation AUC.

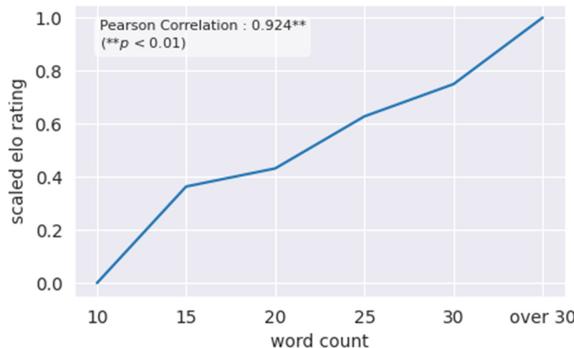
### 3.5 Correlation of LP and Word Problem Performance

The Pearson correlation between the length of the math problem and the average Korean Elo rating score was examined for students who correctly responded to the problems to investigate the relationship between LP and math problem-solving performance. The length of a mathematical problem is defined as the number of words in a sentence. Pearson correlations were examined using Python SciPy 1.10.0 [38].

## 4 Results

### 4.1 Result of the Correlation Analysis Between Students' LP and Their Word-Problem Performance (RQ1)

Figure 3 depicts the Pearson correlation analysis results between the length of the mathematical word problem and the average score of the Korean Elo rating score of students who correctly answered the problem. The length of the word problem sentences correlated strongly with the Korean Elo rating scores (Pearson Correlation = 0.924,  $p < 0.01$ ).



**Fig. 3.** Pearson correlation between students' LP and their word-problem performance.

## 4.2 Experimental Results on KT Models (RQ2)

Table 2 displays the experimental results for baseline models with LP information. Model performance was evaluated using the area under the ROC curve (AUC), accuracy (ACC), and root mean square error (RMSE).

In the experiment, incorporating LP as side information improved the performance of all metrics across all baseline models. The AUC increased from a minimum of 0.7% to a maximum of 7.2% when LP information was included, ACC increased from a minimum of 0.59% to a maximum of 3.78%, and RMSE decreased by a minimum of 0.31% to a maximum of 4.82%.

**Table 2.** Experimental results of KT models.

Model	AUC	ACC	RMSE
DKT	0.748	0.729	0.426
DKT_LP	<b>0.802</b>	<b>0.757</b>	<b>0.405</b>
DKVMN	0.749	0.73	0.426
DKVMN_LP	<b>0.754</b>	<b>0.731</b>	<b>0.424</b>
SAKT	0.787	0.742	0.416
SAKT_LP	<b>0.801</b>	<b>0.755</b>	<b>0.407</b>

To determine the effect of including features on LP information (Elo rating score, time window feature, and word count), an ablation experiment for each feature was performed. Table 3 lists the results of the ablation experiments where ‘elo’ means Elo rating score; ‘wc’ means word count of math problems; and ‘tw’ refers to time windows. The results show that the predictive performance of the model improved when information on LP was incrementally added to each baseline model.

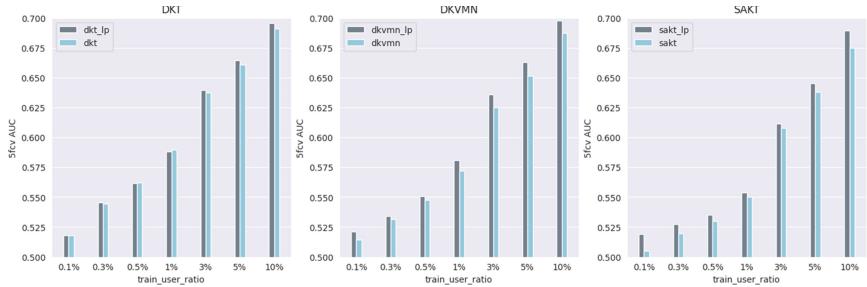
**Table 3.** Ablation experiment results (AUC) for LP-related feature inclusion.

	+elo	+elo+wc	+elo+wc+tw
DKT	0.7989	0.7992	0.8015
SAKT	0.7493	0.7499	0.7538
SAKT	0.7953	0.7983	0.8007

## 4.3 Results of Cold Start Experiment (RQ3)

Similar to many other tasks in machine learning, the KT task has a cold start problem. Cold start problems can be classified into two scenarios [40, 44]. The first scenario is a case of constructing a KT model with a small number of students, while the second

scenario is a case of constructing a KT model with a short learning activity. We examined whether the LP information could alleviate the cold start issue in these two scenarios. In the first scenario, the performance was compared by reducing the number of training users from 10% to 0.1%. In the second scenario, performance was compared by reducing the sequence length of the training data from 50 to 5.

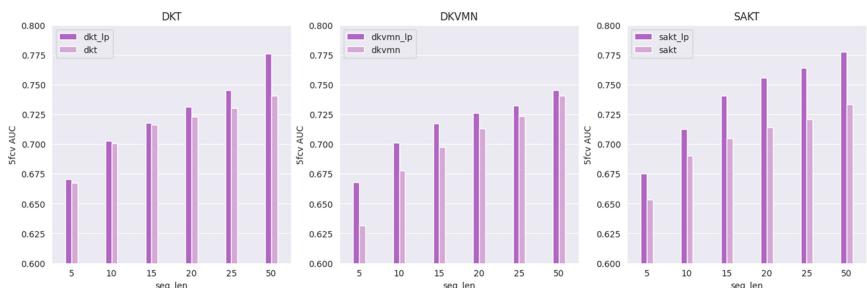


**Fig. 4.** Results of a cold start experiment with a small number of learners.

Figure 4 depicts the results of the cold start experiment conducted with a small number of learners (Scenario 1). As a result of the experiment in which the number of training users was reduced from 10% to 0.1%, the model that utilized LP information outperformed the baseline model. Specifically, performance improved by up to 2.12% when only 10% of the train set was used, 1.47% when 1% was used, and 2.76% when 0.1% was used.

On the other hand, Fig. 5 depicts the results of a cold start experiment conducted with a small number solved problems (Scenario 2). Experiments with sequence lengths reduced from 50 to 5 revealed that models using LP information outperformed baseline models. Specifically, when using LP information, performance improved by up to 5.74% when 5 problem solving histories were used. Using less than 10 improved performance by up to 3.49%, and using fewer than 15 improved performance by up to 5.08%.

In summary, the experiments demonstrate that the AUC performance improved in both scenarios when LP information was included to the three baseline models.



**Fig. 5.** Cold start experiment results for a small number of solved problems.

## 5 Discussion

### 5.1 Effectiveness of LP Information on KT Task

Pearson correlation analysis revealed a strong correlation between the sentence length of the math problem and students' LP. This finding is consistent with previous research showing that LP influences word problem performance in math [12, 37].

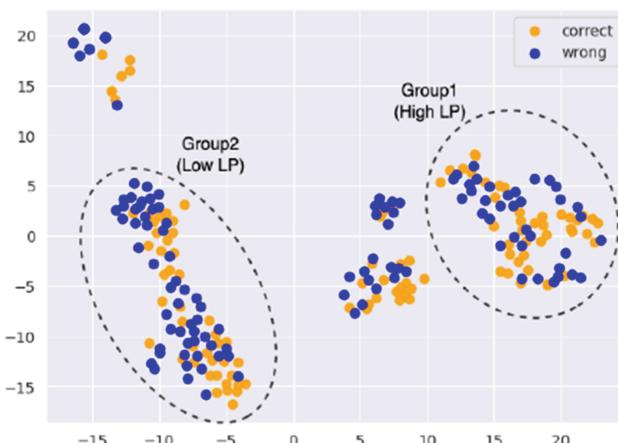
To the best of our knowledge however, it has yet to be determined whether LP information can be useful side information in a KT task. As a result of the experiment in this study, the performance of all RNN-based, MANN-based, and attention mechanism-based KT models was enhanced when LP information was utilized. In addition, the ablation study revealed that all three indicators for measuring LP significantly contributed to performance enhancement. This confirms that the learner's LP information can be meaningfully utilized in the KT task.

Furthermore, experiments were carried out to determine whether the learner's LP information effectively mitigates the cold start problem in the KT field. The experiment confirmed that utilizing LP information to both the scenario of a small number of students and the scenario of a small number of problem-solving alleviated the cold start problem.

In summary, using learners' LP information can improve the performance of KT models, and this information is effective in mitigating the cold start problem.

### 5.2 Possibility as an Assistant for Educational Interventions

Modeling the learner's state of knowledge in light of LP has potential as a supplement for educational interventions. Figure 6 depicts the visualization of LP information from students who solved a random word problem using a T-SNE plot [25]. The students' LP information was encoded using a trained LP-enhanced KT model. Correct responses are marked in orange, while incorrect responses are marked in blue.



**Fig. 6.** T-SNE plot sample for LP information of students.

In Group 1, students with high LP (high Elo score) were distributed. In the case of students who got the problem wrong in this group, it can be determined that the reason was not because of a lack of LP but because of failure to master the concept related to the problem. As a result, the students in this group who got the problem wrong required educational intervention to master related concepts.

Students with low LP (low Elo score) were assigned to Group 2. Students in this group who got the problem wrong, it is likely that they did not comprehend the meaning of the sentence in the problem. As a result, students in this group who were incorrect, required additional educational materials to increase LP as well as additional education on related concepts.

## 6 Conclusion

Language proficiency (LP) is a significant factor influencing students' academic performance. In this study, an LP-enhanced KT model was proposed. The Elo rating scores calculated from the students' Korean problem-solving data and the number of words in the problem were used as LP information. Furthermore, time window features were used to reflect the updating time of the LP information. Data collected from a real-world online learning platform were used for experiments and data analysis. The analysis revealed a strong correlation between the length of math problems and the average LP of the students who correctly answered the questions. The experiments demonstrated that the LP-enhanced KT model proposed in this study outperformed the other baseline models. Furthermore, when the LP information was incorporated, the cold start problem was alleviated in several scenarios. In future work, the performance of the LP-enhanced knowledge tracing model can be validated for various subjects. Furthermore, additional indicators that can measure students' LP can be investigated and applied to KT.

## References

1. Abdelrahman, G., Wang, Q.: Knowledge tracing with sequential key-value memory networks. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 175–184 (2019)
2. Abdelrahman, G., Wang, Q., Nunes, B.P.: Knowledge tracing: a survey. In: ACM Computing Surveys (2022)
3. Adegbeye, A.: Proficiency in English language as a factor contributing to competency in mathematics. Educ. Today **6**(2), 9–13 (1993)
4. Aina, J.K., Ogundele, A.G., Olanipekun, S.S.: Students' proficiency in English language relationship with academic performance in science and technical education. Am. J. Educ. Res. **1**(9), 355–358 (2013)
5. Barkaoui, K.: Examining repeaters' performance on second language proficiency tests: a review and a call for research. Lang. Assess. Q. **14**(4), 420–431 (2017)
6. Bastug, M.: The structural relationship of reading attitude, reading comprehension and academic achievement. Int. J. Soc. Sci. Educ. **4**(4), 931–946 (2014)
7. Batchelder, W.H., Bershad, N.J.: The statistical analysis of a Thurstonian model for rating chess players. J. Math. Psychol. **19**(1), 39–60 (1979)

8. Chan, W.L., Yeung, D.Y.: Clickstream knowledge tracing: modeling how students answer interactive online questions. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 99–109 (2021)
9. Chen, P., Lu, Y., Zheng, V.W., Pian, Y.: Prerequisite-driven deep knowledge tracing. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 39–48. IEEE (2018)
10. Choffin, B., Popineau, F., Bourda, Y., Vie, J.J.: Das3h: modeling student learning and forgetting for optimally scheduling distributed practice of skills. arXiv preprint [arXiv:1905.06873](https://arxiv.org/abs/1905.06873) (2019)
11. Elo, A.E.: The rating of chessplayers, past and present. Arco Pub. (1978)
12. Fuchs, L.S., Fuchs, D., Compton, D.L., Hamlett, C.L., Wang, A.Y.: Is word-problem solving a form of text comprehension? *Sci. Stud. Read.* **19**(3), 204–223 (2015)
13. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
15. Izatullah, S., Nasir, R., Gul, F.: A study to examine the relationship between English language proficiency and academic achievement of students in higher education institutions. *Global Educ. Stud. Rev.* VII (2022)
16. Jun, W.: A study on correlation analysis of academic performance per subject for the gifted children in IT. *J. Gifted/Talented Educ.* **23**(3), 407–419 (2013)
17. Kim, S., Kim, W., Jung, H., Kim, H.: DiKT: Dichotomous knowledge tracing. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 41–51. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-80421-3\\_5](https://doi.org/10.1007/978-3-030-80421-3_5)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
19. Lee, W., Chun, J., Lee, Y., Park, K., Park, S.: Contrastive learning for knowledge tracing. In: Proceedings of the ACM Web Conference 2022, pp. 2330–2338 (2022)
20. Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C.: Improving students' long-term knowledge retention through personalized review. *Psychol. Sci.* **25**(3), 639–647 (2014)
21. Liu, Q., et al.: Ekt: exercise-aware knowledge tracing for student performance prediction. *IEEE Trans. Knowl. Data Eng.* **33**(1), 100–115 (2019)
22. Liu, Q., Shen, S., Huang, Z., Chen, E., Zheng, Y.: A survey of knowledge tracing. arXiv preprint [arXiv:2105.15106](https://arxiv.org/abs/2105.15106) (2021)
23. Liu, Y., Yang, Y., Chen, X., Shen, J., Zhang, H., Yu, Y.: Improving knowledge tracing via pre-training question embeddings. arXiv preprint [arXiv:2012.05031](https://arxiv.org/abs/2012.05031) (2020)
24. Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., Luo, W.: pykt: a python library to benchmark deep learning based knowledge tracing models. arXiv preprint [arXiv:2206.11460](https://arxiv.org/abs/2206.11460) (2022)
25. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *J. Machine Learn. Res.* **9**(11) (2008)
26. Nakagawa, H., Iwasawa, Y., Matsuo, Y.: Graph-based knowledge tracing: modeling student proficiency using graph neural network. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 156–163 (2019)
27. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint [arXiv:1907.06837](https://arxiv.org/abs/1907.06837) (2019)
28. Pelánek, R.: Applications of the ELO rating system in adaptive educational systems. *Comput. Educ.* **98**, 169–179 (2016)
29. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Adv. Neural Inf. Process. Syst. **28** (2015)
30. Racca, R., Lasaten, R.C.S.: English language proficiency and academic performance of Philippine science high school students. *Int. J. Lang. Lit. Linguist.* **2**(2), 44–49 (2016)

31. Shen, S., et al.: Learning process-consistent knowledge tracing. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 1452–1460 (2021)
32. Shen, S., et al.: Convolutional knowledge tracing: Modeling individualization in student learning process. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1857–1860 (2020)
33. Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., Choi, Y.: Saint+: integrating temporal features for ednet correctness prediction. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 490–496 (2021)
34. Song, X., Li, J., Lei, Q., Zhao, W., Chen, Y., Mian, A.: Bi-clkt: bi-graph contrastive learning based knowledge tracing. *Knowl.-Based Syst.* **241**, 108274 (2022)
35. Stoffelsma, L., Spooren, W.: The relationship between English reading proficiency and academic achievement of first-year science and mathematics students in a multilingual context. *Int. J. Sci. Math. Educ.* **17**, 905–922 (2019)
36. Su, Y., et al.: Exercise-enhanced sequential modeling for student performance prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
37. Vilenius-Tuohimaa, P.M., Aunola, K., Nurmi, J.E.: The association between mathematical word problems and reading comprehension. *Educ. Psychol.* **28**(4), 409–426 (2008)
38. Virtanen, P., et al.: SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Method* **17**, 261–272 (2020). <https://doi.org/10.1038/s41592-019-0686-2>
39. Wang, J.: Relationship between mathematics and science achievement at the 8th grade. *Online Submission* **5**, 1–17 (2005)
40. Wu, Z., Huang, L., Huang, Q., Huang, C., Tang, Y.: Sgkt: session graph-based knowledge tracing for student performance prediction. *Expert Syst. Appl.* **206**, 117681 (2022)
41. Yang, Y., et al.: Gikt: a graph-based interaction model for knowledge tracing. In: Machine Learning and Knowledge Discovery in Database: European Conference, ECML PKDD 2020, Ghent, Belgium, 14–18 Sep 2020, Proceedings, Part I, pp. 299–315. Springer (2021)
42. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th international conference on World Wide Web, pp. 765–774 (2017)
43. Zhang, M., Zhu, X., Zhang, C., Ji, Y., Pan, F., Yin, C.: Multi-factors aware dual-attentional knowledge tracing. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management, pp. 2588–2597 (2021)
44. Zhao, J., Bhatt, S., Thille, C., Gattami, N., Zimmaro, D.: Cold start knowledge tracing with attentive neural turing machine. In: Proceedings of the Seventh ACM Conference on Learning@ Scale, pp. 333–336 (2020)



# Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory

Matteo Orsoni<sup>1</sup> , Alexander Pögelt<sup>2</sup> , Nghia Duong-Trung<sup>3</sup> , Mariagrazia Benassi<sup>1</sup> , Milos Kravcik<sup>3</sup> , and Martin Grüttmüller<sup>2</sup>

<sup>1</sup> University of Bologna, Piazza Aldo Moro 90, 47521 Cesena, Italy  
`{matteo.orsoni2,mariagrazia.benassi}@unibo.it`

<sup>2</sup> Leipzig University of Applied Sciences, Karl-Liebknecht-Straße 132, 04277 Leipzig, Germany  
`{alexander.poegelt,martin.gruettmueller}@htwk-leipzig.de`

<sup>3</sup> Educational Technology Lab, German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91C, 10559 Berlin, Germany  
`{nghia_trung.duong,milos.kravcik}@dfki.de`

**Abstract.** Recommending challenging and suitable exercises to students in an online learning environment is important, as it helps to stimulate their engagement and motivation. This requires considering their individual goals to improve learning efficiency on one side and on the other to provide tasks with an appropriate difficulty for the particular person. Apparently, this is not a trivial issue, and various approaches have been investigated in the areas of adaptive assessment and dynamic difficulty adjustment. Here, we present a solution for the domain of mathematics that rests on two pillars: Reinforcement Learning (RL) and Item Response Theory (IRT). Specifically, we investigated the effectiveness of two RL algorithms in recommending mathematical tasks to a sample of 125 first-year Bachelor's students of computer science. Our recommendation was based on the Estimated Total Score (ETS) and item difficulty estimates derived from IRT. The results suggest that this method allowed for personalized and adaptive recommendations of items within the user-selected threshold while avoiding those with an already achieved target score. Experiments were performed on a real data set to demonstrate the potential of this approach in domains where task performance can be rigorously measured.

**Keywords:** Recommender System · Reinforcement Learning · Item Response Theory · Personalized Recommendation · Math Exercises

## 1 Introduction

Conventional university teaching methods usually provide uniform learning exercises for the study groups. Depending on the level of knowledge, exercises can differ in the perception of difficulty by students. For optimal support and challenge of students, an individual selection of tasks is needed, which can be made

based on various metrics, e.g. the level of knowledge or the desired final grade. Individualized learning tries to stimulate the motivation and engagement of students, taking into account theories like the zone of proximal development [16] and flow [4]. The first provides students with tasks beyond their current ability to scaffold the learning process. The second aims to avoid boredom and frustration if the chosen difficulty level does not correspond with the student's ability. Dynamic Difficulty Adjustment (DDA) mechanism, which originated from computer games, is a technique used to automatically adjust the difficulty of online tasks according to the abilities of the user [3, 17], with the goal of keeping the user's attention and engagement. The DDA concept [1] emphasizes the importance of three aspects: the task difficulty (static or dynamic), the user's status (this can include performance or engagement, but also personality and emotions), and the adaptation method, which can be based on rules or data-driven approaches (e.g. probabilistic models, reinforcement learning). Physiologically, user involvement is driven by discovering new knowledge, learning patterns, ideas, and excitement while achieving a particular learning goal [9]. In educational contexts, DDA can ensure that students are presented with tasks suitable for their current level of proficiency, leading to more engaging learning experiences.

One approach to implementing a DDA mechanism is using the Item Response Theory (IRT), a statistical model that estimates an individual's proficiency at a particular task by analyzing their responses to a set of items [5]. This enables to a recommendation of appropriately challenging tasks for the student. However, recommending tasks based on IRT estimates can be suboptimal, as it does not consider the student's learning progress. Therefore, we propose the IRT integration with Reinforcement Learning (RL), which allows for optimizing task recommendations based on the student's past performance.

This study presents a system that utilizes IRT and RL to recommend tasks to first-semester bachelor's degree computer science students taking a mathematics module. Using our proposed method, which employs and compares the Proximal Policy Optimization (PPO) [12], and the synchronous, deterministic variant of the Asynchronous Advantage Actor-Critic [10] algorithm (called A2C), we aim to demonstrate the benefits of personalized task recommendation in the educational settings. In more detail, we incorporated the learner's goals into our recommender system. Literature suggested that specific interventions to set personal academic goals and exam preparation are essential factors contributing to the student's success while in the university [13]. Moreover, goal setting can help students develop a sense of agency, intrinsic motivation, and the ability to manage their learning [11]. We compared the performance of our proposed method to a random baseline, using data from 125 students. The results of our study will provide insight into the effectiveness of using IRT and RL for recommending items in line with the learner's past performance and goals.

In the following, we first reference some related work and background information. Then we present our experiments thoroughly, including the results. Finally, we discuss the outcomes and conclude the paper.

## 2 Related Work

Computerized adaptive assessment methods in well-structured domains like mathematics have a long tradition of selecting tasks according to the student's ability [15], where structured task description schemes allowed for a detailed analysis of student's errors and on-demand generation of task instances facilitated independent student work. During the recent Corona crisis, professional rule-based adaptive learning systems like bettermarks<sup>1</sup> were very popular.

Recent machine learning approaches address the DDA issue also in other domains if there is a significant question bank and users with different competencies [18], considering even individual difficulty levels. This method can be applied when three conditions are met: a discrete action space exists, a feedback signal is a quantitative measure of difficulty, and a target performance value is selected.

DDA can be achieved using statistical models such as IRT [5]. IRT estimates a learner's proficiency based on their responses to a set of items and has been applied in various educational contexts [7]. However, traditional recommendation approaches may not be suitable in educational settings where a student's learning potential changes over time. Reinforcement Learning (RL) addresses this issue by optimizing task recommendations based on the student's past performance and progress [14]. In recent years, the combination of IRT and RL has been proposed as a solution for recommendation in mathematics and cognitive domains. For example, the authors in [8] suggested using an RL system to recommend items based on the student's ability estimates from an IRT model to improve algebra abilities. Also, the study mentioned earlier [18] used IRT to estimate the student's knowledge and RL to adjust task difficulty.

This work is distinct from the previous approaches in recommender systems that combine RL and IRT. It utilizes IRT to estimate the difficulty of items based on the student's past performance and uses this information to compute the expected total score threshold distribution for mathematical modules. This relevant information allowed to integrate into an RL system of the learner's goal to make recommendations that align with the student's objectives.

## 3 Background

In Reinforcement Learning (RL), an agent learns to make decisions by interacting with its environment and receiving feedback through rewards or penalties. The agent's goal is to learn a policy mapping from states to actions that maximize the expected cumulative reward over time [14]. In the present work, we used and compared the performances of two popular RL algorithms: the Proximal Policy Optimization (PPO) [12], and the synchronous, deterministic variant (A2C) of the Asynchronous Advantage Actor Critic (A3C) [10]. PPO is designed to improve the stability and efficiency of policy gradient methods. It is an actor-critic algorithm that uses a value function to estimate the expected cumulative

---

<sup>1</sup> <https://bettermarks.com/>.

reward for a given policy, and it uses a trust region method to optimize the policy. The basic idea of PPO is to optimize the procedure so that the new policy is close to the previous one but with improved expected cumulative reward [12]. The variant of A3C combines the actor-critic method with the advantage function. The actor-critic process separates the policy, which generates the actions, from the value function, which estimates the expected cumulative reward for a given policy. The advantage function estimates the improvement of taking a given action compared to the average action. The term “synchronous” refers to the method of updating the parameters of the actor and critic networks. All agents update their parameters simultaneously using the same synchronous data. In contrast, in the original asynchronous version, each agent updates its parameters independently using its data [10].

## 4 Experiments

### 4.1 Experimental Dataset

This study analyzes a data set collected at Leipzig University of Applied Sciences starting from the winter semester of 2021/22. The data set includes the results of weekly exercises from a mathematics module taken by 125 Bachelor first-year computer science students. To pass the module, students must solve at least 35% of the weekly exercises over the semester. Each weekly practice includes several tasks specific to the topic covered in that week’s lecture. The data set also includes solution attempts made after the semester. The tasks differ slightly for each attempt and student but are assumed to have equivalent difficulty and be based on the same concept. To practice the subject matter, students can work on the exercises and subtasks multiple times. Only the most successful attempt will be counted toward the final grade. The assignments are provided through the OPAL learning management system and ONYX testing software, and some tasks allow using the computer algebra system MAXIMA. The data set is separated into tables for student results and task information. To encourage reproducibility and further investigation, we publish the dataset with the implementation codes on our GitHub repository<sup>2</sup>.

### Result Features

**participant** An ascending number that anonymously references students

**test id** References the weekly exercise (test).

**test attempt** Attempt in which the student solves the weekly exercise

**test score** Points scored by the student

**test pass score** Points to pass the weekly exercise

**test max score** Maximum points of the weekly exercise

**test pass** Status whether the student has passed the weekly exercise

**item id** References the actual subtask in a weekly exercise

---

<sup>2</sup> [https://github.com/MatteoOrsoni/ITS2023\\_Recommending-Math-Tasks](https://github.com/MatteoOrsoni/ITS2023_Recommending-Math-Tasks).

**item attempt** Attempt in which the student solves the subtask

**item datestamp** Timestamp in which the student completed the subtask.

**item sessionStatus** Represents the status of the subtask. (final - The student has solved the task and submitted his/her answers; pendingSubmission - The student has viewed the assignment but has not responded to it; pendingResponseProcessing - The student has entered answers but has not submitted them; initial - The student has not viewed the assignment)

**item duration** Time spent on the subtask

**item score** Points on the subtask scored by the student

**item max score** Maximum points of the subtask

**item candidate responses** Answers from the student

**item correct responses** Correct answers of the subtask

**item candidate responses score** Scores of the student's answers

**item correct responses score** (Maximum)-point scores of the subtask

**item variables** Variable assignments of the subtask execution

## Task Features

**item id** (Equivalent to the result table) references the subtask

**is test** Status whether the item is a test (tests are groupings of subtasks and are usually equivalent to weekly exercises).

**test name** Folder name in which the test file is located

**item description** Tasks description in HTML format

All in all, there are 18576 solutions from a total of 99 different items in a total of 14 modules (including tests for exam preparation) in the data set. Due to the low number of attempts inside some modules and excluding tests for exam preparation, in this study, the analysis focused on 10 modules. On average, the students needed 464 s and achieved an average of 2.18 points per item, with an average maximum score of 3.37 points. Furthermore, students practiced a single item on average 1.85 times, with a maximum of 72 times.

## 4.2 Framework and Baselines

The IRT models have been implemented by using the mirt: a Multidimensional Item Response Theory Package in R [2], while the RL solutions in Python by using the Stable Baseline 3 [6] library. We compared the two RL solutions (PPO, A2C) with a random baseline procedure. According to this, we ran the environment for 1000 episodes, collecting each reward and averaging at the end. For each episode, the actions were taken randomly into the set of those possible. The averaged reward was then taken as baseline values to be compared to the average reward after 1000 episodes estimated by implementing PPO and A2C algorithms. In the following, we will delve deeper into constructing the item difficulty estimation model and the environment in which the RL algorithms were implemented.

### 4.3 Difficulty Level

In the present study, an IRT approach is used to estimate the difficulty of items presented to students in a course each week and to create different thresholds based on the sigmoid distribution of the estimated total score (ETS) of the winning IRT model. It allows us to consider the learners' objectives for that particular module. IRT is a statistical procedure that allows for the discovery of a learner's latent trait for a specific concept and the estimation of different parameters (difficulty, discrimination, and guessing) embedded within the item according to the chosen IRT model. Three other IRT models (1PL, 2PL, 3PL) were compared, and the best one was selected using metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood. The values of these metrics are summarized in Table 1.

**Table 1.** AIC (Akaike Information Criteria), BIC (Bayesian Information Criterion), and LL (Log Likelihood). In bold, the models for each module that reached the significant level  $p < .05$  among others. \* It has not been possible to estimate the parameters due to too few degrees of freedom.

Module	1PL			2PL			3PL		
	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC	LL
1	3452	3550	-1692	<b>3379</b>	<b>3569</b>	<b>-1623</b>	3432	3717	-1617
2	995	1021	-488	<b>932</b>	<b>980</b>	<b>-450</b>	941	1012	-447
3	<b>533</b>	<b>545</b>	<b>-262</b>	533	552	-261	NA*	NA*	NA*
4	<b>1157</b>	<b>1183</b>	<b>-569</b>	1167	1213	-568	1182	1251	-567
5	<b>515</b>	<b>529</b>	<b>-253</b>	521	543	-252	529	562	-252
6	711	728	-350	<b>704</b>	<b>733</b>	<b>-342</b>	713	756	-342
7	844	868	-413	<b>842</b>	<b>884</b>	<b>-405</b>	854	917	-403
8	717	746	-348	<b>711</b>	<b>763</b>	<b>-336</b>	731	809	-336
9	<b>1059</b>	<b>1096</b>	<b>-516</b>	1066	1133	-507	1073	1174	-498
10	<b>664</b>	<b>698</b>	<b>-318</b>	673	735	-310	682	775	-302

The winner over the three possible models was then selected based on the p-value obtained. Only the significantly different model ( $p < .05$ ) from the others was used in further analysis. The estimated total score of the winner IRT model has been then used to estimate the  $\theta$  value, the correspondent threshold difficulty for that specific module. The thresholds have been chosen arbitrarily except for the first, which was the one that allowed the student to pass the module. Two to four thresholds have been set into the RL solution for each module, according to the number of items (number of possible actions) and the steepness of the sigmoid distribution underlying the estimated total score. Moreover, the IRT solution gave us the values of items' difficulty for each module. These values have been used in the RL environment configuration.

#### 4.4 Reinforcement Learning Environment

The recommender system has been developed as a Markov Decision Process (MDP), consisting of a tuple  $(S, A, R, P)$  of elements. The tuple defines the MDP completely, where the agent interacts with the environment. The goal is to find a policy (mapping from states to actions) that maximizes the expected cumulative reward over time. A specific recommender system has a similar MDP representation for each module created. It has been summarized as follows:

1. State Space  $S$ : It represents all possible states of the system. It is related to the answering process of the student according to the item presented in the module. Each state or item in the module has been described as a tuple of five elements  $(d, s, m, dt, t)$ , where:
  - (a)  $d$ : The difficulty of the module items according to the IRT difficulty estimation.
  - (b)  $s$ : The score obtained by the student for each item.
  - (c)  $m$ : The maximum possible score for that specific item.
  - (d)  $dt$ : The difficulty threshold. This parameter does not change until the end of each episode.
  - (e)  $t$ : It is the threshold. This parameter does not change until the end of each episode and is strictly related to the difficulty threshold. It is a numeric value corresponding to the score the student has to obtain by summing the score items.
2. Action Space  $A$ : It represents all possible actions that can be taken in each state.
3. Reward Function  $R$ : It is a function that assigns a numerical reward to each state-action pair  $(s, a)$  and is used to evaluate the quality of different policy choices. We included three different numerical rewards in the environment. A reward is related to the *Difficulty*, *Actions*, and *Episode*.
  - (a) *Difficulty*: For every action chosen by the agent, that is, for every next item chosen, we wanted to create a function that gave a positive reward to the agent if the selected action was in line with the difficulty threshold of the item estimated by the IRT model and the threshold chosen by the user. In this way, we wanted to favor items that had difficulties equal to or lower than the user's needs to reach a certain threshold, discouraging items that were too difficult to achieve the goal.

$$R_D = \begin{cases} \kappa_1 & \text{if } s_t \leq dt, \forall a \in A \\ \kappa_2 & \text{otherwise} \end{cases} \quad (1)$$

In this function, if the action selected by the agent is in line with the IRT estimate and is less than or equal to the user's threshold, the agent will receive a positive reward  $\kappa_1$ . If not, the agent will receive a reward of zero  $\kappa_2$ .

- (b) *Actions*: For every action taken by the agent, this reward function was constructed to track the actions taken and to avoid items for which the

student has received a score equal to the highest possible from being presented again.

$$R_A = \begin{cases} \kappa_3 & \text{if } a_t \in \text{actions\_used}, \forall a \in A \end{cases} \quad (2)$$

where *actions\_used* means the set of actions/items for which the student has already achieved the highest possible score. If the agent recommended an action in the *actions\_used* it received a negative reward.

- (c) *Episode*: The last reward function was related to the episode conclusion. Each episode was set to have a maximum duration related between (54%–150%) longer than the number of possible actions, to allow the agent to present the items again for which the subject had not reached the highest possible score and to reach the thresholds with the items with higher difficulty. If the agent could reach the established threshold within the maximum length of the episode, it received a positive reward; otherwise, it did not receive any reward.

$$R_E = \begin{cases} \kappa_4 & \text{if } s_t + s_{t+1} + \dots + s_{t+n} \geq t \\ \kappa_5 & \text{otherwise} \end{cases} \quad (3)$$

At the end of each episode, the overall reward function was created based on the three functions. If the agent achieved a cumulative score on the items equal to or higher than the set threshold, then the reward function  $R$  included  $R_D + R_A + R_E$ . Otherwise, it only had  $R_D + R_A$ .  $R_D$  and  $R_A$  are considered intermediate rewards that should guide the agent in its choice of future actions.

- 4. Transition Probability Function  $P$ : It defines the probability of transitioning from one state to another after taking a specific action.

## 4.5 Hyperparameters

In this section, we summarized the hyperparameters used in each module. In Table 2, we have included the hyperparameters for configuring the reinforcement learning environment. Specifically, the PPO and A2C algorithms were trained for  $10^5$  timesteps across all modules, each for 1 h. The learning rate was set at  $10^{-7}$ . Finally, the training algorithms were based on a policy object that implements an actor-critic approach, utilizing a 2-layer MLP with 64 units per layer [6]. It is true for some modules, while others utilize a custom network architecture.

Table 3 summarizes the hyperparameters associated with the custom environment, including the maximum length of each episode and its relationship with the number of possible actions. It also shows the number of thresholds considered in each module and the numeric values of the threshold ( $t$ ) based on the estimated total score and the corresponding  $\theta$  value ( $dt$ ) obtained from the winning IRT solution. In addition, it considers  $N$  as the number of complete subjects' recordings for each module. This value has been extracted using the student's first attempt for each task in each module.

**Table 2.** Hyperparameters are implemented in both the PPO and A2C algorithms. lr: learning rate, ts: timesteps, Custom\_net: Custom\_network, policy: the policy implemented.

Module	RL configuration				
	policy	Custom_net	ts	lr	
1	mlp	Yes: [128, 64]	$10^5$	$10^{-7}$	
2	mlp	No	$10^5$	$10^{-7}$	
3	mlp	No	$10^5$	$10^{-7}$	
4	mlp	Yes: [128, 64]	$10^5$	$10^{-7}$	
5	mlp	No	$10^5$	$10^{-7}$	
6	mlp	Yes: [64, 32]	$10^5$	$10^{-7}$	
7	mlp	No	$10^5$	$10^{-7}$	
8	mlp	Yes: [64, 32]	$10^5$	$10^{-7}$	
9	mlp	No	$10^5$	$10^{-7}$	
10	mlp	Yes: [64, 32]	$10^5$	$10^{-7}$	

## 4.6 Experiment Results

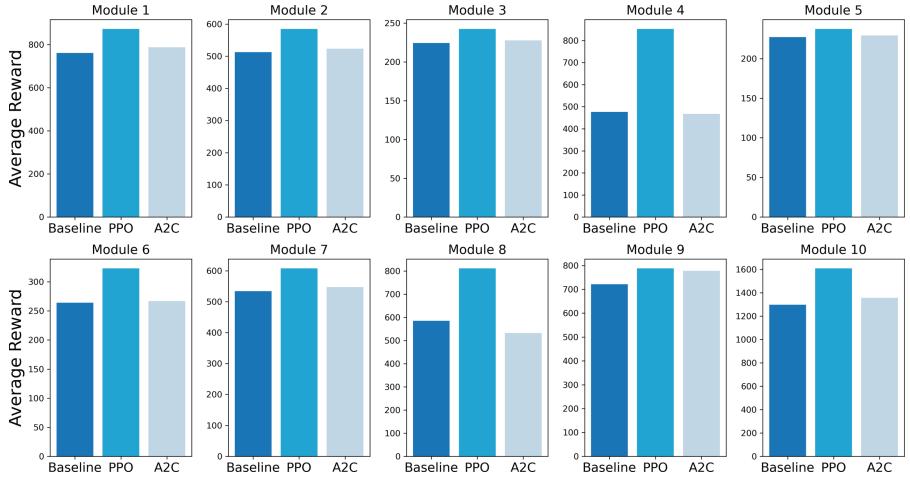
This study evaluated the performance of two reinforcement learning solutions, PPO and A2C, and a random baseline solution in collecting average rewards after 1000 episodes. The results, as illustrated in Fig. 1, demonstrate that the PPO solution outperformed both the A2C solution and the random baseline across all modules presented to subjects. A comparison of the mean improvement in collecting average cumulative rewards among the three solutions is summarized in Table 4. Evidently, the PPO solution achieved, on average, a 22.83% increase in rewards over the random action solution. Furthermore, this advantage in collecting rewards was consistent across all modules, with an improved range of 4.50% to 78.94% compared to the baseline. In contrast, the A2C algorithm demonstrated only a moderate improvement in collecting rewards, with an average increase of 1.29% over the baseline across all modules. This improvement was inconsistent, with a range of -8.99% to 7.69%.

## 5 Remarks and Discussion

The presented research centers on developing a recommender system that utilizes reinforcement learning and item response theory to enhance item recommendations for first-year bachelor's students in computer science taking a mathematics module. The integration of RL and IRT allows for personalized and adaptive recommendations based on the estimated difficulty threshold, enabling the system to suggest items within the user-selected threshold while avoiding items for which the student has already achieved the maximum possible score. In other words, the higher the threshold set by the student, the more complex the recommended items were, according to the  $\theta$  value of the ETS distribution. This aspect

**Table 3.** Hyperparameters in the environment configuration. Length (%) is related to the maximum episode length and the relative percentage compared to the number of possible actions.  $n^o t$ : is the number of thresholds included in the environment for that module.  $dt$ : is the difficulty threshold.  $t$ : is the threshold value.  $N$  is the number of complete subjects' recordings for each module.

Module	Environment configuration				
	length (%)	$n^o t$	$dt$	$t$	$N$
1	20 (+82%)	4	[-2.80, -1.96, .03, .57]	[10.5, 15.5, 25.9, 28.3]	131
2	20 (+150%)	3	[-1.05, 0, 2]	[10.5, 23, 25]	129
3	8 (+100%)	2	[1.11, 2.56]	[10.5, 15.1]	132
4	20 (+150%)	3	[-0.15, 0.63, 2.20]	[10.5, 16, 26]	87
5	10 (+150%)	3	[-0.15, 1.05, 1.60]	[10.5, 15.4, 19.3]	99
6	10 (+100%)	3	[0.63, 1.05, 2.02]	[10.5, 13, 15]	100
7	20 (+150%)	3	[-0.75, 0.33, 1.17]	[10.5, 22.4, 32.2]	53
8	10 (+100%)	4	[18.7, 27, 38, 48]	[-1.24, 0.03, 0.75, 3, 22]	44
9	20 (+54%)	4	[-2.68, -1, 0, 1]	[10.5, 24.3, 33, 40]	50
10	25 (+79%)	4	[-0.27, 0.33, 1.12, 2.62]	[14.525, 17.31, 23, 31.5]	21



**Fig. 1.** Comparing the Performance of RL Algorithms and Baseline Across Modules. Average reward after 1000 episodes comparing Baseline, PPO, and A2C recommendations.

is particularly relevant because of the significance of allowing learners to determine their own difficulty level. As previously mentioned, interventions aimed at establishing personal academic goals are a crucial component in promoting student success. Moreover, by facilitating goal setting, students can develop a stronger sense of agency, intrinsic motivation, and self-directed learning skills.

**Table 4.** Performances comparison in the average reward between PPO and Baseline and A2C and Baseline actions. The values are expressed in percentual terms.

Module	PPO/Baseline	A2C/Baseline
1	+14.61	+3.44
2	+13.98	+2.08
3	+8.00	+1.47
4	+78.94	-1.87
5	+4.50	+0.9
6	+22.39	+1.1
7	+13.83	+2.53
8	+38.72	-8.99
9	+9.20	+7.69
10	+24.11	+1.27
Avg	+22.83	+1.29

The results demonstrate that incorporating RL solutions leads to improved performance, as measured by the average reward collected by the agents over 1000 episodes. Specifically, as highlighted in the results section, the PPO algorithm outperforms the A2C algorithm in every module, achieving an average reward that is 22.83% higher than the baseline.

Nevertheless, some considerations have to be mentioned. Firstly, while we have seen an improvement in the average reward collected, we need to determine if the recommendations benefit students. A future study should investigate this aspect more thoroughly. Secondly, our study used offline students' data for which we had complete answers for a module. It allowed us to use each episode as a new user and the answers as a transition over time for a specific user for that episode. This approach led to a policy strictly dependent on the answers collected, the students who answered all the items in each module, and the sample size and the possible transitions it learned. We only had a few dozen subjects for some modules who answered the entire set of items. In future studies, we plan to use this policy as a starting point and enhance it by incorporating online interaction between the user and the system. In addition, we used arbitrary thresholds derived from the estimated total score of the IRT solution, but there may be better options for achieving better results on test evaluations. In a future study, we plan to integrate this aspect by finding the best possible thresholds for each module that can provide the most informative guide for students to succeed on test evaluations. Lastly, we focused on item difficulty rather than the student's ability to solve a specific task. A future study should include this aspect in the RL environment to suggest items that also consider the student's ability to solve them.

## 6 Conclusion

This study presented a system for enhancing item recommendations for first-year bachelor's computer science students taking a mathematics module. The integration of Reinforcement Learning (RL) and Item Response Theory (IRT) allowed for personalized and adaptive recommendations based on the estimated difficulty threshold, enabling the system to suggest items within the user-selected scale while avoiding items for which the student has already achieved the maximum possible score. Results showed that incorporating RL solutions improved performance as measured by the average reward collected by the agents over 1000 episodes. Specifically, the proximal policy optimization algorithm outperformed the A2C algorithm in every module, achieving an average reward that is 22.83% higher than the baseline. Overall, this study provides valuable insight into the effectiveness of using IRT and RL for dynamic difficulty adjustment and the benefits of personalized task recommendation in educational settings. The proposed method can potentially improve learning outcomes and engagement in the domain of mathematics as well as other areas.

**Acknowledgments.** The authors would like to thank the German Federal Ministry of Education and Research (BMBF) for their kind support within the project *Personalisierte Kompetenzentwicklung und hybrides KI-Mentoring* (tech4compKI) under the project id 16DHB2208.

## References

1. Arey, D., Wells, E.: Balancing act: «the art and science of dynamic difficulty adjustment». In: Game Developers Conference (2001)
2. Chalmers, R.P.: Mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* **48**, 1–29 (2012)
3. Constant, T., Levieux, G.: Dynamic difficulty adjustment impact on players' confidence. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
4. Csikszentmihalyi, M.: Flow. The Psychology of Optimal Experience. New York (Harperperennial) (1990)
5. Embretson, S.E., Reise, S.P.: Item Response Theory. Psychology Press (2013)
6. Hill, A., et al.: Stable baselines3 (2020). <https://github.com/DLR-RM/stable-baselines3>
7. Hori, K., Fukuhara, H., Yamada, T.: Item response theory and its applications in educational measurement part i: item response theory and its implementation in R. *WIREs Comput. Stat.* **14**(2), e1531 (2022)
8. Leite, W.L., et al.: A novel video recommendation system for algebra: an effectiveness evaluation study. Association for Computing Machinery, New York (2022)
9. Lopes, J.C., Lopes, R.P.: A review of dynamic difficulty adjustment methods for serious games. In: Pereira, A.I., Košir, A., Fernandes, F.P., Pacheco, M.F., Teixeira, J.P., Lopes, R.P. (eds.) OL2A 2022. CCIS, vol. 1754, pp. 144–159. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-23236-7\\_11](https://doi.org/10.1007/978-3-031-23236-7_11)
10. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016)

11. AI for Research: Student goal setting: an evidence-based practice student goal setting (2018)
12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. In: International Conference on Learning Representations (2017)
13. Stelnicki, A.M., Nordstokke, D.W., Saklofske, D.H.: who is the successful university student? an analysis of personal resources. *Can. J. High. Educ.* **45**(2), 214–228 (2015)
14. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. *J. Artif. Intell. Res.* **4**, 1–53 (1998)
15. Tvarožek, J., Kravčík, M., Bieliková, M.: Towards computerized adaptive assessment based on structured tasks. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) AH 2008. LNCS, vol. 5149, pp. 224–234. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-70987-9\\_25](https://doi.org/10.1007/978-3-540-70987-9_25)
16. Vygotsky, L.S., Cole, M.: Mind in Society: Development of Higher Psychological Processes. Harvard University Press (1978)
17. Xue, S., Wu, M., Kolen, J., Aghdaie, N., Zaman, K.A.: Dynamic difficulty adjustment for maximized engagement in digital games. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 465–471 (2017)
18. Zhang, Y., Goh, W.: Personalized task difficulty adaptation based on reinforcement learning. *User Model. User-Adap. Inter.* **31**, 753–784 (2021)



# Automatic Execution of the Ontological Piloting Procedures

Marc-Antoine Courtemanche<sup>1</sup>(✉), Ange Tato<sup>2</sup>, and Roger Nkambou<sup>1</sup>

<sup>1</sup> Université du Québec à Montréal, Montréal, Canada

courtemanche.marc-antoine@courrier.uqam.ca, nkambou.roger@uqam.ca

<sup>2</sup> École de technologie supérieure, Montréal, Canada

nyamen\_tato.angea\_drienne@uqam.ca

**Abstract.** Intelligent educational systems currently use ontological knowledge modeling for their expert component. The associated semantics of the ontological languages allow for capturing highly complex domain knowledge. However, the automatic manipulation of knowledge to support the execution of a complex task by humans is a current research issue. This paper presents a solution for interpreting a complex ontological knowledge formalization about aircraft piloting procedures. This specific task is accomplished by human actors where multiple biases may compromise the safe execution. We are developing a solution to support pilots in their tasks' execution. The domain theory interpretation must allow the reference model's automatic execution in a complex environment. A reference model of the domain theory has been formalized with ontological language, a task ontology for the execution taxonomy, and a domain ontology containing the knowledge about the execution environment have also been previously developed. In this work, we propose an interpreter capable of manipulating the reference knowledge of the execution taxonomy within a simulation environment. Semantic rules are used to evaluate environmental constraints, task precondition, and execution to allow this automatic manipulation and execution. Once the solution is connected to a simulation environment, a reasoner manipulates the knowledge in order to evaluate the environment and autonomously execute the required tasks. This work is an intermediary step for providing a complete solution for supporting pilots' execution of reference procedures in a simulation environment.

**Keywords:** Ontology · Task Ontology · Domain Ontology · Intelligent Tutoring Systems · Piloting procedures · Knowledge representation

## 1 Introduction

Air travel is considered a highly efficient and safe means of transportation over long distances. To maintain safe operations and support the public's confidence, multiple procedures must be followed to identify, control and manage the risk associated with this operation. By executing procedures in a very dynamic and

complex environment, pilots have an important role in the safe execution of flight procedures. Since humans accomplish the procedures, multiple human factors can affect the level of attention of the pilots. To support safe execution, aircraft manufacturers are providing piloting procedures that must be followed. We are developing a solution for supporting the execution of piloting procedures. Our work provides intelligent systems for following and monitoring the execution within a complex environment. The goal is to support the pilot, not to replace the human, who is an essential actor that can sometimes be limited by standard human factors.

With this perspective of supporting the execution of piloting procedures, we previously suggested a reference model for piloting procedures [4]. This reference model is a set of production rules with a high semantic link to the execution environment in order to provide a framework for automatic problem resolution. In other words, the normal and abnormal procedures are decomposed with production rules to identify the pilots' expected actions for each situation. This automatic resolution is accomplished by integrating the environmental parameters in the reference framework. Previously, Mitrovic et al. (2003) suggested a set of production rules as the procedural knowledge that can support a process of task execution monitoring once it is linked to declarative knowledge [12].

In this paper, we propose a framework for interpreting and automatically executing the production rules contained in the reference model. This execution framework has to take advantage of the procedural and declarative knowledge available. Since the reference model is highly linked to the execution environment, we suggest integrating the execution with a simulation environment. The next sections are structured as follows. In Sect. 1, issues and related work are presented. Section 2 presents the reference model's general taxonomy structure comprising tasks and domain ontologies. In Sect. 3, the results of the reference model interpreter are detailed. Section 4 concludes the paper and suggests future work perspectives.

## 2 Issues and Related Work

The reference model [4] is formalized with Web Ontology Language (OWL) [20], which is a language that supports a high level of detail with defined semantic links with the elements of the taxonomy. The reference model contains the reference knowledge required for the execution but cannot automatically execute the reference procedures and monitoring capabilities. Previously, ontologies have been used for the formalization of procedural rules. Mizoguchi and Bourdeau [14] support ontological engineering as a levering alternative to the major drawbacks that intelligent tutoring systems face. Using ontology to formalize execution rules, they demonstrated high support specific to procedural knowledge. Multiple intelligent tutors have used production rules as their base framework e.g. Cognitives Tutors [2,3]. Mizoguchi [13] suggests PROLOG for specifying a set of rules that the intelligent tutor can use for interpreting the data by inference reasoning. The rules are the scenario theory that is used by the intelligent tutor

for understanding and interpreting the knowledge acquired by the student. The rules are used in a forwarding and backward reasoning strategy, allowing the system to understand how the student interacts with the knowledge. Others used ontology as the main method for formalizing the theory of a learning system. Specifically, SMARTIES is a software allowing the modelisation of specific knowledge by exploiting the OMNIBUS ontology that is specific to education [5]. One major problem of some intelligent tutoring systems is the lack of quality in designing the scenarios. The SMARTIES software and OMNIBUS ontology support this problem by suggesting a set of tools for supporting the formalization of high-quality domain knowledge. Based on the cognitive architecture ACT-R, Oltramari & Lebriere [16] used ontologies to formalize production rules to exploit their semantic characteristics and allow the cognitive tutor to understand the structures and substructure of the procedural knowledge.

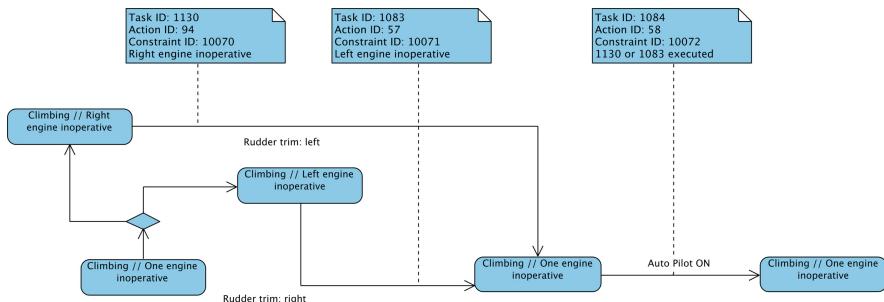
Despite the high quality of procedural knowledge formalization provided by semantic links specific to ontological knowledge, the main issue is the execution of the model. Other works have shown the advantage of using ontologies, but none of these works have provided an execution model specific to ontological procedural knowledge. Our objective is to suggest an alternative to this limitation that some authors are facing. By providing a way of execution, the reference model we previously presented should be more accessible and usable in different simulation environments. However, the automatic execution of the model is not the only issue required for the interoperability of the architecture. Ontological knowledge is typically not supported by aircraft simulator software. Because of that limitation, we have to provide an alternative for keeping the high flexibility of the ontological knowledge but providing a way to receive and send data to the environment with a standard programming language that is typically used by simulation software.

In the aviation domain, ontology engineering is used to support some of the operation processes. Since multiple actors from multiple locations are involved in the process of aircraft operation, the choice of ontology is supported by its flexibility. Since different systems operating with different software are incompatible, the ontology is a good choice for interoperability. Air Traffic Management Ontology (ATMONTO) [7,8] contains knowledge about airspace, flight characteristic, and flight operation condition for the United States. This ontology is supported by the NASA AMES research center to support the interoperability between multiple actors with different standards. Supported by ATMONTO, Sheng et al. [17] suggests an ontology for decision-making about trajectories and weather. In this ontology, a representation of the spatial and temporal factors supports the process of trajectory decisions for safe and efficient destinations. By using an ontology-based system for automating the decision process, a high volume of data cannot be processed by humans to support high-quality decision-making. None of the existing work has focused on an interpreter and automatic execution of the production rules contained in a reference model ontology of piloting procedures.

### 3 The Reference Model

Our reference model ontology comprises two main parts and is formalized with Web Ontology Language (OWL) [20]. The first part is the domain ontology which is the terminological set related to the execution environment. The main objective is to support the execution within the complex environment. The other part is the task ontology, a general taxonomy of the execution domain specific to aviation procedures. Both ontologies are highly related, and multiple semantic connections are formalized between parts of both ontologies, making the execution model anchored to the context.

Before being formalized within the ontologies, expert pilots must decompose and validate the procedure. With this objective in mind, normal and abnormal procedures commonly encountered during takeoff have been decomposed with a standardized visual tool. The statechart diagram from the Unified Modeling Language (UML) [10] has been selected for its capability of capturing knowledge specific to workflow sequence. This choice has been supported by its characteristic of being simple and easily manipulated by domain experts with no prior experience with ontology. Many works have used the UML as a standard for decomposing procedural knowledge in the aviation domain. The Aviation Scenario Definition Language (ASDL), in its context, is a language that shows the possibility of formalizing aircraft navigation scenarios within the environment where multiple actors are involved [6]. However, the ASDL has not demonstrated its effectiveness for modeling the task specifically accomplished by the pilots. Figure 1 shows an excerpt of the abnormal procedure where one engine is inoperative during the takeoff and above the decision speed. Following the UML standard, each box is a specific state of the aircraft and the arrows are a change imposed in the environment by the pilots.

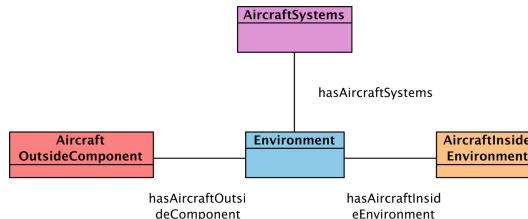


**Fig. 1.** Excerpt of the engine failure procedure after decision speed.

### 3.1 The Domain Ontology

The domain ontology is a terminological set related to the aviation domain. Formalized with the OWL language, this ontology contains the parameter of the execution environment. The parameters stored in this ontology are used as a reference for the production rules stored in the task ontology. Since the execution rules are constrained to specific environment parameters, the reference model stores the specific values that must be evaluated in the domain ontology. The exact process is used for the action that has to be executed in the environment. The changes are pushed to the domain ontology. Since multiple parameters must be stored in this taxonomy, the domain ontology has a general class with three specializations (see Fig. 2). There is a short definition of these classes:

- **Environment:** the main general class composed of the three specialized classes.
- **Aircraft inside environment:** contains all parameters that can be found inside the aircraft. Here we can find the position of different switch and instruments values normally inside the cockpit.
- **Aircraft outside environment:** contains all parameters of the environment outside the aircraft. Wind speed, aircraft position, airspeed, etc. are parameters that can be found in this specialization.
- **Aircraft systems:** some aircraft systems that are not physically outside or inside the aircraft have to be considered. Different systems like the landing gear and engine parameters can be found in this group.



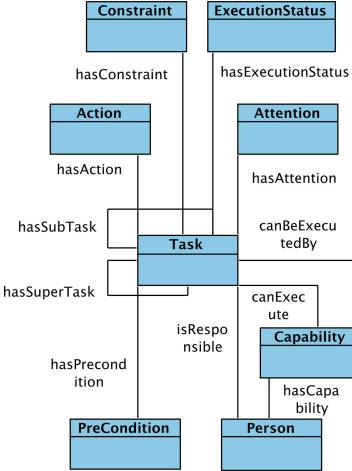
**Fig. 2.** Domain ontology

### 3.2 The Task Ontology

The task ontology is the main terminological set around which the execution is structured. Within this ontology, we can find a general terminology specific to the execution domain. In other words, it is the execution structure used to decompose the reference procedures. This execution decomposition is linked to the execution environment by strong semantic links to the domain ontology. To better understand the objective of a task ontology, this definition shows the important aspect of a task ontology: “Task ontology is a system of vocabulary

for describing the problem-solving structure of all the existing tasks domain-independently. It is obtained by analyzing task structures of real-world problems. The task ontology design is done to overcome the shortcomings of generic tasks and half weak methods while preserving their basic philosophies. The ultimate goal of task ontology research includes providing vocabulary necessary and sufficient for building a model of human problem-solving processes.” [15]. By analysis and decomposition of the general execution process specific to the execution of aircraft piloting procedures, we built a specific taxonomy of task execution. Figure 3 is a visual representation of the task ontology. Each class of the task ontology has a specific meaning. There is a short definition for each of the terms used for decomposing the procedures:

- **Task:** this is the main axiom of the task model. Each task is abstract, with specific identification (ID) and associated parameters.
- **Capability:** each task has to be accomplished by the pilot flying or the pilot monitoring. The goal of the capability is to associate which of the pilot has to accomplish each specific task.
- **Person:** independently of the capability of execution, the pilot flying and pilot monitoring can be the pilot or copilot. This term specifies the role of the pilot and the copilot.
- **Precondition:** this axiom specifies other tasks that must be accomplished before permitting the execution of other tasks. This term allows capturing the chronological order in which the task has to be executed.
- **ExecutionStatus:** this axiom specifies whether or not each specific task is accomplished. This execution status is a way of monitoring the execution process.
- **Action:** with strong links to the domain ontology, this axiom specifies the action to impose on the environment. The action parameter and the value imposed on the environment are specified with this axiom.
- **Constraint:** since the task execution has to be restricted to a specific environment state, this axiom specifies the parameter and the value that has to be reached before allowing the execution. Multiple types of constraints can be formalized. Some restriction is strict value, minimum value, or maximum value. Combined with the precondition restriction, a strong semantic link to the environment from the constraints limits the execution to specific parameters.
- **Attention:** each of the tasks has an attention score associated with the workload required for the execution of the task. This level of attention required for the execution of a task is semantically anchored to the execution model. By linking the attention data to the execution model, we are providing an efficient way of monitoring the level of attention during the automatic execution of the procedures.



**Fig. 3.** Domain ontology

## 4 Results

As previously stated, the literature is not providing a solution for the automatic execution of ontological task models. We built an execution structure based on the task ontology to support this lack of execution alternatives.

### 4.1 The Interpreter

To support the execution model, we are suggesting the integration of Semantic Web Rule Language (SWRL) [19] which integrates the syntax in the form of “Horn” rules that are typically not supported in the semantic of OWL [11]. SWRL is a rule language formalism built on top of ontologies to support some semantic limitations. Integrating this new syntax aims to reason with the knowledge previously formalized with OWL. SWRL rules consist of an antecedent or conditions (body) and a consequent (head). Rule 1 is a generic example showing how we can specify multiple levels of relations between family members with the SWRL syntax.

$$\text{hasParent}(A, B) \wedge \text{hasBrother}(B, C) \rightarrow \text{hasUncle}(A, C) \quad (1)$$

Using the SWRL syntax, we built a set of rules for exploiting the decomposition of execution of the task ontology (see Fig. 3). Once the rules are executed together, the reference model is automatically executable. The execution rules are grouped within one of the three following groups:

- **Constraint evaluation:** this set of rules is specific to evaluating the environment. Since some of the tasks are restricted to a specific environment state,

the task ontology specifies the required parameters for permitting the execution. For each of the task, the constraint is specified in the task ontology (see Fig. 3). Once the environment meets the constraint, the constraint authorization is given to the task. This authorization is later used by the sets of rules specific to the execution. Equation 2 is one of the constraint evaluation rules where the environment parameter has to be between two values. The minimum and maximum values are stored in the task ontology, and the actual value of the environment is pulled from the domain ontology. Once the value from the environment is between the minimum and maximum acceptable value, the constraint permission is attributed.

$$\begin{aligned}
 & \text{task : Constraint(?c)} \wedge \text{task : hasEvaluationCriteria(?c, ?ec)} \\
 & \wedge \text{dom : hasActualValue(?ec, ?val)} \wedge \text{task : hasMinValue(?c, ?minval)} \\
 & \wedge \text{task : hasMaxValue(?c, ?maxval)} \wedge \text{swrlb : lessThan(?val, ?maxval)} \\
 & \quad \wedge \text{swrlb : greaterThan(?val, ?minval)} \\
 \rightarrow & \text{task : hasConstraintPermission} \wedge (\text{?c}, \text{task : ConstraintOk}) \quad (2)
 \end{aligned}$$

- **Precondition evaluation:** since the task execution model has to be executed with a specific chronological order, the precondition is a set of rules restricting the execution to a flexible order. This order is specified by the task ontology, where a specific task has to be executed before permitting the execution of other tasks. Since this precondition evaluation is only between the tasks, the rules are only related to the task structure. This set of rules does not refer to the domain ontology. Equation 3 is one of the rules evaluating the status of execution of the task that has to be executed before permitting the execution of the next task. Once the previous task is executed, precondition permission is attributed to the task currently at evaluation.

$$\begin{aligned}
 & \text{task : Task(?t)} \wedge \text{task : hasPrecondition(?t, ?pc)} \\
 & \wedge \text{dom : hasExecutionStatus(?pc, task : Executed)} \\
 \rightarrow & \text{task : hasExecutionPermission} \\
 & \wedge (\text{?t}, \text{task : PreconditionOk}) \quad (3)
 \end{aligned}$$

- **Task execution:** once the constraint and the task precondition is met, this set of rules is triggered for executing the task. The consequent of the rules is the actual execution in the environment by manipulating the parameters in the domain ontology (see Fig. 2). Since the environment is highly dynamic and the execution is restricted to a specific order, only one task can be executed at a specific time (no parallel execution). Once the execution is triggered, the actions associated with the task are gathered and pushed to the environment. The action is contained in the task ontology where the parameter of the domain ontology and the values to push are specified. The rule retrieves the parameters and values, and the action parameter is directly modified in the domain ontology, which is the specific execution process. Equation 4 is one of the execution rules where the task has two constraints and two actions. We

can see how the actions associated with the task are semantically forwarded to the domain ontology.

$$\begin{aligned}
 & \text{task : Task(?t)} \wedge \text{task : hasNbConstraint(?t, 1)} \wedge \text{task : hasNbAction(?t, 1)} \\
 & \quad \wedge \text{task : hasPreconditionPermission(?t, task : PreconditionOk)} \\
 & \quad \quad \quad \wedge \text{task : hasConstraint(?t, ?c1)} \wedge \\
 & \quad \quad \quad \text{task : hasConstraintPermission(?c1, task : ConstraintOk)} \wedge \\
 & \quad \text{task : hasAction(?t, ?a1)} \wedge \text{task : hasActionParameter(?ap1, ?ap1)} \wedge \\
 & \quad \quad \quad \text{task : hasActionValue(?a1, ?av1)} \\
 & \quad \rightarrow \text{task : hasExecutionStatus(?t, task : Executed)} \\
 & \quad \wedge \text{task : hasActionParameter(task : LiveExecution, ?ap1)} \\
 & \quad \quad \quad \text{task : hasActionValue(task : LiveExecution, ?av1)} \quad (4)
 \end{aligned}$$

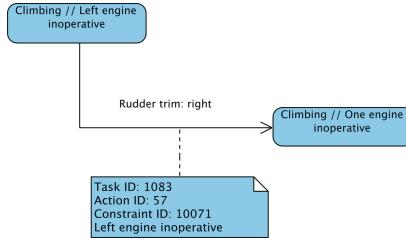
These rules are highly related to the task ontology (see Fig. 3). By following this specific taxonomy, we ensure that by adding production rules in the future, they will be fully supported by the execution model. We can also note that each of the rules set has to make a specific link between the task and domain ontologies. The specific links are formalized in the task ontology and evaluated or executed during the execution of the model.

In order to let the interpreter execute the rules automatically, we have to use a reasoner supporting OWL knowledge and SWRL rules. In other words, the reasoner is activating the rules allowing the automatic manipulation of the knowledge. Pellet has been selected for its ability to support the description logic associated with OWL, and the individuals with its ability to support SWRL rules [18]. Multiple reasoners are available, but the ability to reason over the individuals by using the SWRL rules makes Pellet an efficient choice. According to Abburu and Sunitha [1], the following reasoner supports SWRL rules and individuals: RACER, SWRL-IQ, or HermiT.

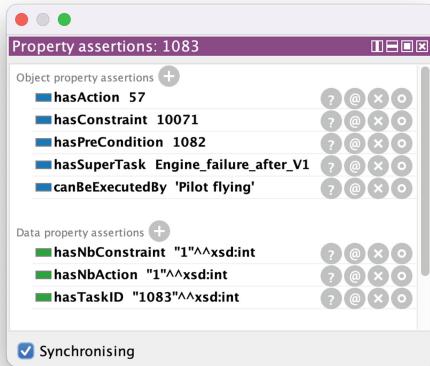
## 4.2 The Execution Process

From Fig. 1, we took an excerpt of task 1083, which is available in Fig. 4. To execute this task, we have to evaluate the task precondition, the constraint from the environment, and then execute the task. For the execution process, data from the task ontology are available in Fig. 5. The process of task execution is as follows:

- 1. Constraint evaluation:** as shown in Fig. 5 and 4, the constraint of task ID 1083 is the constraint with ID 10071 (see Fig. 6). This constraint specifies that the left engine has to be inoperative to allow the execution. For this constraint, the Rule 2 finds the evaluation criteria of the constraint ID 10071 (see Fig. 6), which is the left engine operation state, retrieves the actual value, and if the value shows that the engine is inoperative, the rule is triggered. The constraint permission is attributed to the task.

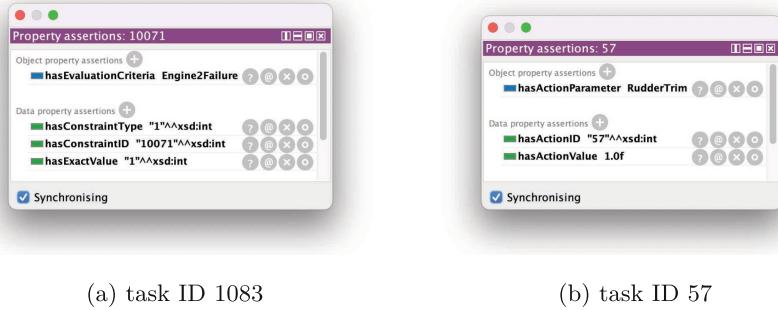


**Fig. 4.** Excerpt of the statechart diagram for task ID 1083



**Fig. 5.** Excerpt of the task ontology for task ID 1083

2. **Precondition evaluation:** from Fig. 5 we can see that the task precondition is another task with the ID 1082. In order to retrieve the execution status of the previous task and authorize the execution, Rule 3 uses the ID of the previous task to find the execution status. If the previous task is executed, task ID 1082 receives the precondition authorization.
3. **Task execution:** once the precondition and constraint authorization are given, the execution rule is triggered and the execution is automatically pushed to the environment. For task ID 1083, Rule 4 evaluates this precondition and constraint and takes the parameter and value from the associated action (action ID 57) (see Fig. 6) specified in Rule 5. Once the rule is triggered, the parameter and value from action ID 57 (see Fig. 6) are pushed to the domain ontology. This new value is automatically executed in the environment, representing a pilot's actual execution.



**Fig. 6.** Excerpt of the task ontology

### 4.3 The Simulation Environment

The rules presented in the previous section allow us to interpret the reference model autonomously. To use the reference model in a simulation environment, we have to provide a standard way of exchanging data. Since ontology languages and SWRL rules are not standard languages actively usable with programming languages, we have to provide a way of using the interpreter with standard programming languages. For supporting this need, the OWLReady [9] library for supporting the usability with the Python language has been selected. OWL-Ready supports SWRL rules and the reasoner Pellet, which allows the manipulation of the ontologies with Python code. The advantage of using Python for the interaction with the simulation environment is the possibility of using the solution with other simulation environments. Currently, the execution model is tested with X Plane which is a basic simulation environment. At another level, the reference model will be connected and tested with a more sophisticated aircraft simulator used by Bombardier and CAE.

## 5 Conclusion

In this paper, we proposed an interpreter capable of manipulating the knowledge about the reference procedures for piloting an aircraft. The task and domain taxonomies and their specific characteristics for the execution have been presented. Semantic links between those two ontologies make the execution model highly related to the execution environment. The interpreter is presented as a set of SWRL rules capable of manipulating the knowledge in order to execute the reference model autonomously. Different groups of rules allow a complete execution strategy: constraint evaluation, precondition evaluation, and execution. Since the model has to be executed within a simulation environment, we suggested a library for using the model with a well-known programming language.

This work is an intermediary step for a solution that will follow the execution of the procedures of a human pilot within a simulation environment. Future work will focus on the task of monitoring the execution of the normal and abnormal

procedures by a human pilot in a simulation environment. For this purpose, we are seeing a monitoring strategy based on a cognitive agent capable of following the execution and giving recommendations to the pilot.

**Acknowledgments.** We acknowledge the support of CRIAQ, the Natural Sciences and Engineering Research Council of Canada (NSERC), CAE, Bombardier, and BMU.

## References

1. Abburu, S.: A survey on ontology reasoners and comparison. *Int. J. Comput. Appl.* **57**(17) (2012)
2. Chang, M., D'Aniello, G., Gaeta, M., Orciuoli, F., Sampson, D., Simonelli, C.: Building ontology-driven tutoring models for intelligent tutoring systems using data mining. *IEEE Access* **8**, 48151–48162 (2020)
3. Corbett, A., Kauffman, L., Maclare, B., Wagner, A., Jones, E.: A cognitive tutor for genetics problem solving: learning gains and student modeling. *J. Educ. Comput. Res.* **42**(2), 219–239 (2010)
4. Courtemanche, M.A., Tato, A., Nkambou, R.: Ontological reference model for pilot-ing procedures. In: Crossley, S., Popescu, E. (eds.) *ITS 2022. LNCS*, vol. 13284, pp. 95–104. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_9](https://doi.org/10.1007/978-3-031-09680-8_9)
5. Hayashi, Y., Bourdeau, J., Mizoguchi, R.: Structurization of learning/instructional design knowledge for theory-aware authoring systems. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) *ITS 2008. LNCS*, vol. 5091, pp. 573–582. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-69132-7\\_60](https://doi.org/10.1007/978-3-540-69132-7_60)
6. Jafer, S., Chhaya, B., Durak, U., Gerlach, T.: Formal scenario definition language for aviation: aircraft landing case study. In: *AIAA Modeling and Simulation Technologies Conference* (2016)
7. Keller, R.: Building a knowledge graph for the air traffic management community. In: *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 700–704 (2019)
8. Keller, R.M.: Ontologies for aviation data management. In: *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pp. 1–9. IEEE (2016)
9. Lamy, J.B.: *Ontologies with Python*. Apress, Berkeley (2021)
10. Larman, C.: *UML 2 et les design patterns: analyse et conception orientées objet* (2005)
11. MacLarty, I., Langevine, L., Bossche, M.V., Ross, P.: Using SWRL for rule-driven applications (2009). Accessed 9 Feb
12. Mitrovic, A., Koedinger, K.R., Martin, B.: A comparative analysis of cognitive tutoring and constraint-based modeling. In: Brusilovsky, P., Corbett, A., de Rosis, F. (eds.) *UM 2003. LNCS (LNAI)*, vol. 2702, pp. 313–322. Springer, Heidelberg (2003). [https://doi.org/10.1007/3-540-44963-9\\_42](https://doi.org/10.1007/3-540-44963-9_42)
13. Mizoguchi, F.: Prolog based expert system. *N. Gener. Comput.* **1**(1), 99–104 (1983)
14. Mizoguchi, R., Bourdeau, J.: Using ontological engineering to overcome common AI-ED problems. *J. Artif. Intell. Educ.* **11**, 107–121 (2000)
15. Mizoguchi, R., Vanwelkenhuysen, J., Ikeda, M.: Task ontology for reuse of problem solving knowledge. In: *Towards Very Large Knowledge Bases*, pp. 46–59. IOS Press (1995)
16. Oltramari, A., Lebiere, C.: Mechanisms meet content: integrating cognitive architectures and ontologies. In: *2011 AAAI Fall Symposium Series* (2011)

17. Sheng, Y., Chen, X., Mo, H., Chen, X., Zhang, Y.: An ontology for decision-making support in air traffic management. In: Liang, Q., Wang, W., Mu, J., Liu, X., Na, Z., Chen, B. (eds.) Artificial Intelligence in China. LNEE, vol. 572, pp. 458–466. Springer, Singapore (2020). [https://doi.org/10.1007/978-981-15-0187-6\\_55](https://doi.org/10.1007/978-981-15-0187-6_55)
18. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: a practical OWL-DL reasoner. *J. Web Semant.* **5**(2), 51–53 (2007)
19. W3C: SWRL: a semantic web rule language combining owl and RuleML (2004)
20. W3C: OWL 2 web ontology language: structural specification and functional-style syntax (2nd edn) (2012). <https://www.w3.org/TR/owl2-overview/>



# Method and Tools to Supporting Math Learning in Inclusive Education of Blind Students

Michał Maćkowski<sup>1</sup> , Mateusz Kawulok<sup>1</sup> , Piotr Brzoza<sup>1</sup> , and Dominik Spinczyk<sup>2</sup>

<sup>1</sup> Department of Distributed Systems and Informatic Devices, Silesian University of Technology, Gliwice, Poland  
michal.mackowski@polsl.pl

<sup>2</sup> Department of Medical Informatics and Artificial Intelligence, Silesian University of Technology, Zabrze, Poland

**Abstract.** This study presents a math tutoring platform designed to teach sighted and blind students using audio-tactile graphics in an inclusive educational setting. The platform<sup>4</sup> provides alternative means for presenting information typically included as images in mathematical exercises via tablets and an audio-touch interface. The research used a prepared set of 40 exercises on the properties of geometric figures, and it covered 5 inclusive classrooms at the primary school level during one semester of the school year. The proposed evaluation method of the platform considered several factors: students' learning progress (stored using the knowledge vector and error vector mechanisms), teacher-student and student-student communication, a detailed assessment of the student's way of solving the exercise via a module used by the teacher, and the possibility of group work. The results indicate that the developed method and accompanying tools support group work and increase the independence of blind students in the process of learning mathematics. The platform also increases communication with and detailed ongoing assessment of the student's progress by the teacher.

**Keywords:** inclusive education · math tutoring platform · collaborative math learning · alternative method of audio-tactile presentation · augmented learning strategies

## 1 Introduction

Education is supposed to prepare students for the challenges of the modern world. Inclusive education can achieve this goal. It can be defined as a teaching model whereby all students, regardless of their ability, learn together in one environment [1]. Inclusive education creates a space to ensure the participation of every child and student in the learning process [2, 3]. Inclusive education does not assume the preparation of a separate education path, but aims to meet individual needs through curricular and extracurricular group activities [3, 4]. The teacher uses several methods to teach the same content [5, 6], allowing each student can find the optimal way of learning and fully participate in classes.

The popularity of inclusive education has grown in recent years. Most children with blindness, and especially low vision children, now start their education in mainstream schools [4]. This trend has been confirmed by the European Agency Statistics on Inclusive Education (EASIE) [7]. The statistics tracked the placement of children/learners with recognized special educational needs (SEN) in inclusive or special settings for almost all European countries over the last decade. They showed that the number of students with SEN educated in mainstream (pre-)schools as compared to special (pre-)schools has been growing year by year [7].

Despite the mentioned advantages, inclusive education encounters many problems and obstacles. A systematic review on inclusive education of students with visual impairment focused on two aspects: perceptions of general education teachers, and challenges faced by students with blindness [1]. Research shows that teachers and students need effective teaching-learning tools and access to learning materials [5, 8]. The teachers must consider the barriers faced by blind students when designing instructional methods for them. In the context of classroom activities, students with blindness may encounter difficulties related to school activities [2].

These challenges inspired us to develop a tool and method to support inclusive mathematics education for people with blindness while learning science subjects (our study focuses on learning mathematics). The method assists both the teacher and student while simultaneously supporting the teaching process.

This paper presents our math tutoring platform designed to teach sighted and blind students in inclusive education using audio-tactile graphics. The platform uses augmented learning strategies to adapt the presentation of math exercises to meet the needs and overcome the limitations of students with blindness. It allows students with blindness to solve the same exercises as those solved by sighted students, while remotely supervised by a teacher. Educators using the platform can identify students' difficulties by monitoring their progress while solving exercises, as well as track how students with blindness explore images.

## 2 Related Works

In recent years, there has been rapid development of tools such as educational software, e-learning platforms, and assistive technologies. The latter include screen readers and magnifiers, which allow everyone to use digital materials. The basic source of information in the humanities is text, which can be easily scanned, recognized, and read using speech synthesizers.

People who are blind process cognitive information in the same way as sighted students. However, they must assemble information through haptic or auditory channels, which as other researchers have noted [9, 10], can be a disadvantage. Although the modes of information collection are clearly different between blind and sighted students, their cognitive structures and operation process are the same [11]. It is necessary to prepare materials, media, or techniques that help all students accomplish learning goals. Teachers should provide various assessment methods, materials on how to teach students with blindness, and other aids to ensure equal access to the learning process for all students [5, 12].

Despite the rapid development of information technologies, inequality persists in the access to information for sighted students and those who are blind. This disparity is especially apparent in the case of multimedia platforms or technical data including specific symbols, diagrams, or other formats inaccessible to blind people. In other words, extant technologies in many ways are not capable of automatically transforming such data into forms accessible to those who are blind [13]. Mathematics and math science face this problem. Both presenting such information to blind students and gauging their understanding of the material remains challenging.

Images play a significant role in science subjects, as the presentation of information is mainly visual and contains many formulas, charts, and diagrams. Thus, two questions present themselves: how to present such materials to people who are blind, and how to assess their learning progress [14, 15]? A literature review identified many studies that described alternative methods of presenting graphical information to blind students, especially in interactive audio-tactile form [16–18]. These papers presented several methods for identifying elements of a picture that a blind user selects by touch to provide a description.

Several tools have been developed to provide alternative presentations of mathematical formulas and expressions for blind students. For example, one approach is to automatically convert a mathematical expression into an audio description or mathematical Braille notation describing its structure [19, 20]. Methods for automatic conversion of expressions using, e.g., MathML, are known [21]. Other researchers suggested ways of implementing math notation in electronic form using special tools such as Lambda editor, InftyEditor, and Duxbury Braille Translator [22, 23]. Creation of new mathematical notations are based on extensions to already established software products such as BlindMoose and Math Type for Microsoft Word, odt2braille for Open/Libre Office Writer, and the MathML extension of the DAISY format.

Another group of solutions uses alternative means to present graphics used in science education. These include touch tablets with dedicated software such as Talking Tactile Tablet produced by Touch Graphics and the IVEO by ViewPlus2 [24, 25]. Other researchers showed how to use popular tablets with a large diagonal together with proprietary software for interactive audio and touch presentation of graphic information [26]. Moreover, in paper [27] User-Centered Design approach was used to develop an effective solution for visually impaired students. In many such solutions, mathematical objects such as formulas and images are supplemented with text or audio descriptions, which is a form of Augmented Reality (AR) [19].

The aforementioned solutions allow teachers to share tailored learning materials with students during lessons, but they fail to assess the student's work progress when solving exercises. The teacher must spend extra time conversing with the student to monitor their progress or identify problems the student is experiencing. Such solutions work best in special schools or in integration classes where the main teacher is supported by an assistant in accessible education. Such an assistant should have experience preparing adapted educational materials (tactile images, alternative descriptions, etc.) and the use of Braille notebooks.

However, the foregoing solutions are ineffective and often insufficient for inclusive education. Students in the classroom should be treated equally and must not feel

disrespected. We developed a multimedia platform enabling joint learning of sighted and blind students in mainstream mathematics classes as a response to the previously identified problems in inclusive education [2–4, 6, 10, 28].

Our contribution is to develop a proposed method and provide tools supporting math learning in inclusive education of blind students. The platform includes a mechanism for indirect assessment of student skills by combining the vector of knowledge metric with logged information about mistakes made while solving exercises. Next, we proposed audio-tactile descriptions with different levels of detail to manage the increasing number of terms occurring in audio description (cognitive load problem).

### 3 Materials and Methods

#### 3.1 Proposed Solution

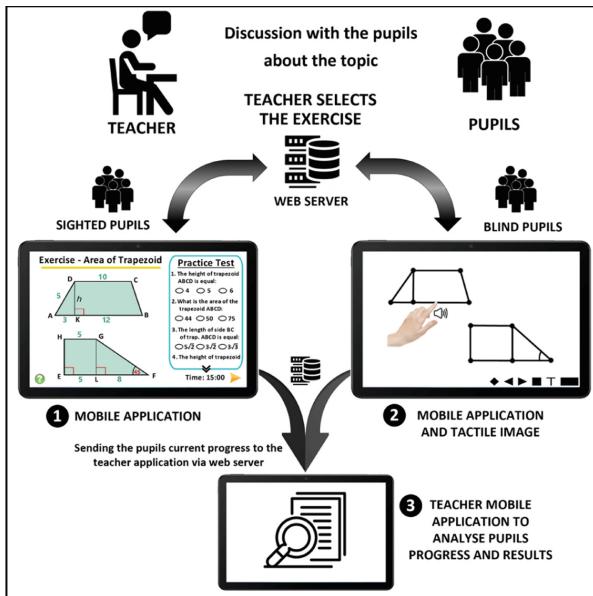
The developed platform uses a client-server architecture. The server application stores and distributes educational materials and mediates communication between the teacher and the student during class. The server application also allows the teacher to monitor the student's work while solving exercises. The platform is equipped with a web application responsible for managing the learning process (learning management system) and preparing educational materials in the form of SVG audio-tactile images. A detailed description of the platform are presented in authors' prior papers [29, 30]. In the previous study we demonstrated improvements in learning effectiveness using the developed platform [31]. Using a dedicated client application on a mobile tablet, the student could solve exercises using audio-tactile images and solve tests whose results were sent to the server. Using a web application, the teacher could view the list of exercises solved by the student along with completed tests—the principle of operation was based on the asynchronous mode of work between the teacher and the student.

Prior to its use in integration classes, the system was quantitatively and qualitatively evaluated for functionality and usefulness in special schools for blind students. The system's impact on learning efficiency in comparison with traditional Braille books and relief drawings was studied. The evaluation was carried out on 2 groups of students: 20 from primary school and 24 from high school. The number of mistakes made while solving exercises with and without the system were recorded. The exercises concerned the properties of geometric figures and the properties of functions based on their graphs. In both cases, the number of errors made and the time needed to complete the exercises decreased by a statistically significant amount. Details can be found in an earlier publication [29, 31].

The present study concerns an extension of the existing platform with mobile client modules. These are mobile applications dedicated to both sighted and blind students as well as teachers. The teacher's application allows for ongoing (synchronous) observation of student progress while solving exercises. This feature makes it possible to use the system in inclusive education.

Figure 1 shows a schematic for solving the same exercise by both sighted students and those students who are blind in our developed solution.

First, the teacher selects an exercise from a previously prepared set. This exercise is displayed on tablets (phones) of sighted students via a graphical user interface. The



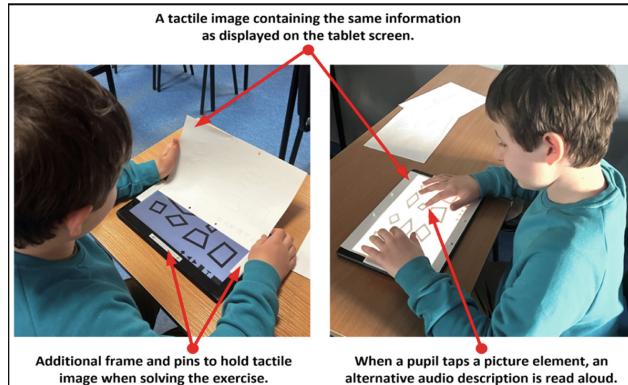
**Fig. 1.** Use of the developed platform in the inclusive education: 1) a sighted student's mobile application; 2) blind student's mobile application; 3) teacher's mobile application.

student sees the test text and image(s) with the necessary information to solve the exercise (e.g., designations of the vertices of the figure, lengths of the sides, angles, etc.). The mobile application for a student who is blind has an audio-touch interface. At the bottom of the screen there is a 6-button interface which allows the blind student to control the application. The student also places a tactile image on the tablet which conveys the same information as the content displayed on the tablet. This allows a blind person to explore the tactile image via touch as well as listen to alternative audio descriptions of the selected image elements by tapping them. The mobile application for a blind student has two operating modes.

- Image exploration – A student who is blind can get acquainted with the image. The question module is used to provide information directly related to understanding the layout of the image, and not to the exercise itself.
- Solving an exercise – A student with blindness solves an exercise by completing a test using an audio-tactile interface.

Figure 2 shows a student using the mobile app. A tactile image is placed on the tablet screen (immobilized by an applied frame). The student, upon sensing an element of the image under his finger, can tap it to prompt the mobile application to read the appropriate alternative description. A sheet with a tactile image is not a barrier for the tablet to detect a finger tap on the image.

The mobile application provides three types of gestures (1, 2, and 3 taps), which allow students to listen to alternative descriptions with different levels of detail. For



**Fig. 2.** The real use of the system during tests by a blind pupil.

example, selecting the lower base of the trapezoid marked in Fig. 1 (hand icon) prompts the following alternative descriptions:

- 1-tap: “Trapezoid ABCD with height h”;
- 2-taps: “Section KB which is part of the base of the trapezoid ABCD”;
- 3-taps: “Section KB equals 12”.

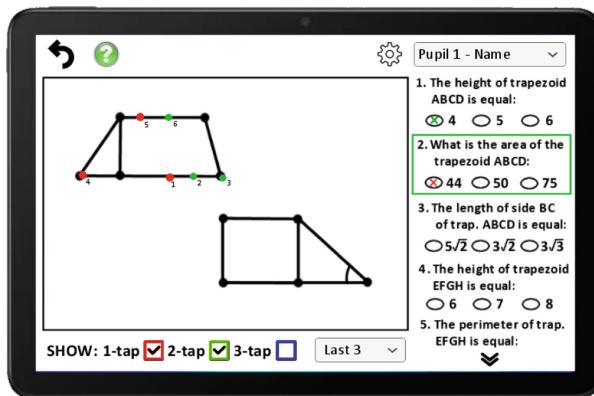
The gestures used are simple and repeatable. They were selected in consultation with blind users of the touch interface. In addition, the mobile application allows a user to solve tests using an audio-tactile interface. The student hears the questions and possible answers and selects an answer using the touch interface button. The interface is simplified and provides the same functionality as the GUI in the application for sighted students. There is also a button in the application that allows the student to request individual help from the teacher.

The last client module is an application for teachers to view the current phase of the exercise being solved by a particular student. The teacher can select a student from the list provided and see his progress. The teacher can also observe the tactile exploration of the image by a blind student. Colored markers show the last gestures made by the student (1, 2, and 3 taps) at each stage of solving the exercise. If necessary, the teacher can view previous gestures made by the student. Figure 3 shows an example screen from the teacher’s mobile application. The red and green dots with numbers show which elements of the tactile picture were tapped, the order of selection, and the number of taps.

The web application and mobile application dedicated to the teacher included also:

#### **Analysis of the results of STUDENT’s answers (a module for a detailed personalized assessment of the student)**

This mode is used both when exploring the image and when solving an exercise. The teacher sees the answers, their order, and the time at which they were entered. Mistakes made by the student are classified. These errors are aggregated in an error vector [32].



**Fig. 3.** Teacher's mobile application—the teacher can observe a student's progress as they solve the exercise. For a selected question in the test, the teacher can observe which elements of the picture were touched by the student with different gestures.

The completeness of the material mastered by the student is represented in the knowledge vector [33], which is also used to select the next exercise.

#### Analysis of interaction with the graphic image while solving the exercise (taps maps—static and dynamic with timestamps)

In the user interaction assessment mode, the teacher first obtains information about the student's progress in solving the exercise. The system shows an interaction map for the current stage which identifies the number of different gestures, their correctness, and the student's ability to use them. In addition, the tool collects user interaction data in three modes.

- Raw mode: This mode is mainly used to assess motor and gesture skills. The teacher can observe if the student has problems with touching particular elements on the picture (for example due to uncontrol shaking of hands or fingers).
- Gesture recognition mode: This mode maps assigned to individual primitives with preserved chronology. (The order is indicated by numbers, and it is also possible to display relative differences in time between gestures). In the case of a large number of taps, it is possible to control the time interval of the presented gestures.
- Pattern recognition mode: This mode highlights information about the completeness or redundancy of a user's session in relation to a valid interaction pattern.

### 3.2 Research Group

The system was evaluated in 5 inclusive classrooms at the primary school level. There were about 25 students in each class, including 1 to 3 who are blind. The system was used in mathematics at the primary school level for one semester (5 months). In addition, the system was used by 7 teachers experienced in teaching inclusive classes. All student research participants were aged from 10 to 13.

### 3.3 Platform Evaluation

The research used a previously prepared set of 40 exercises on the properties of geometric figures. The exercises were solved by both sighted and blind students, while a teacher supervised progress. Evaluation was based on a questionnaire containing a list of questions which are listed in Table 1. We used the Likert scale, which enables quantitative assessment of qualitative research.

**Table 1.** List of detailed questions of the questionnaire.

No	Questions for teachers from integration classes and special schools for the blind
1	How do you assess the way of communication and cooperation with the student using the developed method?
2	How do you rate the support in evaluating the student's current progress?
3	How do you assess the impact on the student's learning progress?"
4	How do you assess the possibility of joint solving of exercise by sighted and blind students?
<b>Questions for a student who is blind</b>	
5	How do you assess the usefulness of the proposed tools?
6	How do you rate the ergonomics of the audio-tactile interface?
7	How do you assess the way of communication with the teacher when solving exercises (presentation of content, exchange of information, accuracy of the teacher's hints)?
8	How do you assess the possibility of self-learning with the use of the proposed tools?
9	How do you assess the possibility of working together on solving exercises with sighted students?
<b>Questions for a sighted student</b>	
10	How do you assess the usefulness of the proposed tools?
11	How do you assess the possibility of self-learning with the use of the proposed tools?
12	How do you assess the possibility of working together on solving exercises with blind students?

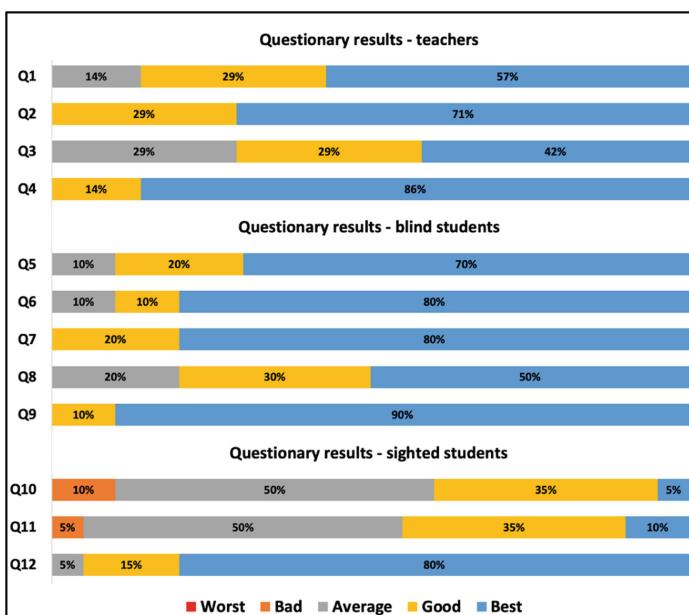
Furthermore, we asked teachers to express their opinion about the tool using open-ended questions about: (1) Overcoming barriers to the integration of blind and sighted people during joint learning; (2) Ease of implementation and integration of the system during learning; and (3) System completeness – strengths and weaknesses.

## 4 Results

Figure 4 presents the results of the questionnaire on a percentage scale. All study groups generally confirmed the usefulness of the developed method and tools. Teachers rated the improvement of their ability to assess student progress most highly. Improvement of

student results was rated relatively low. This result was caused by the limited time for evaluation of the tool (1 semester), which made it impossible to assess the durability of acquired knowledge in the form of skills their application to other areas of life.

Students with blindness rated the ergonomics of the audio-tactile interface and the possibility of cooperation with peers the highest. They rated the possibility of self-learning relatively low. Self-study requires overcoming all the barriers mentioned in the introduction and changing one's attitude. This is the hardest goal to achieve for less skilled students. It was observed that lower values in this category correlate with the test score averages. Sighted students rated the possibility of cooperation with blind peers the highest. They scored the lowest on usability, which may indicate that they perform at a similar level when solving exercises in the traditional way.



**Fig. 4.** Questionnaire research results using Likert scale.

We also collected the most important answers to the open questions posed:

### **Overcoming barriers to the integration of blind and sighted people during joint learning**

Most of the respondents reported a positive impact of the tool. Blind students appreciated the possibility of working independently on the same material after the lesson, along with the possibility of remote consultation with the teacher. Teachers emphasized the usefulness of the work evaluation module as a detailed tool enabling quick identification of challenges students face while solving the exercise. Sighted students noticed the

possibility of learning together and the possibility of consolidating their knowledge by helping students with blindness learn.

### **Easy to implement and integrate the system while learning**

Teachers emphasized the relative ease of implementing the tool on a ready-made set of sample exercises. However, they remarked that preparing new exercises required considerable effort. This result suggests that future work include the development of a wide, generally available set of exercises and the development of tools for semi-automated preparation of new exercises.

### **Completeness of the system – strengths and weaknesses**

Currently, the method requires quite a lot of involvement from the teacher when using the system (e.g., in exercise selection). In the future, more individual reports on the student and the selection of homework tailored to the kinds of errors made could occur automatically by using the knowledge vector and the error vector.

## **5 Discussion and Conclusions**

Thanks to the possibility of simultaneously solving the same exercises by sighted and blind students, they can learn together. The teacher can initiate a class discussion, and the students can work together and solve the presented problems. The ability to work in a group is developed. At the same time, the teacher has tools to assess the student's learning progress, including the knowledge vector, error vector, and various interaction maps.

Although in our opinion the presented solution can be useful in improving math learning and usefulness in inclusive classrooms, the study has some limitations. The biggest disadvantage of our method is the time-consuming preparation of the audio descriptions in various levels of detail, and formulating various tests with answers. Moreover, the presented study is based on the parameters included in the Guidelines and Standards for Tactile Graphics [34]. In our opinion, more research is needed regarding this issue, as we encounter some problems with limited tablet screen size, resolution, and tap detection accuracy. However, we believe these limitations have not influenced the primary outcome of the study.

We hope that the presented math tutoring platform designed to teach sighted and blind students will facilitate inclusive classrooms. The platform allows students with blindness to solve the same exercises as those solved by sighted students, while remotely supervising this process by a teacher in the classroom. The obtained results indicated that the developed platform supports group work and increases the independence of a blind person in the process of learning mathematics. The tools also increase the possibilities of cooperation and integration in the group, as well as the detailed ongoing assessment of the student's progress by the teacher.

## **References**

1. Miyauchi, H.: A systematic review on inclusive education of students with visual impairment. *Educ. Sci.* **10**, 346 (2020)

2. Lintangsari, A.P., Emaliana, I.: Inclusive education services for the blind: values, roles, and challenges of university EFL teachers. *Int. J. Eval. Res. Educ.* **9**, 439–447 (2020)
3. Asamoah, E., Ofori-Dua, K., Cudjoe, E., Abdullah, A., Nyarko, J.A.: Inclusive education: perception of visually impaired students students without disability, and teachers in ghana. *SAGE Open* **8**, 1–11 (2018)
4. Wehmeyer, M.L., Shogren, K.A., Kurth, J.: The state of inclusion with students with intellectual and developmental disabilities in the United States. *J. Policy Pract. Intellect. Disabil.* **18**, 36–43 (2021)
5. Opertti, R., Brady, J.: Developing inclusive teachers from an inclusive curricular perspective. *Prospects* **41**(3), 459–472 (2011)
6. Roos, H.: Inclusion in mathematics education: an ideology, a way of teaching, or both? *Educ. Stud. Math.* **100**, 25–41 (2019)
7. European Agency Statistics on Inclusive Education. <https://www.european-agency.org/activities/data>
8. Oyebanji, M.S., Idiong, U.S.: Challenges of teaching mathematics to students with visual impairment. *Malikussaleh J. Math. Learn. (MJML)* **4**(1), 1 (2021)
9. Lahav, O., Hagab, N., El Kader, S.A., Levy, S.T., Talis, V.: Listen to the models: sonified learning models for people who are blind. *Comput. Educ.* **127**, 141–153 (2018)
10. Bishop, V.E., Barraga, N.C.: *Teaching Visually Impaired Children*. Charles C Thomas Publisher (2004)
11. Amedi, A., Raz, N., Pianka, P., Malach, R., Zohary, E.: Early “visual” cortex activation correlates with superior verbal memory performance in the blind. *Nat. Neurosci.* **6**, 758–766 (2003)
12. Klingenberg, O.G., Holkesvik, A.H., Augestad, L.B.: Research evidence for mathematics education for students with visual impairment: a systematic review. *Cogent Educ.* **6**, 1626322 (2019)
13. Regec, V.: Mathematics in inclusive education of blind students in secondary schools in the czech republic. *Procedia-Soc. Behav. Sci.* **174**, 3933–3939 (2015)
14. Moreno-Guerrero, A.J., Aznar-Díaz, I., Cáceres-Reche, P., Alonso-García, S.: E-learning in the teaching of mathematics: an educational experience in adult high school. *Mathematics* **8**, 840 (2020)
15. De Prado, J.E.L., Gago, A.R.A., López, M.E.: Theoretical review of the teaching of mathematics in student with visual functional diversity. In: ACM International Conference Proceeding Series, 201–208 (2019)
16. Janu, A., Retnawati, H.: Developing audio tactile for visually impaired students. *Int. J. New Trends Educ. Their Implic.* **6**, 18–30 (2015)
17. Papadopoulos, K., Barouti, M., Charitakis, K.: A university indoors audio-tactile mobility aid for individuals with blindness BT-computers helping people with special needs. In: Presented at the (2014)
18. Griffin, E., Picinali, L., Scase, M.: The effectiveness of an interactive audio-tactile map for the process of cognitive mapping and recall among people with visual impairments. *Brain Behav.* **10** (2020)
19. Mikułowski, D., Brzostek-Pawlowska, J.: Multi-sensual augmented reality in interactive accessible math tutoring system for flipped classroom. *Intell. Tutoring Syst.* **12149**, 1 (2020)
20. Bier, A., Sroczyński, Z.: Rule based intelligent system verbalizing mathematical notation. *Multimed. Tools Appl.* **78**, 28089–28110 (2019)
21. Salamoneczyk, A., Brzostek-Pawlowska, J.: Translation of MathML formulas to polish text, example applications in teaching the blind. In: 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF), pp. 240–244 (2015)

22. Fichten, C.S., Asuncion, J.V., Barile, M., Ferraro, V., Wolforth, J.: Accessibility of e-learning and computer and information technologies for students with visual impairments in postsecondary education. *J. Vis. Impair. Blind.* **103**, 543–557 (2009)
23. Maćkowski, M., Brzoza, P., Żabka, M., Spinczyk, D.: Multimedia platform for mathematics' interactive learning accessible to blind people. *Multimed. Tools Appl.* **77**, 6191–6208 (2018)
24. Gardner, J.A., Bulatov, V.: Scientific diagrams made easy with IVEO<sup>TM</sup>. In: Lecture Notes in Computer Science (including Subser. Lecture Notes in Artificial Intelligence Lecture Notes Bioinformatics). 4061 LNCS, 1243–1250 (2006)
25. Minhat, M., Abdullah, N.L., Idrus, R., Keikhosrokiani, P.: TacTalk: talking tactile map for the visually impaired. In: ICIT 2017 - 8th International Conference Information Technology Proceedings, 475–481 (2017)
26. Melfi, G., Müller, K., Schwarz, T., Jaworek, G., Stiefelhagen, R.: Understanding what you feel: a mobile audio-tactile system for graphics used at schools with students with visual impairment. In: Conference on Human Factors in Computing Systems - Proceedings (2020)
27. Shoaib, M., Khan, S., Fitzpatrick, D., Pitt, I.: A mobile e-learning application for enhancement of basic mathematical skills in visually impaired children. Universal Access Information Society 1–11 (2023)
28. Klingenberg, O.G., Holkesvik, A.H., Augestad, L.B.: Digital learning in mathematics for students with severe visual impairment: a systematic review. *Br. J. Vis. Impair.* **38**, 38–57 (2020)
29. Maćkowski, M., Brzoza, P., Kawulok, M., Knura, T.: Mobile e-learning Platform for Audio-Tactile Graphics Presentation. In: Miesenberger, K., Kouroupetroglo, G., Mavrou, K., Manduchi, R., Covarrubias Rodriguez, M., Penáz, P. (eds.) Computers Helping People with Special Needs, pp. 82–91. Springer International Publishing, Cham (2022)
30. Maćkowski, M., Brzoza, P., Kawulok, M., Meisel, R., Spinczyk, D.: Multimodal presentation of interactive audio-tactile graphics supporting the perception of visual information by blind people. *ACM Transactions on Multimedia Computing, Communications and Applications* (2022)
31. Maćkowski, M., Brzoza, P.: Accessible tutoring platform using audio-tactile graphics adapted for visually impaired people. *Sensors* **22**(22), 8753 (2022)
32. Spinczyk, D., Maćkowski, M., Kempa, W., Rojewska, K.: Factors influencing the process of learning mathematics among visually impaired and blind people. *Comput. Biol. Med.* **104**, 1–9 (2019)
33. Maćkowski, M., Żabka, M., Kempa, W., Rojewska, K., Spinczyk, D.: Computer aided math learning as a tool to assess and increase motivation in learning math by visually impaired students. *Disabil. Rehabil. Assist. Technol.* **17**, 559–569 (2020)
34. Guidelines and Standards for Tactile Graphics. <http://www.brailleauthority.org/tg/>. Last accessed 06 Nov 2022



# Understanding the Usages and Effects of a Recommendation System in a Non-formal Learning Context

Anaëlle Badier<sup>(✉)</sup>, Mathieu Lefort, and Marie Lefevre

Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, 69622 Villeurbanne, France  
[{anaelle.badier,mathieu.lefort,marie.lefevre}@liris.cnrs.fr](mailto:{anaelle.badier,mathieu.lefort,marie.lefevre}@liris.cnrs.fr)

**Abstract.** Learning support faces many challenges in a non-formal context: the pedagogical recommendations delivered aimed at learners with very different objectives and studying methods, and the analyses of learners' behavior toward the recommendations are based almost exclusively on learning traces from the learning platform. In this context, we propose a recommendation system for educational resources in an extracurricular mobile application. Our system is based on a pedagogical score, on the learner history and ensures a renewal of the recommended chapters. By analyzing the usage traces of more than 10,000 learners over 4 months, we show that delivered recommendations are mainly used as a revision tool when learners face difficulties, and that they mostly remain in chapters of their current level grade. Usage traces also reveal that following recommendations has a positive impact on the learning experience.

**Keywords:** Pedagogical Recommendations · Traces Analysis · Usage and Practice Analysis · Self-Regulated Learning

Recommendation Systems in Education dedicated to learners aim to deliver personalized learning content to adapt to learner's needs and competencies. We are working with an after-school e-learning mobile application. This application gathers millions of students from all grade levels. The learners are currently totally autonomous in their working experience, and use the application as a micro-learning tool: they work not regularly, with very small working sessions. We want to study if guiding learner with personalized recommendations to continue working with the application would improve their studying experience. To do so, we implemented a recommendation system for Mathematics, that have been pedagogically validated by experts and which is used by the students from middle school and high school grade levels [1]. In this article, we focus on the learning experience with this system, toward the following research questions:

**Q1- How do learners select which of the recommended resources to follow?**

**Q2- How does our recommendation system change the learning experience?**

After having briefly recalled the functioning of the existing system and positioned it in the scientific context (Sect. 1), we present the experimentation protocol and the collect of usage data (Sect. 2). We then analyze according to several criteria the recommendations followed by the learners among those proposed (Sect. 3) and end our study on the changes observed in the learning experience (Sect. 4). Finally, we discuss these results in Sect. 5 and present the resulting perspectives in Sect. 6.

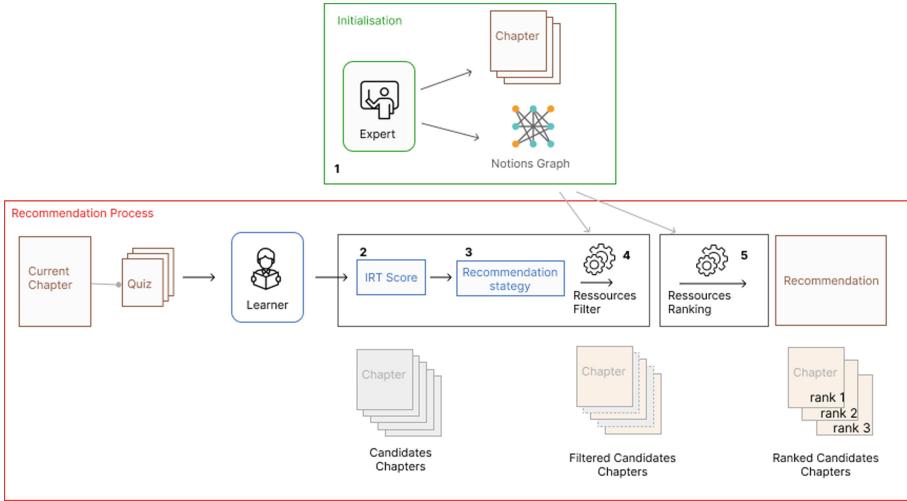
## 1 Delivering Recommendations in a Self-regulated Learning Context

We are working with an educational mobile application organized into disciplines (Mathematics, Biology...), divided into chapters (Arithmetic, Functions...). Each chapter contains small courses, 4 quizzes of 5 multi-choices questions each and revision sheets. We propose a recommendation system to guide learners from chapter to chapter. Considering a student currently working on a given chapter (*input chapter*), the system recommends among all the available chapters of the same discipline in the application (*candidates chapters*) a small list of chapters (*recommended chapters*) to continue working on. According to Brusilovsky's classification of educational systems [6], our system is an Adaptive Navigation Support System. The recommendation system workflow is presented on Fig. 1. To deliver recommendations and guide students in their use of the application, the system is based on 3 modules: a **Pedagogical Module**, an **Historic Module** and a **Novelty Module**.

### 1.1 Pedagogical Selection and Ranking of Learning Resources

The **Pedagogical Module**, described in [1], selects a recommendation strategy (step 3 on Fig. 1), filters the *candidates chapters* according to this strategy (step 4) and compute a pedagogical relevancy score to rank the filtered chapters and recommend the best (step 5). We organized our learning items (*candidate chapters*) in a notions graph (step 1). Other works rely on competency-based approaches as in the ComPer project [16], or use ontology models [10]. In our system, we only recommend application internal chapters, that are labelled with key notions by the pedagogical experts that created the content. "We call notion a piece of knowledge useful to understand the current chapter. A notion is labelled as prerequisite if the learner must already understand part of the concept described by the notion to master the current chapter. The notion is expected if the chapter strategy is either to discover this notion, or to go further with more difficult questions on this notion." [1]. With this notions-labelling system, the chapters are linked to each others inside a notions graph. Two chapters are closer as they share the same specific notions.

We defined 3 recommendations strategies: revision, continuation and deepening, that are assigned to learners depending on they mastery of the current chapter. Other recommendation systems are based on the Zone of Proximal



**Fig. 1.** Workflow of the recommendation system. Based on the notions graph established by experts and depending on a recommendation strategy computed from learner results, we filter and rank the candidates chapters to deliver recommendations. Steps with numbers are explained in the next sections.

Development (ZPD) concept [21] to assign learning resources to students based on their difficulty level and on the learner estimated mastery [3]. Others use Knowledge Tracing Models (KTM) to model the level of knowledge of a student and recommend suitable content. Among these KTM we can cite the Corbett and Anderson's Bayesian Knowledge Tracing Model (BKT) [9] or other KTM systems as the one developed by Vie and Kashima to predict learner's performance in MOOCs [19]. In our case, we estimate the learner's level of mastery for a given chapter depending on the quizzes done by the student in this chapter. We use the Item Response Theory (IRT) [2] to compute  $\theta$  which represents the level of mastery on the quizzes and by extension, on the input chapter for a given learner (step 2 on Fig. 1). We define 3 level groups depending on the  $\theta$  score and assign a recommendation strategy for each group (step 3 on Fig. 1). The recommendation strategies are defined and computed for each learner for each chapter. Then, we filter the available candidate chapters for the recommendation depending on this selected strategy (step 4). Finally, we compute a pedagogical relevancy score (Eq. 1) for each candidate chapter, based on the shared notions ( $score_{similarity}$ ) and the grade level distance of the candidate chapters from the current working chapter ( $penalty_{distance}$ ). We rank chapters mainly depending on this computed pedagogical score. Filtering and computation details are explained in the previous article [1]. The resulting recommendations have been validated by experts and used by all the students.

$$score_{pedagogical} = score_{similarity} * (1 - penalty_{distance}) \quad (1)$$

## 1.2 Delivering Recommendations Based on Learner's History

As we recommend learning content inside an extracurricular mobile application, the only knowledge we have about learners are the learning items explored in the application, especially the quizzes done and their results. We do not have any information about the chapters already or currently studied in class. However, to maintain consistency in the learning experience on the application, we do not want to recommend recently already studied and mastered learning resources. Students may find the system irrelevant if it recommends a content studied a couple hours ago. For this reason, we included an **Historical Module**. This historical indicator is made up of 2 components: the completion of candidate chapters and the grade obtained on these chapters. While some recommendation systems decide to not recommend at all already studied contents [4], we decided to introduce a time parameter to keep this possibility. Indeed, the temporal parameters is an important factor in several recommendations systems: Choffin *et al.* studied the effect of repetition on forgetting and learning [7], Jiang *et al.* used a Long Short-Term Memory model to include the forgetting parameter [13]. The distribution of learning and memory over time, described by Ebbinghaus [11] have been used in various works since [20].

We introduced in this historical module a forgetting factor  $w_t$ , which was computed from the Ebbinghaus forgetting curve adapted to our context of non regular use: in our case, the forgetting curve estimates 70% retention of information after 7 days, reduced to 12% after 30 days. The historical score is computed with Eq. 2:

$$score_{history} = 1 - [w_t * score_{completion} * score_{grade}] \quad (2)$$

with  $w_t \in [0, 1]$ : forgetting factor calculated according to the adapted Ebbinghaus formula,  $score_{completion} \in [0, 1]$  the ratio of ended quizzes in the candidate chapter and  $score_{grade}$  the average grade on the ended quizzes in the candidate chapter.

With this module, we consider the previous activity of the learner on the application, promoting content not yet studied. However, we can recommend content that was already covered but not yet completed or poorly mastered. We can also recommend a chapter already studied a long time ago, assuming that a reminder of the knowledge of this chapter can help to better master the chapter being studied.

## 1.3 Encouraging Curiosity by Varying the Recommendations

Our recommendation system is applied to a self-regulated learning application. We introduced a **Novelty Module**, so that the learner sees the recommendation offer evolve, and thus mitigate a lassitude effect. Indeed, Oudeyer *et al.* demonstrated that curiosity and novelty are important to enhance learning and retention [15]. Serendipity was show to be an important criterion for making appealing and useful recommendations [24].

As with the historical indicator, we introduce a time factor to allow us to recommend again resources that were declined a long time ago. To limit the number of parameters in the model and because we have no previous studies about our learner's forgetting curves, the  $w_t$  is the same as for the historical indicator, which can be discussed. The novelty score is computed with Eq. 3:

$$score_{new} = 1 - [w_t * \frac{1}{n} \sum_{i=1}^n \frac{3 - k_i}{3}] \quad (3)$$

with  $n$  number of times the chapter was recommended and  $k_i \in [0, 3]$  [the resource rank in the top 3 displayed recommendations (0 being the best) the  $i^{th}$  time].

To initialise our model and rank chapters, we combine these 3 modules in a weighted multiplication with manual initialisation of weights, as in [23]. We decided to give more importance to the pedagogical module. This manual setup of the initial weights and the form of this Eq. 4 are discussed in Sect. 5.

$$score_{reco} = 0.5 * score_{pedago} + 0.4 * score_{history} + 0.1 * score_{new} \quad (4)$$

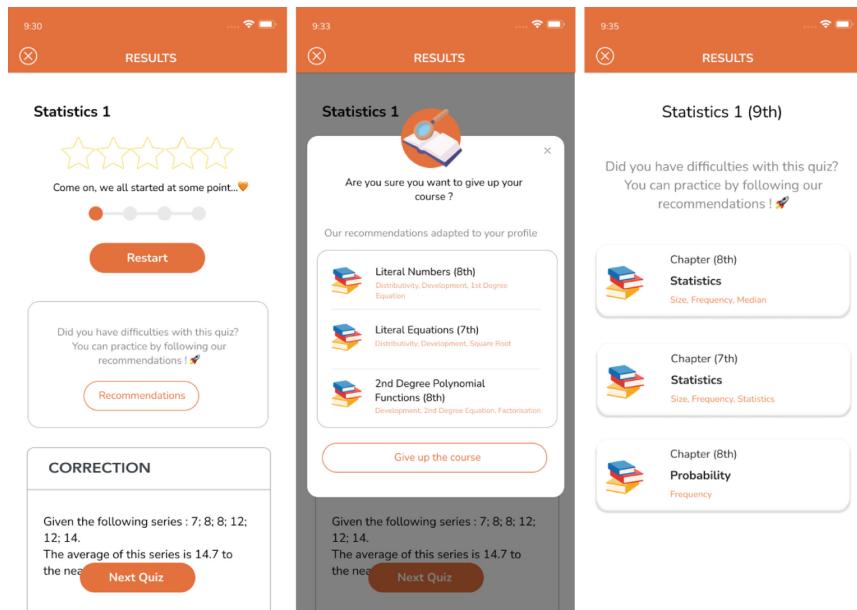
## 2 Evaluation Protocol

From the literature, numerous recommendation systems are used and evaluated during classroom experiments [17, 18, 22] or with virtual simulated students [8] to collect enough data to train learning models, which is not our case. Because of our context and to understand learners needs and behaviours, we evaluate directly the system in the real context of use. The main drawback of this validation "in real-life" is the difficulty to collect data, as highlighted by [12]. This difficulty is even more significant with self-regulated learning platforms because learners can quit their learning session at any time, and never come back.

The recommendation system has been implemented for Mathematics, from 6th to 12th grade. It was available for learners during their first month using the app. The recommendation are displayed from 3 entry points (Fig. 2):

- **Quiz Failed.** If the learner totally failed one of the 4 quizzes in the chapter, a button appears to access directly the revision strategy recommendations.
- **Path Interrupted.** Learners are supposed to end the 4 quizzes in the chapter. If they interrupt their path, a pop-up suggests to continue working with our recommendations, with the strategy computed from the already collected answers.
- **Path Ended.** Recommendations are displayed on the result page, at the end of the 4-quizzes path, depending on the grade obtained on these 4 quizzes.

Qualitative analysis to collect students perceived usefulness of the recommendations and analysis of students re-use of recommendations can be found in [1]. To analyse learner's usage, we implemented an internal system collecting learning traces similar to xAPI used in several works [5, 14]. This tracking system identifies and stores each learner action, e.g we collect the actions linked with



**Fig. 2.** Recommendation interface for the chapter *Statistics* (9th grade) for the revision strategy: on the left from the quiz failed entry point, in the middle for the path interrupted, on the right at the end of the path. Grade levels are displayed in parenthesis. Common notions between chapters are displayed in orange. (Color figure online)

our recommendation system as *Recommendation interface viewed*, *Recommendation chosen*, *Recommendation interface left*, but also more generic activities as *Connection to the app*, *Quiz started*, *Course read*, with associated context information (learner grade level, content discipline, working session identifier, date and time, *etc*). Learners are anonymized to respect data privacy. In the following sections we present the results observed between September 2022 and January 2023.

### 3 Selection of Recommendations by Learners

We collected usage traces from 11,962 learners who encountered a recommendation at least once, representing 39,420 recommendation traces seen, including 5,609 recommendations chosen by 3,448 learners. We analyzed the data according to several criteria.

The first criterion studied is the entry point of the recommendation display. We ran a Chi-2 test that highlights a significant influence of the recommendation interface entry point on the follow-up decision. The results presented on the Table 1 show that recommendations are more followed in the context of a failed quiz (residuals +34.9) or a completed course (residuals +20.0). Recommendations from interrupted courses are rejected more than others.

**Table 1.** Influence of recommendation source on follow-up decision (Pearson's Chi-2 test.  $\chi^2 = 2714.6$ , df = 2,  $p_{value} < 2.2e - 16$ ). A recommendation is *chosen* if the learner clicks on one of the recommended chapters, *rejected* otherwise.

Entry point	Recommendation <i>observed (residuals)</i>	
	Chosen	Rejected
Quiz failed	904 (+34.9)	1202 (-14.2)
Path interrupted	1724 (-26.6)	21041 (+10.8)
Path ended	2981 (+20.0)	11568 (-8.2)

The second criterion is the influence of the recommendation strategy. The results are presented in Table 2. We see that recommendations associated with the revision strategy tend to be followed more than continuity or deepening strategies (residuals +10.9). As the revision strategy is assigned for the *Quiz failed* entry point, this result is consistent with the previous result.

These analyses show that learners preferentially choose recommendations when they recognize gaps in their knowledge.

**Table 2.** Influence of the recommendation strategy on the follow-up decision (Pearson's Chi-2 test,  $\chi^2 = 207.19$ , df = 2,  $p_{value} < 2.2e - 16$ ).

Strategy	Recommendation <i>observed (residuals)</i>	
	Chosen	Rejected
Revision	1682 (+10.9)	7384 (-4.4)
Continuity	1815 (-7.6)	13432 (+3.1)
Deepening	2112 (-0.8)	12996 (+0.3)

As the grade level of recommended chapters are displayed on the interface, we analysed the influence of this grade level factor on learner's choice. We observe on Table 3 that inside each strategy, chapters of equal grades are more chosen than chapters of different grades regarding the recommended content. For example in revision strategy, even if 73.8% of recommended chapters are of lower grade, only 68.5% of chosen chapters are of lower grade, while for 26.2% of equal grade chapters recommended, 31.5% of chosen chapters are of equal grades. The same tendency is observed for the continuation and deepening strategies.

From these analysis we see that learners select recommendations when they failed a quiz or end a chapter especially when poorly mastered (the revision strategy is more followed than others). Among the different chapters recommended, they tend to prefer chapter from their current grade level.

**Table 3.** Distribution of chosen and recommended chapters among the recommendations followed, depending on the strategy. CG = Current Grade. (Unknown grades refer to technical issues.)

Strategy	Chapters	
	Chosen	Recommended
<b>Revision</b>		
Lower grade ( $<CG$ )	1152 (68.5%)	3577 (73.8%)
Equal grade (CG)	529 (31.5%)	1272 (26.2%)
<i>Unknown grade</i>	1	
<b>Continuation</b>		
Lower grade ( $CG - 1$ )	929 (51.2%)	2938 (56.9%)
Equal grade (CG)	884 (48.8%)	2223 (43.1%)
<i>Unknown grade</i>	2	
<b>Deepening</b>		
Equal grade (CG)	956 (45.3%)	2489 (40.3%)
Higher grade ( $>CG$ )	1156 (54.7%)	3684 (59.7%)

## 4 Effects of Recommendations on the Learning Experience

After identifying some criteria for selecting recommendations, we investigate the activities performed within the chosen recommended chapters. To do this, we studied the 5,609 traces of the 3,448 students who have chosen a recommendation at least once and observe the activities performed within the chapter they chose. The results are presented in Table 4. Within each chapter, the learner can study 2 different resources: the quizzes and the courses. We see that 32.3% of the recommendations chosen did not lead to any activity in the corresponding chapter: either because the recommendation did not correspond to the learners' expectations, because they followed the recommendations only out of curiosity or because we remain in a micro-learning context with limited time dedicated to

**Table 4.** Activities completed within chapters for recommended and followed resources. Inside a chosen chapter, a learner can start quizzes (ended or not) and/or read courses (*total*  $\geq 100\%$ ).

Activity in the chosen chapter	Observed recommendations N = 5609
No activity	1813 (32.3%)
At least 1 course read	1668 (29.7%)
At least 1 quiz started (ended or not)	3061 (54.6%)
At least 1 quiz started and ended	2027 (36.1%)

study. Recommendations that were followed without any activity in the chapter consulted could be described as “visit only”. In 67.7% of cases, the recommendation followed led to some activity: learners started at least 1 quiz and/or read at least 1 course (partial follow-up) and in 29.7% of cases they completed at least one quiz in the chapter followed (in-depth follow-up).

Beyond the activity done in a chosen recommendation chapter, we question the impact of following recommendations in the learning experience. In our context of extracurricular use, we are not looking for learning gains but rather for changes in working behavior with the application. To qualify the learning experience, we use the indicators of *number of working sessions*, *duration of working sessions*, *number of quizzes started*, *number of quizzes ended*, *number of courses read*, *average time to read courses*. A working session is defined as all the activities recorded in the application between the opening and closing of the application. A learner can have several working sessions within the same day.

We had at our disposal 7,693 full-session data from 1,126 learners who followed recommendations at least once, and started at least one math quiz, to compare the learning experience between sessions with and without following recommendations. Using a paired Student’s t test on each of the indicators, we investigate the impact of following recommendations on the learning experience within the application. The results are presented in the Table 5.

**Table 5.** Student’s test comparison of learning indicators between sessions with and without recommendation following. (\*\**pvalue* < 0.001, \**pvalue* < 0.05)

Indicators by member	Sessions with follow-up N = 1471	Sessions without follow-up N = 6222	pvalue
Number of sessions	1.31	5.53 ***	<2.2e - 16
Average session time (min)	11.06 ***	6.72	<2.2e - 16
Average number of quizzes started per session	3.97 ***	2.28	<2.2e - 16
Average number of quizzes ended per session	3.27 ***	1.85	<2.2e - 16
Average number of courses read per session	3.01 ***	1.67	<2.2e - 16
Average time spent reading courses per session (sec)	49.4 *	37.9	0.02

We find that sessions with one or more recommendation chosen are more active working sessions. Learners work almost twice as long, even though we remain in a micro-learning context. They also start more quizzes, and read more courses. Two explanations can be made from this analysis: the recommendations encourage the learner to work more, or when the learner decides to spend more time studying with the application, they will tend to follow more recommendations. In both cases, our system has a positive impact by offering a new learning experience in the application.

## 5 Discussion

### 5.1 System Design

The recommendation score described in Eq. 4 was designed to consider the influence of the 3 modules. We chose to initialise the weights with arbitrary values to prioritize the pedagogical score, then the historical score and finally the novelty score. We could evaluate precisely the impact of each module in the learner's choice with ablation studies. Similarly, the choice of modeling time with the parameter  $w_t$  can be discussed. Further studies could vary this parameter to study its relevance in a mobile application learning context. In an extracurricular learning context with very little regularity of use, we have chosen to first conduct an exploratory approach to study the suitability of such a system to our context. Our priority was to collect usage data to understand learners' behaviors before studying deeper the model parameters.

### 5.2 Influence of Recommendations on the Learning Experience

We have chosen to analyze learner behavior through the prism of certain criteria and indicators, because of the totally free context of this form of learning, several other criteria could have been relevant to continue the analysis.

In our context of extracurricular use of the recommendation system, it is difficult to understand the motivations that drive learners to use the app and our recommendations. Recommendations seem to be appropriate in some contexts (failed quiz, revision strategy) but less so in others (interrupted course, deepening strategy). Moreover, the recommendation system was designed to overcome the grade level barrier: even if they are in 6th grade, 5th or 7th grade level chapters may be relevant. However, our study shows that learners tend to keep studying content of their current grade level. This can be partially explained by our out of school mobile-app learning context: learners want to work quickly with something they currently study in class.

When we look at the influence of recommendation following on the learning experience, the Table 5 shows that the number of courses accessed is higher for sessions with recommendation chosen, but the reading time is only slightly higher. We can hypothesize that the learners skimmed over the recommended content but not really focused on it. Indeed, the short session times are a good indication of the very auxiliary nature of the application, which is used as a support tool and not as the primary source of learning. This "exploratory" behavior is illustrated by the Table 4 with recommendations that are followed but whose relative pedagogical activities are not systematically exploited.

## 6 Conclusion and Future Work

In this paper we studied a recommendation system deployed for a non-formal learning context on a mobile application. The recommendations, built from three

modules (Pedagogical, History and Novelty), are presented to learners at different moments of their learning experience, and are based on 3 strategies: revision, continuity and deepening. We analyzed the learning traces and showed that learners mainly use the system's recommendations in the revision strategy, primarily when they have failed a quiz or finished studying the current chapter. We also highlighted different modalities of recommendation follow-up (visit only, partial or in-depth) and showed that the working sessions with recommendation follow-up were more committed than those without. Indeed, the session time as well as the number of quizzes started, completed and course consultation are significantly higher in the sessions with follow-up.

As this first study shows promising results, we will continue the analysis of the recommendation system applied to other disciplines, for which the system is already implemented but not yet made available to learners. Our first research question focused on the selection of recommendations proposed by learners. We will deepen the analysis with other indicators that can influence this selection, such as the display or not of the grade level in the recommendations. In connection with the first study published in [1], we will examine whether these selection criteria are common to all learners, or vary according to the audience. Our second research question focused on the learning experience regarding our recommendation system. Having observed different levels of follow-up, in our further work we will seek to understand the behavior of two target audiences: those who never consult the proposed recommendations and those who remain only in a “visit-only” mode. This better understanding may allow us to propose more relevant recommendations to them.

## References

1. Badier, A., Lefort, M., Lefevre, M.: Recommendation model for an after-school e-learning mobile application. In: CSEDU 2023, Czech Republic (2023)
2. Baker, F.B.: The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation (2001)
3. Baker, R., Ma, W., Zhao, Y., Wang, S., Ma, Z.: The results of implementing zone of proximal development on learning outcomes. In: The 13th International Conference on Educational Data Mining (2020)
4. Bothe, M., Renz, J., Rohloff, T., Meinel, C.: From MOOCs to micro learning activities. In: 2019 IEEE Global Engineering Education Conference (EDUCON), pp. 280–288. IEEE (2019)
5. Branthôme, M.: Pyrates: a serious game designed to support the transition from block-based to text-based programming. In: Hilliger, I., Muñoz-Merino, P.J., De Laet, T., Ortega-Arranz, A., Farrell, T. (eds.) EC-TEL 2022. LNCS, vol. 13450, pp. 31–44. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-16290-9\\_3](https://doi.org/10.1007/978-3-031-16290-9_3)
6. Brusilovsky, P., Peylo, C.: Adaptive and intelligent web-based educational systems. Int. J. Artif. Intell. Educ. (2003)
7. Choffin, B., Popineau, F., Bourda, Y., Vie, J.J.: DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills. In: Junior Conference on Data Science and Engineering, France (2019)

8. Clément, B., Roy, D., Oudeyer, P.Y., Lopes, M.: Multi-armed bandits for intelligent tutoring systems. *J. Educ. Data Min.* **20**–48 (2015)
9. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. In: *Proceedings of UMUAI*, pp. 253–278 (1994)
10. De Maio, C., Fenza, G., Gaeta, M., Loia, V., Orciuoli, F., Senatore, S.: RSS-based e-learning recommendations exploiting fuzzy FCA for knowledge modeling. *Appl. Soft Comput.* **113**–124 (2012)
11. Ebbinghaus, H.: Memory: a contribution to experimental psychology. *Ann. Neurosci.* **155**–156 (2013)
12. Erdt, M., Fernandez, A., Rensing, C.: Evaluating recommender systems for technology enhanced learning: a quantitative survey. *IEEE Trans. Learn. Technol.* **326**–344 (2015)
13. Jiang, W., Pardos, Z.A., Wei, Q.: Goal-based course recommendation. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 36–45. Association for Computing Machinery (2019)
14. Nikolayeva, I., et al.: How to help teachers adapt to learners? Teachers' perspective on a competency and error-type centered dashboard. In: Pammer-Schindler, V., Pérez-Sanagustín, M., Drachsler, H., Elferink, R., Scheffel, M. (eds.) *EC-TEL 2018*. LNCS, vol. 11082, pp. 596–599. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-98572-5\\_53](https://doi.org/10.1007/978-3-319-98572-5_53)
15. Oudeyer, P.Y., Gottlieb, J., Lopes, M.: Intrinsic motivation, curiosity, and learning. In: *Progress in Brain Research*, vol. 229, pp. 257–284. Elsevier (2016)
16. Pierrot, L., Michel, C., Broisin, J., Guin, N., Lefevre, M., Venant, R.: Promoting self-regulated learning strategies for first-year students through the COMPER service. In: *18th International Conference on Cognition and Exploratory Learning in Digital Age, Virtual*, Portugal (2021)
17. Sablayrolles, L., Lefevre, M., Guin, N., Broisin, J.: Design and evaluation of a competency-based recommendation process. In: Crossley, S., Popescu, E. (eds.) *ITS 2022*. LNCS, vol. 13284, pp. 148–160. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_14](https://doi.org/10.1007/978-3-031-09680-8_14)
18. Venant, R., Teyssie, C., Marquié, D., Vidal, P., Broisin, J.: A Competency-Based Model to Bridge the Gap Between Academic Trainings and Industrial Trades (2015)
19. Vie, J.J., Kashima, H.: Knowledge tracing machines: factorization machines for knowledge tracing. In: *Proceedings of the AAAI Conference on AI* (2019)
20. Vlach, H.A., Sandhofer, C.M.: Distributing learning over time: the spacing effect in children's acquisition and generalization of science concepts. *Child Dev.* **83**, 1137–1144 (2012)
21. Vygotskī, L.S., Cole, M.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press (1978)
22. Yanes, N., Mohamed Mostafa, A., Ezz, M., Naif Almuayqil, S.: A machine learning-based recommender system for improving students learning experiences. *IEEE Access* **8**, 201218–201235 (2020)
23. Zhao, Z., et al.: Recommending what video to watch next: a multitask ranking system. In: *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 43–51. ACM (2019)
24. Ziarani, R.J., Ravanmehr, R.: Serendipity in recommender systems: a systematic literature review. *J. Comput. Sci. Technol.* **36**(2), 375–396 (2021). <https://doi.org/10.1007/s11390-020-0135-9>



# Detecting Mental Fatigue in Intelligent Tutoring Systems

Alyssa Hajj Assaf, Hamdi Ben Abdessalem<sup>(✉)</sup>, and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal,  
Montréal H3C 3J7, Canada

{alyssa.hajj.assaf, hamdi.ben.abdessalem}@umontreal.ca,  
frasson@iro.umontreal.ca

**Abstract.** Mental fatigue happens when a cognitive activity is maintained without the availability of adequate intellectual capabilities. This leads to change in mood, motivation, and temporary deterioration of various cognitive functions involved in goal-directed behavior. In this paper, we present our framework to successfully generate and measure mental fatigue through completion of cognitive tasks in a virtual simulated environment. Participants showed significant decline in pupil diameter and theta/alpha score during the various cognitive tasks. We trained an RBF SVM classifier from Electroencephalogram (EEG) data classifying mental fatigue with 95% accuracy on the test set. We believe that this Intelligent Tutoring System can aid in providing intelligent assistance on one’s mental resources available for learning and productivity.

**Keywords:** Mental fatigue · ITS · Machine Learning · Mental Workload · Task-engagement · Virtual Reality · EEG · Pupil Diameter

## 1 Introduction

Mental fatigue is a complex state that can manifest in mood variation, change in motivation, and a temporary decline of cognitive functions [1, 2]. Mental fatigue can negatively impact workplace performance and cause potential danger for oneself or/and others. It is in fact the most frequent cause of accidents in the workplace [3].

Previous work in the field of mental fatigue generation and detection mostly uses classic laboratory settings where users are asked to engage in a cognitive exercises or simulations on a computer screen. These settings are rarely representative of real-world contexts in which mental fatigue arises and thus challenge the integrity of the data collected. In our work, we intend to generate and measure mental using virtual reality to best reproduce contexts and emotions in which mental fatigue arises. Virtual Reality (VR) session paired with Electroencephalogram (EEG) recording has been used in several experiments to assess emotional response from a virtual simulated environment [4, 5].

The goal of our experiment is to (1) generate and measure mental fatigue using VR session paired with EEG and Eye tracking and (2) develop an Intelligent Tutoring

System (ITS) capable of providing user feedback on 3 different levels of mental fatigue. We believe that an intelligent tutoring system (ITS) can aid in providing intelligent assistance on one's mental resources available for learning and productivity and motivate learners to rest until cognitive capacities are restored. In order to best reproduce settings in which mental fatigue can occur, participants will engage in many cognitive exercises for a prolonged period of time in a virtual environment resembling a closed working office/room.

Kamińska investigated the use of EEG signals to classify a subject's mental stress level using virtual reality environment [6]. Participants were immersed in two alternating VR interactive simulations: stress inducing and relaxing. The stress inducing environment consisted of the Stroop test, while relaxing environment consisted of interactive relaxing scene based on scenarios created for psychotherapy treatment. During the session, brain wave activity was continuously monitored using EEG, and participants were asked to fill a questionnaire to assess their mood and level of stress, before and after the session. The experimenters used a convolutional neural network (CNN) to classify the level of stress of the participants and matched the subjective stress assessment of the participants with 96.42% accuracy [6].

Like many other related studies, Kamińska established their labels based on subjective assessment of fatigue via the questionnaire. However, subjective feedback questionnaires are time-consuming and unreliable for real-time fatigue detection. Ren Ziwu and colleagues [7] developed a Radical Basis Function (RBF) Neural Network to detect fatigue in driving simulation using EEG signals. Instead of using questionnaire, they used eye closure, a well-known fatigue indicator to label fatigue and alert segments. Ren Ziwu et al. achieved 92.71% mean accuracy on their RBF neural network.

This paper presents an exploratory study of generation and detection of mental fatigue. We will first present a short resume of key concepts in mental fatigue followed by our results in terms of the chosen fatigue indicators: pupil diameter, theta/alpha ratio, task engagement index. Finally, we will present the machine learning model obtained to classify mental fatigue.

## 2 Predicting Mental Fatigue

There are some factors and task demand characteristics that might influence whether an individual will become mentally fatigued by a task. For instance, how much time have they spent doing that task? Is the task interesting/engaging this person? How focus do they need to be to perform well in this task?

Several factors such as time on task, workload, task engagement, and fatigue type can influence mental fatigue.

**Time on Task (TOT):** The effect of task duration on mental fatigue and performance is known as the time-on-task effect (TOT). In general, mental fatigue increases as the time spent on a task increases. This relation can be observed through task performance: the more time we've spent on a task, the more fatigue we become and thus perform less at it. However, it should be noted that the relation between TOT and performance is not linear: during the first blocks of a task, an improvement in performance can be observed

as a result of learning or automatization of performance [8]. Nonetheless, this peak in performance is generally followed by a decrease in performance caused by mental fatigue, which makes task duration an important task-related factor for mental fatigue.

**Workload (WL):** The **mental workload** can be defined as the number of cognitive resources or/and energy required to execute a cognitive task, requiring attention, memory, alertness or decision making [9]. In general, working under high levels of mental workload over prolonged periods results in an individual's depletion of cognitive resources and energy and, eventually, mental fatigue [10]. Thus, mental workload is also an important task-related factor in predicting mental fatigue.

**Task Engagement (TE):** The level of attention, involvement, and interest one dedicates to a particular task is one of the many factors affected by mental fatigue. Hence, mental fatigue can result in an unwillingness for further efforts, abandoning behavior, where one becomes disengaged with the current task [2]. Consequently, **task engagement** decreases with TOT effects (mental fatigue).

**Fatigue type (Passive vs. Active Fatigue):** An important theory proposed by Desmond and Hancock's (2001) suggested there are two types of fatigue: active and passive. In driving studies, **active fatigue** is characterized by elevated stress and results from a continuous and prolonged demanding interaction vehicle control requiring constant perceptual and motor adjustments. On the other hand, **passive fatigue** is characterized by task disengagement and is the result of prolonged hours of little to no perceptual-motor response or interaction with vehicle control [11, 12]. Thus, active fatigue appears to occur in higher workload conditions while passive fatigue occurs in lower workload conditions. Research by Saxby and Matthews (2008) later confirmed that the passive fatigue induced by low workload condition resulted in a significantly greater task disengagement over time compared to the active fatigue and control group.

### 3 Measuring Mental Fatigue

#### 3.1 Pupil Diameter in Mental Fatigue and Workload

Pupil diameter (PD) can be used to detect workload and mental fatigue. As mental fatigue increases the pupil diameter decreases with respect to baseline measurements [13]. While pupil diameter is also sensitive to changes in arousal and fatigue, it is also sensitive to changes in mental workload. Gonca Gokce Menekse Dalveren conducted a study to measure changes in mental workload in surgical residents during surgery where participants were subjected to a computer-based simulation of surgical task. This study found that pupil diameter grows in direct proportion with mental workload [14]. Thus, during the **transition** state between high mental workload and mental fatigue, pupil diameter is expected to decrease.

### 3.2 Electroencephalogram (EEG) Measures Fluctuation

The electroencephalogram is a non-invasive way to measure the electrical activity originating from the brain from a set of electrodes spaced on the scalp. The analysis of power spectral density with EEG is a common way to assess and monitor mental workload and fatigue because of the fluctuation in EEG waveforms – delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–18 Hz)-in various scalp areas [15]. Transition state between increasing workload and mental fatigue generally results in an overall **increase in delta, theta and alpha** frequency band, and a **decrease in beta** frequency bands. Various EEG indexes have been derived to isolate and monitor changes in EEG waveform that are associated with mental fatigue.

Pope and colleagues (1995) at the NASA developed an **engagement task index** based on EEG frequency bands applied in a closed-loop system to modulate task allocation. This index is defined by the ratio of frequency bands **Beta/(Alpha + Theta)** [16]. This index decreases as mental fatigue increases.

The ratio of theta power in frontal areas over alpha power in parietal area is a well-known ratio index to measure workload. This index captures theta power increase in frontal regions and alpha power decrease in parietal regions during increases in task demands [17]. Thus, increases in the **theta frontal to alpha parietal ratio** indicate an increase in task load perception. However, from transition states from high mental workload to increasing mental fatigue levels, this ratio decreases as alpha power starts increasing [18].

## 4 Methods and Materials

We aimed to successfully generate mental fatigue in a VR session while collecting EEG and eye tracking data from participants. EEG data from the experiment were transformed and fed to multiple machine learning models to select the one which best fits mental fatigue EEG measures, and eye tracking data was used to isolate the pupil diameter to derive the labels. In addition, we expect to track the fluctuation of workload and task engagement to analyze predictive patterns and possible formulas for mental fatigue.

### 4.1 Development of Mental Fatigue

We create and increase mental fatigue of participants performing two types of cognitive tasks set in a simulated VR environment: one with **distractors** and one **without distractors**. Distractors are fake answers or wrong hints that will be displayed in the user's virtual visual field. Their purpose is to distract participants in order to increase the difficulty of the exercises. During these tasks, fatigue is measured via **pupil dilation, workload, and task engagement**. The exercises chosen to generate mental fatigue are aligned with the **active fatigue** type: mental fatigue is generated through cognitive tasks which aim to exert a high cognitive load on the participants. Participants will perform both sets of tasks consecutively (25 min).

The set of cognitive tasks **with distractors** consists of mental arithmetic tasks, anagram tasks and backward digits span (BDS) tasks. The user must perform these tasks

within the time allocated for each. For each cognitive tasks, fake answers in the form of “hints” are displayed to distract the users.

Then another set of cognitive tasks **without distractors** is presented to the participant. The goal here is to show that the eventual fatigue has been previously generated and is no more due to the distractors. These exercises continue to generate mental fatigue as they require concentration, attention, memory, and other cognitive resources. This set is composed of an attention exercise, a naming exercise, and three different memory tests to evaluate contextual/visual memory, working memory, and short-term memory which are all 5 min duration.

## 4.2 Participants, Data Preprocessing and Labels

31 participants (15 female and 16 male) aged between 19 and 29 years old were invited to a room at Beam Me Up office, partner of the project, 5925 Monkland Ave, H4A1G7, Montréal, to complete the different steps of the experiment and a few real-time and offline outcome measures.

**Selection of participant:** Due to a too large number of failed electrodes during the experiment, the central region electrodes and 10 participants were additionally discarded from the dataset. Thus, the final dataset was thus composed of 12,398 EEG power spectral density segments of prefrontal, frontal right, frontal left, temporal right, temporal left, parietal, and occipital electrodes from the 21 participants (10 female and 11 male).

The EEG signal was band-pass filtered with a fourth-order Butterworth filter and passed onto a wavelet denoising filter to remove signal noise. The power spectral density was computed using the Welch method and used to compute the absolute power of theta, alpha, beta and delta at each electrode. Workload and task engagement were computed using the theta frontal to alpha parietal ratio and the NASA task engagement EEG indexes.

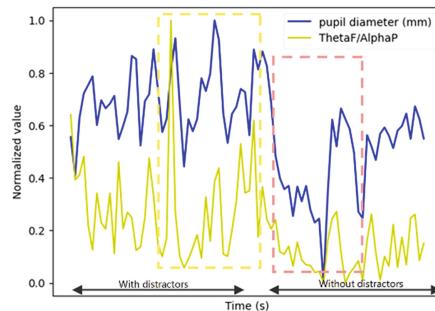
Labels have been established based on variation of the pupil diameter, which is a well-known physiological indicator of mental fatigue and mental workload: **pupil increases** with respect to baseline when subjects are under **high mental charge** and **decreases** when **fatigue rises** [13, 14]. Baseline range is calculated at the first 60 epoch for each participant. 10 epoch moving segments by 1 epoch increments are assigned to the label 0 if the mean of the segment falls between the baseline range (no signs of fatigue), 1 if the mean falls above baseline range (slow increase in fatigue), and 2 if the mean falls below the baseline range (important increase in fatigue).

## 5 Results and Discussion

The experimental results are presented in the following subsection. The first section will be concerned with analysis of the pupil size with respect to the method chosen to classify mental fatigue. The second part will be concerned with the fatigue indicators, workload and task engagement, progression throughout the experiment. Finally, the third part will address selection of the best machine learning model to classify mental fatigue.

## 5.1 Pupil Size Analysis

Labels were assigned with respect to pupil size variation to identify segments of increasing workload and fatigue. Assigning fatigue labels according to eye measurements is a methodology that was employed by Ren Ziwu and colleagues [7] to classify mental fatigue using an RBF neural network. The correlation between pupil diameter and workload/fatigue progression was verified against a well-known EEG indicator of workload and fatigue: **Theta F/Alpha P** (theta bands of frontal electrodes over alpha bands of parietal electrodes). Thus, **increases in the theta frontal to alpha parietal ratio** indicates an increase in workload. An **increase of this index followed by a decrease** indicates **transition states** from high mental workload to increasing mental fatigue levels [18]. Figure 1 shows the measure of such a ratio during cognitive exercise for participant #2. We see, in dashed yellow, an increase of both indexes followed by a decrease in dashed red which indicates transition states from high mental workload to increasing mental fatigue levels.



**Fig. 1** Normalized values of the smoothed ThetaF/AlphaP ratio and pupil diameter

The correlation value between the smoothed and normalized pupil diameter and thetaF/alphaP ratio over all participants is 0.54, which is a strong relation for the nature of the data (see Fig. 1). The distribution of the labels during the first half and the second half of the cognitive tasks part of the experiment shows a significant decrease in absence of fatigue segments and slow fatigue progression proportion. Moreover, we noted a significant increase of rapid fatigue segment proportion during the second half of cognitive tasks compared to the first half. To compare the different fatigue segments, we used the Wilcoxon non-parametric test. This test was chosen (instead of paired t-test) because the distribution of fatigue segments did not follow the normality assumption from parametric tests (Table 1). Results show a significant decrease in non-fatigue and slow fatigue signs, and significant increase in rapid fatigue progression signs.

## 5.2 Workload and Task Engagement

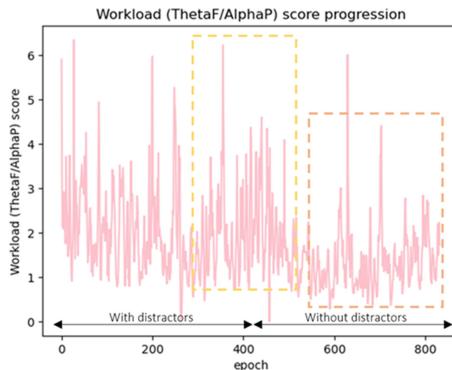
To follow the evolution of workload during the tasks, we compare the measures of the first half of tasks with the measures of the second half. Normally we have generated fatigue using distractors in the first tasks, and the second group of tasks are executed

**Table 1** Wilcoxon test of the proportion of mental fatigue signs comparing the first half of the experiment and the second half

Label number	Median first half	Median second half	Alternate hypothesis	Residual statistic	Pvalue
0	0.888	0.607	$\mu_0 > \mu_1$	168	0.00002
1	0.025	0.001	$\mu_0 > \mu_1$	5	0.005
2	0.041	0.393	$\mu_0 < \mu_1$	0	0.000004

without distractors but the participant is already tired. Hence, we observe a significant decrease in workload score during the second half of the cognitive tasks compared to the first half ( $p < 0.05$ ). However, we did not observe a significant decrease ( $p < 0.05$ ) in task engagement for the majority of participant. This phenomenon can be explained by the nature of the fatigue generated by the cognitive tasks: active fatigue. Active fatigue (fatigue induced by high workload condition) results in a slower task disengagement overtime than passive fatigue (fatigue induced by low workload condition) [12]. Thus, a longer period of cognitive tasks would have been needed to see a change in task engagement.

Figure 2 shows the progression of workload. This index is known to decrease as mental fatigue increases. In dashed yellow, an increase of theta/alpha score followed by a decrease, shown in dashed red, indicates transition states from high mental workload to increasing mental fatigue levels.



**Fig. 2** ThetaF/AlphaP score of participant #1 during the mental fatigue generation tasks

### 5.3 RBF SVM for Mental Fatigue Classification

Our goal was to classify states of mental fatigue segment during a period where participants were asked to perform in various cognitive tasks. All 8 electrodes cerebral regions with their transformation, except for central electrode region, were used in the

feature matrix for classification. The balanced accuracy as opposed to standard accuracy was chosen as one of the parameters to evaluate model performance, as the proportion of labels across the dataset was unbalanced. Various machine learning (ML) algorithm candidates (Nearest Neighbors, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, Naïve Bayes, QDA) were trained and tested in order to select the ML algorithm which best fits our data with respect to evaluation metric ‘balanced accuracy’ and ‘f1’ measures. Among the 8 different classifiers candidates, RBF SVM showed a better performance with respect to balance accuracy (87.0%) and f1 (0.821) evaluation metrics. After parameter tuning achieved through grid search algorithm, the balanced accuracy of the RBF SVM on the test data was 95%.

## 6 Conclusion

Mental fatigue is a complex multi-faceted state resulting from a change of emotions and cognitive capacity. Although it is impossible to fully recreate real-world contexts and emotions from which mental fatigue and recuperation can arise, VR session enables us to achieve laboratory setting that are closer to the ones observed in real world situations. Our results showed significant decrease in the workload score (theta/alpha ratio) and pupil size across participants during the completion of cognitive tasks in the VR environment, indicating increasing of mental fatigue. While we did not find significant changes in task engagement, we believe this phenomenon is the results of the fast rotation of different cognitive tasks, which did not allow enough time for participant to get bored by them. Using EEG for feature data and pupil diameter to derive labels, we were able to develop an RBF SVM classifier able to detect **signs of mental fatigue** with 95% balanced accuracy. Thus, we believe that this intelligent tutoring system can aid in providing intelligent assistance on one’s mental resources available for learning and productivity and motivate learners to rest until cognitive capacities are restored.

**Acknowledgements.** We acknowledge NSERC-CRD (National Science and Engineering Research Council Cooperative Research Development), Prompt, and BMU (Beam Me Up) for funding this work.

## References

1. Ishii, A., Tanaka, M., Watanabe, Y.: Neural Mechanisms of Mental Fatigue. *Reviews in the Neuroscience* **25**(4), 469–479 (2014)
2. Hopstaken, J.F., van der Linden, D., Bakker, A.B., Kompier, M.A.J.: The window of my eyes: task disengagement and mental fatigue covary with pupil dynamics. *Biological Psychology* **110**, 100–106 (2015)
3. Hopstaken, J.F., van der Linden, D., Bakker, AB., Kompier, M.A.J.: A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology* **52**(3), 305–315 (2015)
4. Horvat, M., Dobrinić, M., Novosel, M., Jerčić, P.: Assessing emotional responses induced in virtual reality using a consumer EEG headset: A preliminary report. In: Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1006–1010. Opatija, Croatia (2018)

5. Tarrant, J., Viczko, J., Cope, H.: Virtual reality for anxiety reduction demonstrated by quantitative EEG: A pilot study. *Front. Psychol* **9**(1280) (2018)
6. Kamińska, D., Smołka, K., Zwoliński, G.: Detection of Mental Stress through EEG Signal in Virtual Reality Environment. *Electronics* **10**(22), 2840 (2021). <https://doi.org/10.3390/electronics10222840>
7. Ziwu, R., et al.: EEG-based driving fatigue detection using a two-level learning hierarchy radial basis function. *Frontiers in Neurorobotics* **15** (2021). <https://doi.org/10.3389/fnbot.2021.618408>
8. Csathó, A., Linden, D.V., Hernádi, I., Buzás, P., Kalmár, G.: Effects of mental fatigue on the capacity limits of visual attention. *Journal of Cognitive Psychology* **24**(5), 511–524 (2012). <https://doi.org/10.1080/20445911.2012.658039>
9. Chaouachi, M., Jraidi, I., Frasson, C.: Modeling Mental Workload Using EEG Features for Intelligent System. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) *User Modeling, Adaption and Personalization. UMAP 2011* (2011)
10. Fan, J., Smith, A.P.: The impact of workload and fatigue on performance. In: Longo, L., Leva, M. (eds.) *Human Mental Workload: Models and Applications. H-WORKLOAD 2017. Communications in Computer and Information Science*, pp. 726 (2017)
11. Desmond, P.A., Hancock, P.A.: Active and passive fatigue states. In: Hancock, P.A. (ed.) *Desmond Stress, workload, and fatigue*, pp. 455–465 (2001)
12. Saxby, D.J., Matthews, G., Hitchcock, E.M., Warm, J.S., Funke, G.J., Gantzer, T.: Effect of active and passive fatigue on performance using a driving simulator. *Proc. Hum. Fact. Ergonom. Soc. Ann. Meet.* **52**(21), 1751–1755 (2008)
13. Bafna, T., Bækgård, P., Hansen, J.P.: Mental fatigue prediction during eye-typing. *PLoS One* (2021). <https://doi.org/10.1371/journal.pone.0246739>
14. Menekse Dalveren, G.G., Cagiltay, N.E., Ozcelik, E., Maras, H.: Insights From Pupil Size to Mental Workload of Surgical Residents: Feasibility of an Educational Computer-Based Surgical Simulation Environment (ECE) Considering the Hand Condition. *Surgical innovation* **25**(6) (2018). <https://doi.org/10.1177/1553350618800078>
15. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience and Biobehavioral Review* **44**, 58–75 (2014)
16. Pope, A.T., Bogart, E.H., Bartolome, D.S.: Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology* **40**, 187–195 (1995)
17. Raul, F.R., et al.: Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments. *Frontiers in Neuroscience* **14** (2020). <https://doi.org/10.3389/fnins.2020.00040>
18. Cao, T., Wan, F., Wong, C.M., da Cruz, J.N., Hu, Y.: Objective evaluation of fatigue by EEG spectral analysis in steady-state visual evoked potential-based brain-computer interfaces. *Biomed Eng Online* (2014). <https://doi.org/10.1186/1475-925X-13-28>.



# An Approach for Detecting Gaming the System Behavior in Programming Problem-Solving

Hemilis Joyse Barbosa Rocha<sup>1</sup>(✉),  
Patricia Cabral de Azevedo Restelli Tedesco<sup>1</sup>, Evandro de Barros Costa<sup>2</sup>,  
and Julios Suruagi Rocha<sup>2</sup>

<sup>1</sup> Computer Center-Cin, Federal University of Pernambuco, Recife, Brazil  
`{hjbr,pacrt}@cin.ufpe.br`

<sup>2</sup> Institute of Computing-IC, Federal University of Alagoas, Maceió, Brazil  
`{evandro,jslr}@ic.ufal.br`

**Abstract.** In ITS, for students' success in carrying out their activities, it is essential to provide scaffolding, such as hints and feedback. Although using the ITS has increased student engagement and effort in the classroom, some students have responded to the ITS's support facilities with an inappropriate behavior called gaming the system. Thus, in this article, we explore the phenomenon of gaming the system behavior, studying underlying factors related to automatically detecting when students game the system. We aim to develop a model for detecting gaming the system behavior in a computer programming student during problem-solving activities. Particularly, we have had special attention on testing the influence of variables such as (i) the student's level of belief and the system's level of belief in the difficulty level of the problem in detecting gaming the system; and (ii) partial submission of the problem associated with complete submission, in program form, in the detection of system manipulation. To conduct the detector's development, we developed an appropriate environment for data collection and preparation of these data to be used by supervised machine learning algorithms, allowing the detection of the behavior. The development of the detector involved training single and ensemble machine learning algorithms to classify the system's gaming behavior and obtain models with the best accuracy, including exploring, evaluating, and comparing different algorithms. The results show, considering the average of all algorithm results concerning all measures, the XGBoost ensemble classifier achieved the best performance.

**Keywords:** gaming the system · programming problem · machine learning

## 1 Introduction

Students have used Intelligent tutoring systems (ITS) with different needs, preferences, and characteristics, being applied to a variety of domains [2]. In these

systems, for students' success in carrying out their activities, it is essential to provide scaffolding, such as hints and feedback. The feedback feature is a central scaffolding for student learning and may be required from an ITS whenever necessary. Although using ITS has increased student engagement and effort in the classroom [4], some students have responded to the ITS's help, feedback, and support with an inappropriate behavior called "gaming the system" [3]. This behavior is defined in the literature as "trying to succeed in an educational environment by exploiting the properties of the system rather than learning the material and trying to use that knowledge to answer correctly" [1] and has been reported in several studies [5, 6].

Previous work has increasingly linked gaming behavior to poor student learning outcomes [13]. Therefore, information from students' interactions with the ITS generated a student model with "game the system" behavior. Such models were created using knowledge engineering techniques [7–9, 11, 16] and machine learning [5, 6, 11]. However, the previous work inspected class versus student [14], finding that, to determine the game, knowing the variables related to the students has much less predictive power than knowing the variables of the tutor's classes. Other previous work [8] found that gaming system behavior is more related to learner variables than knowledge domain skill variables. In this sense, there is still no consensus on the relevant variables to predict the behavior of the gaming system. In this article, we explore the phenomenon of gaming the system behavior, studying underlying factors related to automatically detecting when students game the system. Thus, we aim to develop a model for detecting gaming the system behavior in a computer programming student during problem-solving activities. Specifically, we have had particular attention on testing the influence of variables such as (i) the student's level of belief and the system's level of belief in the difficulty level of the problem in detecting gaming the system; and (ii) partial submission of the problem associated with complete submission, in program form, in the detection of gaming the system. To conduct the detector's development, we developed an appropriate environment for data collection and preparation of these data to be used by supervised machine learning algorithms, allowing the detection of the behavior. The development of the detector involved training machine learning algorithms, single and ensembles, to classify gaming the system behavior and obtain models with better accuracy, including exploring, evaluating, and comparing different single classifier algorithms, such as decision tree, K-NN, neural network, and SVM.

Furthermore, we consider ensemble classifiers such as Random Forest (RF) [27], Gradient Boosting Machine (Gradient Boosting), [25] Adaptive Boosting (AdaBoost) [26] and XGBoost [24] to proceed with the exploratory analysis. As one of the results of this study, We evaluated the classifiers using six metrics: ROC curve, accuracy, precision, recall, F-measure, and Kappa. The results show a strong relationship between the partial submission of the problem associated with the complete submission, in program form, in detecting gaming the system. Thus, regarding the classifiers, we noticed that considering the average of all algorithm results concerning all measures, the XGBoost ensemble classifier achieved the best performance.

## 2 Related Work

Focusing on unwanted behavior in the search for help or feedback, some authors have created models to detect a behavior called gaming the system. There is a machine learning Latent Response Model (LRM) to discern which students frequently gaming the system in a way correlated with low learning [12]. This model corroborates the hypothesis that students gaming the system are more likely to do so under challenging stages. As the work focuses on verifying if the student games more in the difficult moments of the problem, the variables are focused on characterizing such information. Thus, students were divided into three groups: Students who never showed gaming the system behavior, students who showed gaming behavior but had poor performance, called “GAMED-HURT” and students who showed gaming the system behavior and performed well, called “GAMED NOT-HURT”. In another work, the authors study high school students’ goals, attitudes, behavior, and learning in a mathematics tutor [17]. For the study, two sources of data were combined: a questionnaire about students’ motivations and beliefs and records of each student’s actions with the tutor. In addition, the system gaming instances were divided into two cases: gaming without impact and gaming with negative impact. The authors found that the frequency of gaming the system does not correlate with a known measure of performance goals; rather, gaming is related to not liking computers and the tutor.

Using the cognitive algebra tutor, in [19] built a cognitive model based on system data and using knowledge. The data contained information about the actions of the students while using the system, such as time, tips, the context of the problem, input that the user entered, and evaluation of the system, and with that, they can identify whether gaming the system exists or not. In this system, the tutors provide hints, which can be given whenever the student needs help. The more he asks, the more specific the hints are. The Decimal Point system focuses on mathematics for 5th and 6th-grade students and aims to help develop the students by using some game mechanisms [18]. Detectors were developed based on logs using text replays. The game has characters that cause a fantasy to the user using this system. The coding was divided into three phases, and after having accepted reliability, the data was added to machine learning algorithms to understand the encoders. In the previous work, 13 patterns [19] were identified, in this current study [20], a model was built based on these patterns that aim to detect the student’s use of the platform, managing to detect the order in which they are performed.

## 3 Study Design

In this article, we will discuss the work of detecting games in undergraduate students in the field of computer programming through the use of a cognitive tutoring system. Thus, we have had special attention on testing the influence of variables such as (i) the student’s level of belief and the system’s level of

belief in the difficulty level of the problem in detecting gaming the system; and (ii) partial submission of the problem associated with complete submission, in program form, in the detection of system manipulation. For this, we discuss the characteristics of the data and the whole data labeling process.

### 3.1 ADA Cognitive Tutor

The ADA cognitive tutor was developed to be used in a computer programming discipline of an undergraduate computer science course. ADA offers the following features: i) Problem-solver: In Fig. 1, there is an example of the screen shown to students during the resolution of a problem. Each problem was elaborated following the structure in Fig. 1. For each problem, the student must make two types of submission: i) submission of the alternative (partial solution): the student only submits the alternative corresponding to some question about the solution of the problem, ii) submission of the solution complete: the student submits the program developed as a solution to a problem (complete solution). ii) Tips: As shown in Fig. 1, the ADA provides the student with four levels of hints with different levels of abstraction. In tip 1, there is a more superficial explanation until in tip 4 the student can receive the answer to the problem. Struggling students can ask for help anytime, and the tutor will provide increasingly specific, multi-step tips. The student is free to request any level of hint during problem-solving. ii) Assessment of the level of the problem: After the student submits the partial solution to the problem, the system asks the student to evaluate the problem's difficulty level, indicating among the levels: very easy, easy, medium, difficult, and very difficult, as shown in Fig. 1.

### 3.2 Data Set

For this study, we used data from ADA, a web-based computer programming tutoring system. The data are from 32 students between 18 and 22 years of age in a university's undergraduate computer science course in one semester in 2022. Students answered an average of 41.3 problems ( $SD = 25.0$ ) with an average of 4.3 times ( $SD = 2.7$ ) resolutions for each problem. The total dataset consisted of 5,307 problem solutions.

**Data Labeling.** When we compare our amount of data with previous work, we realize that the number of solution attempts per problem can be significant in ADA, as it requires two types of answers: partial solutions (multiple choice activity) associated with a complete solution (for example, programs written in a programming language) for each problem, rather than reifying each step of the problem individually. Thus, our dataset consisted of approximately an order of magnitude fewer data per student-time than was used in previous detectors of gaming behavior [5, 10, 18, 19, 21]. As the success of any machine learning algorithm depends on the relevance of attributes, selecting attributes to include or exclude is an important exercise. As the work of Baker et al. [12] was used as

The screenshot shows a digital learning environment for a cognitive tutor. At the top, the title "ADA" is displayed in a dark blue header bar. Below it, a light gray navigation bar contains the text "Problem 1". The main content area has a white background. A problem statement is presented in a box, followed by a list of four multiple-choice options labeled A through D. To the right of the problem is a "Tip 1" box containing a note about Lucas's program. At the bottom of the screen are two buttons: "Skip" on the left and "Submit" on the right.

**Problem 1**

Lucas wanted to buy a new cell phone, so he searched for cell phone prices on the Amazon.com website and liked one worth R\$1190.00. Consulting the payment options, he discovered that if the payment is made in cash, he will get a 12% discount on the total value of the cell phone. However, if payment is made in installments, there will be a 15% surcharge. Lucas created a program to find out the amount he should pay in cash on his cell phone. Which alternative corresponds to the sequence of steps to calculate the discount value that Lucas used in his program?

Select the CORRECT alternative Alternatives

A. 1 - Store the value 1190 in a "price" variable; 2 - Multiply the "price" variable by 0.12 and store it in a "discount" variable; 3 - Subtract the variable "price" from the variable "discount" and store it in the variable "finalprice";

B. 1 - Multiply the "price" variable by 0.12 and store it in a "discount" variable; 2 - Store the value 1190 in a "price" variable; 3 - Subtract the variable "price" from the variable "discount" and store it in the variable "precofinal";

C. 1 - Store the value 1190 in a "price" variable; 2 - Subtract the "price" variable by 0.15 and store it in a "discount" variable; 3 - Multiply the "price" variable by the "discount" variable and store it in the "finalprice" variable;

D. 1 - Store the value 1190 in a "price" variable; 2 - Multiply the "price" variable by 0.15 and store it in a "discount" variable; 3 - Subtract the variable "price" from the variable "discount" and store it in the variable "finalprice".

**Tip 1**

Lucas's program was developed only to calculate the cash value with the 12% discount and does not consider any additions.

**Hint 1**    **Hint 2**  
**Hint 3**    **Hint 4**

**Skip**    **Submit**

**Fig. 1.** ADA - Cognitive Tutor

a partial guide for this process, some of the attributes used are very similar to theirs, just adapted to the particularities of the Assistance system and variable time windows. We generated a base dataset with 145 attributes using log data of students' actions. The attributes are broken down and defined below:

- Alternatives that the student clicked (amount, mean, variance and standard deviation);
- Alternatives that the student did not click (quantity, mean, variance and standard deviation);
- The correct answers (quantity, mean, variance and standard deviation);
- Incorrect answers (quantity, mean, variance and standard deviation);
- Tips requested (quantity, mean, variance and standard deviation);
- Unsolicited hints (quantity, mean, variance and standard deviation);
- Solved problems (quantity, mean, variance and standard deviation);
- Unsolved problems (quantity, mean, variance and standard deviation);
- Solved problems that the student considered [very easy, easy, medium, difficult, very difficult] (quantity, mean, variance and standard deviation);
- Solved problems that the system considered [very easy, easy, medium, difficult, very difficult] (quantity, mean, variance and standard deviation);
- Unsolved problems that the system considered [very easy, easy, medium, difficult, very difficult] (quantity, mean, variance and standard deviation);
- Date and time for each alternative that the student clicked (amount, mean, variance and standard deviation);

- Date and time for each alternative that the student did not click (quantity, mean, variance and standard deviation);
- Date and time for each correct answer (amount, mean, variance and standard deviation);
- Date and time for each incorrect answer (quantity, mean, variance and standard deviation);
- Date and time for each requested tip (amount, mean, variance and standard deviation);
- Date and time for each unsolicited tip (amount, mean, variance, and standard deviation);
- Date and time for each solved problem (quantity, mean, variance and standard deviation);
- Date and time for each unsolved problem (quantity, mean, variance and standard deviation);
- Date and time for each solved problem that the student considered [very easy, easy, medium, difficult, very difficult] (amount, mean, variance and standard deviation);
- Date and time for each solved problem that the system considered [very easy, easy, medium, difficult, very difficult] (quantity, mean, variance and standard deviation);
- Date and time for each unsolved problem that the system considered [very easy, easy, medium, difficult, very difficult] (quantity, mean, variance and standard deviation);
- How much time the student spend on the whole task;
- How much time the student spend in each section;
- How much time the student spend on each problem in the section;
- How long the student spent on each trial of the problem.

All log files were labeled using text replays concerning whether the student was gaming. The text replays represent a segment of student behavior from the log files in a textual format. Text replays offer several advantages: text replays can be classified extremely quickly, achieve acceptable inter-rater reliability [29], and have previously been used to train gaming the system detectors [28]. We follow the thirteen gaming the system behavior patterns presented in previous works [19]. The text repetitions were conducted by two labelers (first and fourth authors) in two rounds. Thus, a small set of text repetitions was performed, and the inter-rater reliability was low - Cohen's Kappa was 0.41 in differentiating between gaming behavior and non-gaming behavior and 0.33 in distinguishing all behaviors from each other. In the second round, the two labelers recoded the observations they disagreed on, discussed their interpretation of the behaviors, and then conducted additional labeling. In the second set of text replays, Cohen's Kappa was 0.85 in differentiating gambling and non-gaming behavior and 0.82 in distinguishing all behaviors. In the end, 986 randomly selected student behavior sequences were coded, representing 4321 problem-solving attempts.

### 3.3 Machine Learned Models

In this Section, we present an overview of proposed models for automatically detecting gaming behavior in the computer programming domain. However, we clean the dataset before building the detection models, removing missing data.

#### Database Pre-processing

First, we removed the attributes that carry information related to the system, such as, for example, student id, problem id, start time for solving each problem, etc. Then, to select numerical data in the same format, the format conversion of the type hour, minutes, and seconds was performed to store only the seconds of the data that saved the moment when a tip was requested. The same happened at the moment of submitting a response. In addition, the fields referring to problem difficulty (system belief), problem difficulty (student belief), partial solution, and complete solution from categorical to numerical data were converted using the Python pandas library. Finally, to deal with the amount of data, the imblearn<sup>1</sup> library was used, which manages to simulate the data in a smaller amount using close neighbors for this purpose, preventing the model from being biased on a class.

#### Algorithms

To develop the models, we analyzed the performance of both single and ensemble classifiers. For the single classifiers, we apply the decision tree algorithms [1] SVM [6], KNN, and neural networks (MLP). In terms of ensemble learning algorithms [23], the use of random forest has appeared in the literature among the most used approaches for student classification tasks in other contexts [22]. So, to build our model, we employed several concurrent ensemble methods, as follows: Random Forest (RF) [27], Gradient Boosting Machine (Gradient Boosting), [25] Adaptive Boosting (AdaBoost) [26] and XGBoost [24] to proceed with the exploratory analysis.

#### Assessment

We evaluated the classifiers using six metrics: ROC curve, accuracy, precision, recall, F-measure, and Kappa, at the level of significance ( $p < 0.05$ ). First, the ROC curve area is a feature that relates true positive and false positive rates. The area under this curve varies between 0.5 and 1 and shows the ability of a model to perform correct and incorrect classifications [28]. This measure reveals the probability that if the detector compares two sequences, one involving games and one not involving games, it will correctly identify which sequence is which. Accuracy indicates the overall performance of the model. Recall, also called true positive rate or sensitivity, is the percentage of true positives overall positive examples. Precision is the percentage of true positives over all those classified as positive. The F-measure measure relates to precision and recall. So we used the second, third, fourth, and fifth measurements to assess whether the detector was properly balancing between identifying games and avoiding false positives. Kappa evaluates whether the detector determines the correct action sequences as games rather than chance. A Kappa of 0 indicates that the detector works randomly, and a Kappa of 1 indicates that the detector works perfectly.

---

<sup>1</sup> <https://imbalanced-learn.org/stable/>.

## 4 Results and Discussion

In 76.3% of the attempts to submit a resolution to the problems made by the students, there is a disagreement between the system's belief and the student's belief about the difficulty level. Among resolution attempts with controversy, 68.9 % showed at least one gambling behavior distributed according to Table 1. As we can observe in Table 1, the most extreme problem difficulty levels present a higher percentage of disagreement. For example, 7.34% of problem resolutions the system assigns as "very easy" the student believes are only easy, 10.67% medium, 7.7% difficult, and 3.3% very difficult. In this same table, when analyzing the other extreme in 13.93% of the resolutions, the student believes that the problem considered by the system as very difficult is, in fact, very easy, 11.89% easy, 9.98% medium, and 11.7% difficult. When we analyzed the impact of the combination of the partial solution variables associated with the complete solution variable separately, we noticed that in 44% of the resolutions with cases of gaming the system, there was a partial submission, and there was not a complete one. Furthermore, in 23% of the resolutions, there was a partial solution with a complete solution copied from the internet or an incomplete solution taken from another problem.

**Table 1.** Percentage of disagreement between the system's belief and the student's belief about the problem's difficulty level.

		Student belief				
		Very easy	Easy	Medium	Difficult	Very difficult
System belief	Very easy	–	7.34	10.67	7.7	3.3
	Easy	1.45	–	7.01	5.05	4.1
	Medium	1.56	0.34	–	2.43	3.45
	Difficult	2.9	1.55	1.2	–	1.9
	Very difficult	13.93	11.89	9.98	11.7	–

### 4.1 Gaming the System Behavior Detectors

Having developed a reasonably effective detector of gaming behavior in general, we discuss our attempts to differentiate types of gaming behavior in this section. We compared the performance of the detectors using three different combinations of datasets: (i) dataset 1: standard dataset with the variables cited in Sect. 3.2; (ii) dataset 2: dataset 1 plus the variables system belief and student belief about the level of the problem; and (iii) data set 3: dataset 2 plus variables referring to the partial and total solution of the problem. In all cases, the same cross-validation method was used as above. For this, we present in Tables 2, 4 and 3 all the values of the measurements retrieved from each algorithm, and we will follow the discussion route: dataset 1, dataset 2, dataset 3 and all together.

**Table 2.** Base detectors

	Accuracy	Precision	Recall	F-Score	Kappa	AUC
Decision tree	0.84	0.88	0.85	0.84	0.34	0.79
K-NN	0.78	0.79	0.79	0.79	0.42	0.79
SVM	0.68	0.81	0.67	0.63	0.33	0.72
MLP	0.84	0.85	0.84	0.84	0.53	0.76
RF	0.91	0.90	0.89	0.90	0.49	0.82
Gradient Boosting	0.90	0.90	0.90	0.91	0.62	0.81
AdaBoost	0.90	0.89	0.89	0.89	0.63	0.83
XGBoost	0.91	0.90	0.90	0.90	0.47	0.82

**Default Dataset:** By differentiating all the algorithms, according to the values in Table 2 for accuracy, the level of accuracy of the results obtained by applying each classifier, the XGBoost, and Random forest algorithms also had the best efficiency measures 91%. However, if we consider the accuracy, recall, and F-Score, the model was more accurate in capturing the behavior related to the gradient boosting game (Precision = 90%, Recall = 90% and F-Score = 91%) presents slightly better performance. Considering the cases of false positives and true positives, XGBoost (AUC = 0.83) performed better. AdaBoost achieved a better performance considering the Kappa (0.63).

**Table 3.** Detectors considering the variables delivering the partial solution and total solution

	Accuracy	Precision	Recall	F-Score	Kappa	AUC
Decision tree	0.85	0.87	0.85	0.84	0.64	0.83
K-NN	0.79	0.79	0.79	0.79	0.44	0.79
SVM	0.80	0.81	0.79	0.63	0.63	0.81
MLP	0.84	0.85	0.84	0.84	0.53	0.83
RF	0.96	0.95	0.96	0.94	0.61	0.86
Gradient Boosting	0.93	0.93	0.92	0.91	0.62	0.88
AdaBoost	0.92	0.85	0.94	0.94	0.64	0.81
XGBoost	0.96	0.94	0.94	0.94	0.68	0.82

**Variables Delivering Partial Solution and Total Solution Dataset:** When we analyzed the algorithms' results with combining partial solution and complete solution variables, according to Table 3, most achieved better performance. The classifier that presented the best performance considering measures of accuracy, precision, recall, and F-Score, was the Random Forest (Accuracy = 96%, Precision = 95%, Recall = 96%, and F-Score = 94%). On the other hand,

if we analyze the Kappa measure, AdaBoost (0.63) has a better performance. XGBoost performs better considering the area under the ROC curve (0.83).

**Table 4.** Detectors considering the variables system belief and student belief about the difficulty level of the problem

	Accuracy	Precision	Recall	F-Score	Kappa	AUC
Decision tree	0.84	0.88	0.85	0.84	0.64	0.73
K-NN	0.78	0.79	0.79	0.79	0.44	0.74
SVM	0.68	0.81	0.67	0.63	0.63	0.68
MLP	0.84	0.85	0.84	0.84	0.53	0.77
RF	0.90	0.92	0.92	0.91	0.49	0.81
Gradient Boosting	0.89	0.91	0.91	0.91	0.62	0.82
AdaBoost	0.91	0.89	0.90	0.90	0.63	0.87
XGBoost	0.92	0.91	0.91	0.92	0.63	0.89

**Variables System Belief and Student Belief About the Difficulty Level of the Problem Dataset:** Considering the dataset 3 Table 4, with the variables system belief and student belief about the level of the problem, the performance of the models was a little better than with the standard data set and a little worse than using dataset 3 with the variables: partial solution and complete solution. Among them, on average, the best performance was achieved by XGboost with (Accuracy = 91%, Precision = 90%, Recall = 90%, and F-Score = 90%). However, according to Kappa (0.63) and AUC (0.83), AdaBoost performs better than the others. After analyzing the results of the measurements of each classifier, we realized that the inclusion of the variables partial solution and complete solution improved the performance of the models by an average of 7.3%. On the other hand, when we include the system belief and student belief variables regarding the system's difficulty level, performance dropped by an average of 2.2% but improved by 4.1% in relation to the standard dataset.

## 5 Conclusion

The purpose of this work was to study the influence of some features related to identifying gaming the system behavior and, based on these features, to develop a computational solution using supervised machine learning techniques for a classification task, which automatically detects when students use the gaming the system behavior during the process of solving programming problems. Thus, the results indicate that creating effective predictive models for this task is feasible. In this sense, among the contributions of this work are: (1) the development of an environment for data collection; (2) the creation and annotation of a database with the data; (3) the generation of machine learning models with

good performance for classifying the presence of gaming the system behavior. Regarding the classifiers, we noticed that considering the average of all algorithm results concerning all measures, the XGBoost ensemble classifier achieved the best performance. Furthermore, according to all measures, the ensemble classifiers performed better than the single classifiers.

Including the variables partial solution, complete solution, student belief, and system belief regarding the problem's difficulty level, they have improved the detection of gaming the system behavior. However, although including the system belief and student belief variables regarding the problem's difficulty level did not achieve the best result among the classifiers, it revealed valuable information. For example, in 76.3% of the attempts to submit a solution to the problems made by the students, there is a disagreement between the system's belief and the student's belief about the difficulty level. And among resolution attempts with disagreement, 68.9% showed at least one gambling behavior. And there is more significant disagreement at the most extreme levels (very easy and very difficult), which leads us to suppose that there are indications that students tend to game more when they believe that the problem is very easy or very difficult.

## References

1. Baker, R.S.J., de Carvalho, A.M.J.A.: Labeling student behavior faster and more precisely with text replays. In: Proceedings of the EDM, vol. 2008, pp. 38–47 (2008)
2. Ghaleb, E., Popa, M., Hortal, E., Asteriadis, S., Weiss, G.: Towards affect recognition through interactions with learning materials. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 372–379 (2018). <https://doi.org/10.1109/ICMLA.2018.00062>
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 (2004)
4. Schofield, J.W.: Computers and Classroom Culture. Cambridge University Press, Cambridge (1995)
5. Baker, R.S.D., Corbett, A.T., Roll, I., Koedinger, K.R.: Developing a generalizable detector of when students game the system. User Model. User Adap. Inter. **18**, 287–314 (2008)
6. Baker, R.S.J., Mitrović, A., Mathews, M.: Detecting gaming the system in constraint-based tutors. In: De Bra, P., Kobsa, A., Chin, D. (eds.) UMAP 2010. LNCS, vol. 6075, pp. 267–278. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13470-8\\_25](https://doi.org/10.1007/978-3-642-13470-8_25)
7. Aleven, V., McLaren, B.M., Roll, I., Koedinger, K.R.: Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. Int. J. Artif. Intell. Educ. **16**, 101–130 (2006)
8. Gong, Y., Beck, J.E., Heffernan, N.T., Forbes-Summers, E.: The fine-grained impact of gaming (?) on learning. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 194–203. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13388-6\\_24](https://doi.org/10.1007/978-3-642-13388-6_24)
9. Johns, J., Woolf, B.: A dynamic mixture model to detect student motivation and proficiency. In: Proceedings of the AAAI 2006, pp. 163–168 (2006)

10. Muldner, K., Burleson, W., Van de Sande, B., VanLehn, K.: An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impact. *User Model. User Adap. Interact.* **21**, 99–135 (2011)
11. Walonoski, J.A., Heffernan, N.T.: Prevention of off-task gaming behavior in intelligent tutoring systems. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 722–724. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_80](https://doi.org/10.1007/11774303_80)
12. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 531–540. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30139-4\\_50](https://doi.org/10.1007/978-3-540-30139-4_50)
13. Fancsali, S.E.: Data-driven causal modeling of gaming the system and off-task behavior in cognitive tutor algebra. In: NIPS Workshop on Data Driven Education (2013)
14. d Baker, R.S.: Differences between intelligent tutor lessons, and the choice to go off-task. In: International Working Group on Educational Data Mining (2009)
15. Aleven, V., McLaren, B., Roll, I., Koedinger, K.: Toward tutoring help seeking: applying cognitive modeling to meta-cognitive skills. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 227–239. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30139-4\\_22](https://doi.org/10.1007/978-3-540-30139-4_22)
16. Walonoski, J.A., Heffernan, N.T.: Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 382–391. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_38](https://doi.org/10.1007/11774303_38)
17. Baker, R.S., et al.: Do performance goals lead students to game the system? In: AIED (2005)
18. Richey, J.E., et al.: Gaming and confusion explain learning advantages for a math digital learning game. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021, Part I. LNCS (LNAI), vol. 12748, pp. 342–355. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78292-4\\_28](https://doi.org/10.1007/978-3-030-78292-4_28)
19. Paquette, L., de Carvalho, A.M., Baker, R.S.: Towards understanding expert coding of student disengagement in online learning. In: CogSci (2014)
20. Paquette, L., Baker, R.S.: Variations of gaming behaviors across populations of students and across learning environments. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 274–286. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-61425-0\\_23](https://doi.org/10.1007/978-3-319-61425-0_23)
21. Paquette, L., Baker, R.S.: Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interact. Learn. Environ.* **27**(5–6), 585–597 (2019)
22. Alamri, A., Sun, Z., Cristea, A.I., Stewart, C., Pereira, F.D.: MOOC *next week* dropout prediction: weekly assessing time and learning patterns. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 119–130. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-80421-3\\_15](https://doi.org/10.1007/978-3-030-80421-3_15)
23. Yang, D., et al.: Turn on, tune in, drop out: anticipating student dropouts in massive open online courses. In: Proceedings of the 2013 NIPS Data-Driven Education Workshop, vol. 11 (2013)
24. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
25. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001)

26. Hastie, T., et al.: Multi-class adaboost. *Stat. Its Interface* **2**(3), 349–360 (2009)
27. Liu, Y., Wang, Y., Zhang, J.: New machine learning algorithm: random forest. In: Liu, B., Ma, M., Chang, J. (eds.) ICICA 2012. LNCS, vol. 7473, pp. 246–252. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-34062-8\\_32](https://doi.org/10.1007/978-3-642-34062-8_32)
28. Gonçalves, L., et al.: ROC curve estimation: an overview. *REVSTAT-Stat. J.* **12**(1), 1–20 (2014)
29. Baker, R.S., Corbett, A.T., Wagner, A.Z.: Human classification of low-fidelity replays of student actions. In: Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, vol. 2002 (2006)



# An Approach to Generating Adaptive Feedback for Online Formative Assessment

Fuhua Lin<sup>(✉)</sup> and Supun De Silva

Athabasca University, Alberta T9S 3A3, Canada  
`{oscarl,sdesilva}@athabascau.ca`

**Abstract.** In this paper, we propose a novel approach to generating adaptive feedback by identifying the chain of weakest learning objectives to a learner working in a domain. It combines the domain model based expert-driven model with question-answering based data-driven model. The domain model is an AND/OR graph of domain knowledge structure based on the revised Bloom's taxonomy, defining the learning objectives of the domain and the corresponding pre-requisite relationships. The adaptive formative assessment process uses an improved Top-Two Thompson sampling algorithm for solving the best arm identification problem in the multi-armed bandit framework. The simulation results show the feasibility and performance of the proposed approach.

**Keywords:** Formative assessment · adaptive feedback generation · domain modeling · bandit algorithms · simulated learners · online learning · adaptive learning

## 1 Introduction

Formative assessment is a process used by teachers and learners to keep learning moving forward. It has been proposed to make education more effective through identifying and closing learning gaps. Research reports that the key element of formative assessment is feedback, awareness of gaps between current students' knowledge and their learning objectives, and where to go next to alleviate those deficiencies. However, most existing research for conducting formative assessment is for classroom-based educational environments, in which the most efficient way of gathering evidence about student weaknesses and taking actions for improvement is face-to-face interaction between teachers and students [1]. However, in online education, esp. Self-paced online learning, face-to-face discussions are not feasible. Thus, using curriculum-embedded assessments or incorporating a formative assessment tool into an online course is desirable for formative assessment.

In an adaptive learning environment, formative assessment must be designed by considering learners in different knowledge states. This requires that the question selection rules and stopping rules of the formative assessment process should be adaptive to the knowledge state of the learner. It also requires a balance or tradeoff between efficiency and reliability of formative assessments. The proficiency profile of a learner can be

modeled with learning outcomes or learning objectives based on the revised Bloom's taxonomy [2].

Considering the efficiency of formative assessment, the first one is to identify the weakest learning objective from among a set of learning objectives to be assessed or the student's current Zone of Proximal Development (ZPD) [3]. Also, to recommend an action for remedying the weakness effectively, it is needed to identify the pre-requisite of learning objectives that affect the student's mastery of the weakest learning objective. Such a chain of learning objective will be recommended to the learner as feedback as they have the most improvement potential and need remediation first. The second one is to identify and remedy all the unmastered learning objectives. However, this strategy may overwhelm students if the proficiency levels of many learning objectives are still low. Also, this strategy needs a lengthy questioning process.

In this paper, we propose a novel approach to identifying the weaknesses among a set of learning objectives the student is learning and generating adaptive feedback to the student, including learning objectives with most improvement potential. It combines the domain model based expert-driven model and question-answering based data-driven model. The domain model is represented as an AND/OR graph in Artificial Intelligence [4]. It defines the learning objectives of a domain and their pre-requisite relationships. To realize the trade-off between efficiency and reliability of formative assessment process, we use the multi-armed bandit (MAB) framework [5] to model the adaptive formative assessment process. The MAB framework has been applied to solve problems in adaptive learning [6]. We tested the feasibility and performance of the proposed approach through simulation [7].

The advantages of the proposed approach:

- (1) It can generate adaptive feedback a chain of learning objectives that have most improvement potential to a learner working in a domain in any knowledge state and at any time during her/his study.
- (2) It utilizes an efficient and easy-to-implement Bayesian approach for best arm identification to realize adaptive formative assessments.

The rest of the paper is organized as follows. Related work is reviewed in Sect. 2. Section 3 formulates the automated feedback generation problem and the proposed algorithm to solve the problem. Next, our simulation model, experimental setup, and result analysis are presented in Sect. 4. Finally, we conclude the paper and point out future research directions in Sect. 5.

## 2 Related Work

Systematic literature reviews of automated feedback generation have been done in [8, 9]. Our approach for adaptive feedback generation is involved with domain modeling, adaptive formative assessment, and weakest learning objective identification (WLOI) in an online learning environment. A domain model is fundamental to generating adaptive feedback as it tells where feedback is deduced [10, 11]. Adaptive formative assessment in online learning is a process of sequential and combinatorial questioning and decision-making under uncertainty and randomness. It can be modeled as a Beta Bernoulli Bandit

model [12]. WLOI is an optimal adaptive allocation of assessment sequence to identify the weakest among a finite set of learning objectives. The agent sequentially chooses questions relevant to one or more learning objectives and observes independent noisy responses by the student. The goal is to allocate assessment questions intelligently so that the weakest learning objective can be identified confidently after a small number of questions. WLOI is modeled as a best arm identification (BAI) problem which is a particular pure exploration problem [13]. We use Thompson sampling (TS) algorithm [5] for adaptively selecting assessment learning objectives and relevant questions. Such a Bayesian algorithm begins with a prior distribution over unknown mastery of the learning objectives. The agent learns as the responses are gathered, and beliefs are updated to form a posterior distribution. This posterior distribution gives a principled mechanism for reasoning about the uncertain mastery of learning objectives, and for assessing the probability any given learning objective is weakest.

Thompson sampling (TS) has attracted a great deal of recent interest [12]. But it is designed to maximize the cumulative reward earned while sampling. As a result, in the long run it allocates almost all effort to measuring the estimated-best arm and requires a huge number of total measurements to certify that none of the alternative designs offer better performance. In the BAI problem with the fixed-confidence setting, the goal is to identify the best arm with as few as actions possible, thus the standard TS is not suitable for solving the problem. Russo (2018) proposed a top-two TS to avoid the low efficiency issue [14]. By choosing between two different candidate arms in each round, these sampling rules enforce the exploration of sub-optimal arms, which would be under-sampled by vanilla Thompson sampling due to its objective of maximizing rewards. Shang et al. (2020) proposed TTTS/T3C algorithm, a computationally more favorable variant of TTTS algorithm [15]. We adopted and augmented the TTTS/T3C algorithm by using the structural information among arms (i.e., learning objectives) and increase the efficiency of adaptive formative assessment [16].

### 3 Problem Formulation

In online education, as an integral and important part of learning, formative assessments provide students with constructive feedback to close existing learning gaps. Minimizing the time spent on formative assessment is vital to incentivize this repetitive process and make it more efficient. Thus, question selection rules and stopping rules of a formative assessment should be *adaptive* to the knowledge state of the learner and pursue a balance or tradeoff between efficiency and reliability of the assessment. Thus, the goal of this research is to design algorithms for adaptive feedback generation for online adaptive learning. It can dynamically select questions and stop the process according to the learner's responses so that the learner answers as few questions as needed and then can receive feedback for improvement.

#### 3.1 Domain Modelling

We consider a *domain*  $D$  consisting of a set of topics of a subject that a student is learning. The mastery criteria of the topics in  $D$  is specified as  $K$  ( $K > 0$ ) learning objectives (or

learning outcomes). These learning objectives, denoted as  $LO = \{lo_j\}_1^K$ , are designed based on the revised Bloom's taxonomy [2]. There are pre-requisite relationship among these learning objectives. We use an AND/OR graph to model the learning objectives and their pre-requisite relations. We denote such a LO-Graph as  $G(V, E)$ , where the nonempty set  $V$  of vertices is the disjoint union of two subsets  $V_{AND}$  of AND-vertices and  $V_{OR}$  of OR-vertices.

### 3.2 The Student Model

For any  $lo \in LO$ , a set of questions which is denoted as  $Q_{lo}$  is designed for testing it. For each question  $q \in Q_{lo}$ , it may significantly associate with one or more learning objectives at least including  $lo$ . Given student  $s$ , and question  $q$ , the correctness of the response  $r$  to  $q$  by  $s$ , called the reward,  $x_s(q, r) \in \{0, 1\}$ . If the answer is correct,  $x_s(q, r) = 1$ . If the answer is wrong, then  $x_s(q, r) = 0$ . For each learning objective  $lo_k$ , the success probability or mean reward of student  $s$  answers to questions for testing  $lo_k$  is denoted as each  $\mu_k(s)$ . For simplicity, we assume that all questions in  $Q_{lo_k}$  have the same difficulty level. That is,  $p(x_s(q, r) = 1 | r; q \in Q_{lo_k}) = \mu_k(s)$ . We call  $\{\mu_k(s), lo_k \in LO\}$  *student model*, which are unknown prior and to be learned. We assume that they are fixed over time in a formative assessment process.

### 3.3 Weakest Learning Objective Identification

A feedback generation is started by the student who is learning topics in a ZPD or her/his teacher at any time during her/her study in any knowledge state. To generate feedback, two tasks must be done. The first task is to identify the gaps between the target mastery criteria and the current proficiency levels through adaptive assessment. The second is to provide/recommend suitable remedial learning material or learning activities to the learner according to the weaknesses identified. This paper focuses on the fist task.

Considering efficiency, we can identify the learning objective that has most improvement potential or needs remediation first. In formative assessment, the key trade-off is between efficiency and reliability: reducing the number of formative assessment questions that get asked vs. making sure that the questions that were asked were the one that should be asked. For example, if the goal is to identify the weakest learning objective of the student's knowledge within  $LO$ , then the key is to choose questions as few as possible for student  $s$  to identify one learning objective for which  $\mu_k(s)$  attains its smallest value, i.e.,  $lo^* = argmin_{k \in K} \mu_k$ . To realize the trade-off, we use the notion of multi-armed bandits (MAB) because the MAB algorithms are data-driven and can balance exploration and exploitation and make sequential decisions under uncertainty [5]. We modeled the adaptive formative assessment as a Beta-Bernoulli bandit [5]. We adopted the recommendation rule and stopping rule in the TTTS-T3C algorithm [15] to identify the weakest/strongest learning objective among a set of learning objectives.

### 3.4 The Chain of Weakest Learning Objectives

From our simulation, the number of questions needed for MAB based WLOI algorithms scales linearly with the number of learning objectives  $K$  in the domain  $D$ . Also, it is not

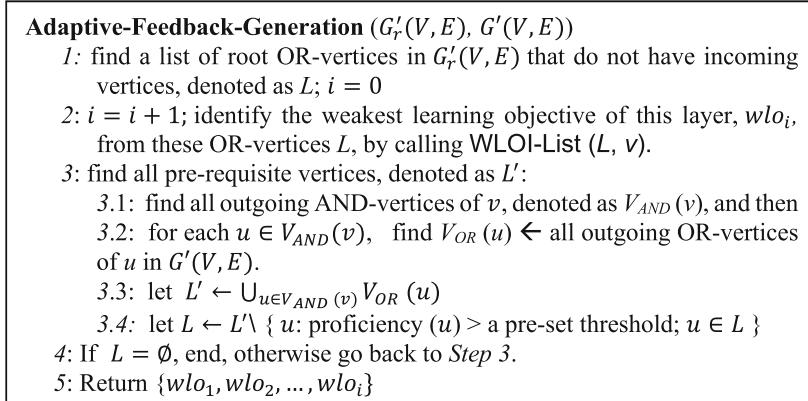
enough to just provide the weakest learning objective among a set of learning objectives to be assessed. Usually, a learning objective is weak to the learner because there exist some weak pre-requisites of this learning objective. To generate constructive feedback, we need to tell the learner which pre-requisites they should work on in terms of maximizing the overall increase in the mastery of domain by the student. To solve this problem, our strategy is a layer-based approach that uses the structural information of LO-Graph and the breadth-first search (BFS) of graphs. First, we identify which learning objective among the latest learning objectives that do not serve as a pre-requisite of other learning objectives in LO. The result is called the first layer weakest learning objective, denoted as  $wlo_1$ . After this, our method will further identify which branch of the pre-requisites of  $wlo_1$  should be chosen for the student. We repeat this until we find the foremost pre-requisite learning objective of  $wlo_1$  that needs to be improved.

### 3.5 The Proposed Algorithm

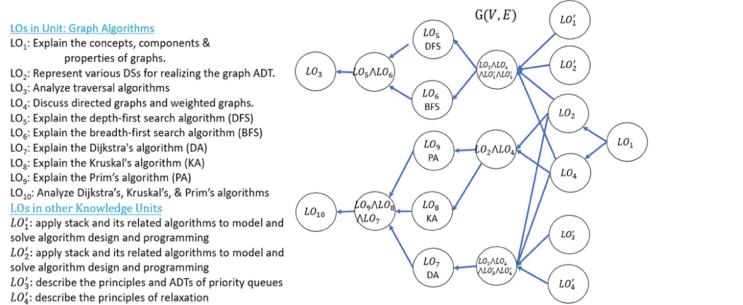
The learning objectives in the current ZPD is a sub-graph of  $G(V, E)$ , denoted as  $G'(V, E)$ . To find the chain of weakness and generate feedback for remediation, we use a backward process and thus first reverse the LO-Graph  $G(V, E)$ . The reversed LO-Graph is denoted as  $G_r(V, E)$ , which is still an AND/OR graph with the same  $V$  as in  $G(V, E)$ . Accordingly, the  $G'(V, E)$  for the current ZPD is reversed to a sub-graph of  $G_r(V, E)$ , denoted as  $G'_r(V, E)$ . The algorithm for generating the formative feedback is illustrated in Fig. 1. From the above algorithm, it starts from some vertices in  $G'_r(V, E)$  and then generates a path from the weakest learning objective in the first layer to the weakest learning objective identified in the last time of calling WLOI-List in the domain  $G'_r(V, E)$ . The learning objectives in the path are then used as the adaptive feedback to the student. The first output vertex from calling WLOI-List is defined as the weakest learning objective identified. The other weakest learning objectives output from calling WLOI-List are viewed as learning objectives that should be improved to remedy the gaps. The last output from the algorithm is the foremost weakest learning objective that has most improvement potential. It is worth mentioning that the output learning objectives may be beyond the current ZPD and still in  $G'(V, E)$ . From these identified weakest learning objectives, some remedial learning material can be recommended to the student. Algorithm WLOI-List for a list of learning objectives uses TTS/T3C algorithm in [15].

In our implementation, we use the observation dependencies amongst the arms in  $G(V, E)$ , selecting a question for testing a learning objective yields reward of not only this learning objective, but also of those learning objectives that are connected to it by an edge.

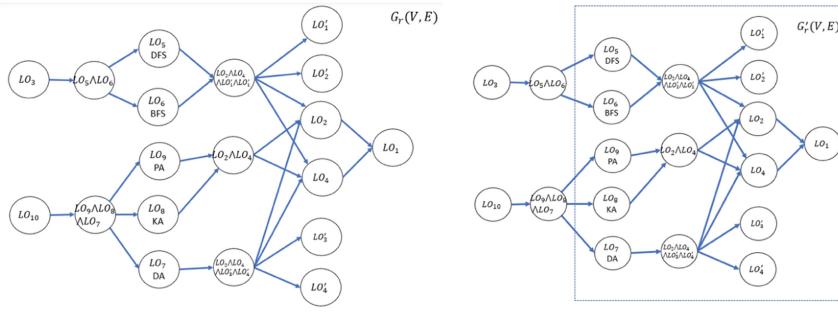
Figure 2 shows an example of the feedback generation process using a set of the learning objectives of unit ‘Graphs’ in Data Structure and Algorithms in computer science. The ZPD is  $\{LO_1, LO_2, LO_4, LO_5, LO_6, LO_7, LO_8, LO_9\}$ .  $LO'_3, LO'_3, LO'_3, LO'_3$  are learning objectives in other units beyond the current ZPD. At the end of the formative assessment process,  $LO'_3$  is identified as the foremost learning objective that has most improvement potential and needs remediation first.



**Fig. 1.** The pseudocode of the adaptive feedback generation process.



(a)



(b)

(c)

**Fig. 2.** (a) The LOs of unit ‘Graphs’ in a computer science course called Data Structure and Algorithms and its  $G(V, E)$ ; ; (b) the  $G_r(V, E)$ ; ; (c) an example of  $G'_r(V, E)$ ; ; (d) exploring the vertices without descendants in  $G'_r(V, E)$  . . ; (e)  $LO_7$  is identified as *wlo-1*; (f) exploring the  $LO_7$ ’ pre-requisite vertices; (g)  $LO'_3$  is identified as *wlo-2*; end the process.

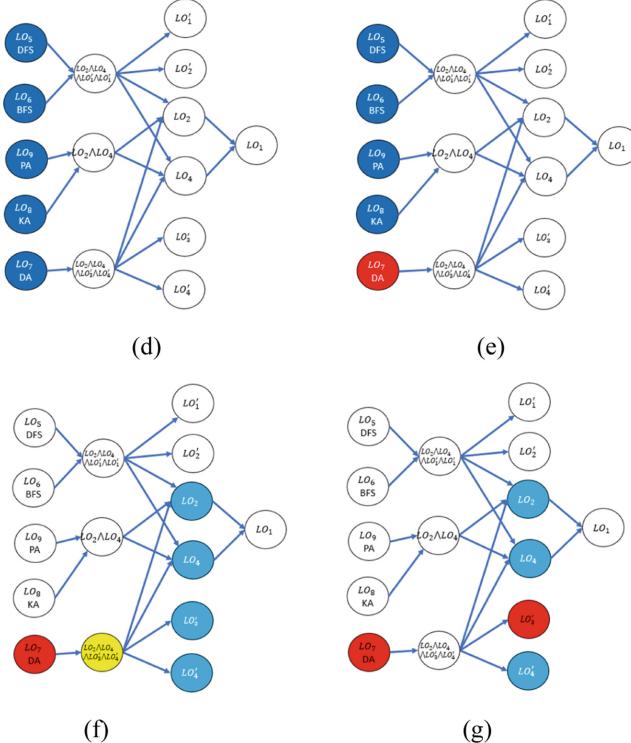


Fig. 2. (continued)

## 4 Experiment

### 4.1 Simulated LO-Graphs

For each round, a LO-Graph to be assessed with the associated questions is randomly generated for a simulated learner reflecting his/her proficiency and formative assessment request. We simulated these graphs randomly for our simulation as this would facilitate repeating the experiment many times. To simulate LO-graphs, we utilize the Networkx library (<https://networkx.org/>) in Python to help randomly generate the graphs. We will generate graphs containing 2~15 vertices. Each vertex represents a learning objective. A probability of 0.05 is used to determine whether there will be an edge between two vertices. The main reason such a low edge probability is used is to ensure that we will not have bidirectional edges in our randomly generated graphs. If we consider a real situation, learning objective #1 might be a prerequisite for learning objective #2. However, we cannot also have learning objective #2 be a prerequisite for learning objective one. Therefore, we cannot have bidirectional edges in our graph and setting the edge probability to 0.05 ensures none of our randomly generated graphs will have bidirectional edges. Also, we set one learning objective significantly weaker than others for each layer. We do this by checking all the randomly generated probabilities of answering a learning objective correct for all arms within each layer.

Once  $\{\mu_{lo}(s) : lo \in LO\}$  has been generated, it is used as parameters of a Bernoulli Bandit function which simulates the question-answering process and return a reward. For simplicity, we assume that all questions for testing a learning objective has the same difficulty level.

## 4.2 Simulated Learners

A learner  $s$  is simulated by pre-setting the proficiency values of each learning objective in a simulated LO-Graph. Each node  $lo$  has been assigned a proficiency level of  $s$ . The proficiency value is randomly assigned using the following formula:

$$\mu_{lo}(s) \triangleq w_1 \times apt(s) + w_2 \times pre(lo) \quad (1)$$

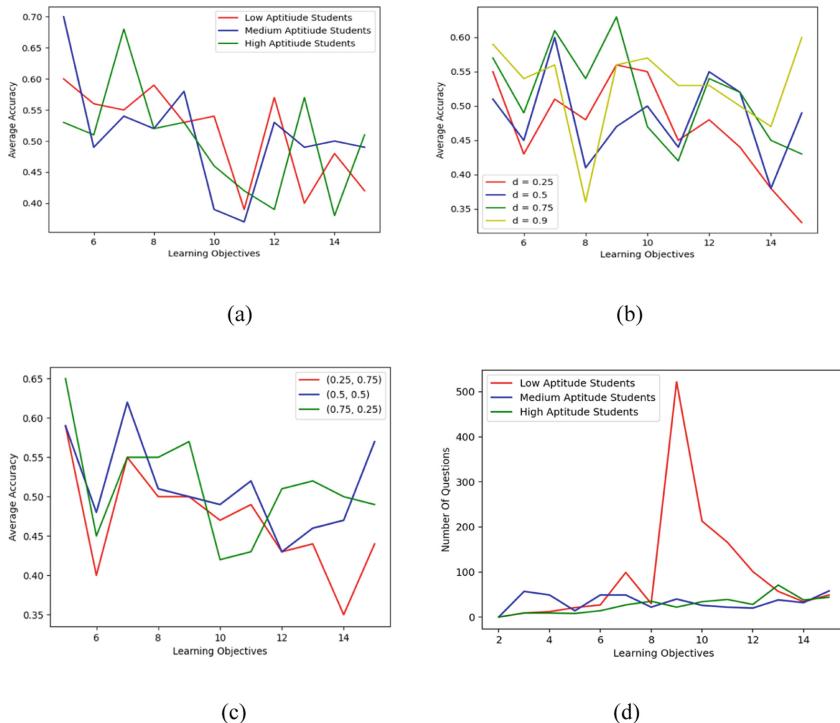
To model the learners' proficiency, we capture that they differ from one another, calling this difference their aptitude. The aptitude of student  $s$  is denoted as  $apt(s)$  and represented with a number between 0.1 and 1, where a higher number means that in the evaluation function the learner will "master" learning objective more easily. To keep this model simple, we say that an individual learner's aptitude does not change as they learn. The second term,  $pre(lo)$ , is computed as follows. If  $lo$  has no prerequisites, then the number 1 is taken as the value for this term. If  $lo$  has a single prerequisite  $pr$ , the number 1 is also taken if the learner has mastered (e.g.,  $p(s, pr) \geq$  a pre-set threshold) the prerequisite; otherwise, 0 is taken. If  $lo$  has multiple OR prerequisites, it takes the ratio of the number of the prerequisites the learner mastered and the total number of pre-requisites. If one of the OR prerequisites of  $lo$  is an AND-vertex, it takes 1 if all pre-requisites in the AND-vertex has been mastered by the learner, otherwise it takes 0. Each term is weighted as to its overall importance in determining the probability in (1).  $w_i \in [0, 1]$ , ( $i = 1, 2$ ), thus,  $p(s, lo) \in [0, 1]$ . After pre-setting  $\{\mu_{lo}(s), s \in LO'\}$ , the weakest learning objective has been determined and to be identified by the formative assessment algorithm.

To test the performance of the approach for the different weights, we set the weights ( $w_1, w_2$ ) for two items in (1) as three sets (0.2.5, 0.7.5), (0.5, 0.5) and (0.75, 0.25). Also, to test the performance of the approach for different learners, we simulate three types of learners (scenarios): Low aptitude students with  $apt()$  values from 0.1 to 0.3; Medium aptitude students with  $apt()$  values from 0.4 to 0.6; High aptitude students with  $apt()$  values from .7 to 1. 1000 learners were created for each scenario. Lastly, the WLOI-List algorithm forces exploration by randomly choosing between the best performing arm and the second-best performing arm based on a parameter  $\beta = 0.5$  [15]. Also, the threshold values for stopping was set at four different values 0.25, 0.5, 0.75, and 0.9.

## 4.3 Results and Analysis

A Python program was developed for tests and experiments using simulated learners. Figure 3(a), 3(b), and 3(c) shows the performance of the proposed algorithm in three scenarios with different parameters. The accuracy is defined as the average percentage of successfully identified the pre-set weakest learning objectives of the simulated learners, that is, the pre-set weakest learning objective is identified in the chain of the weakest

learning objectives. We can see that accuracy depends on many factors including the number of edges of the LO-Graph (i.e., the sparseness of the LO-Graph). The accuracy of predicting the weakest learning objective of the student decreases as the number of learning objectives increase. This is expected because the algorithm will have to test more learning objectives to determine the weakest learning objective. Also, as the number of learning objectives increase there is a greater chance of having more equally weak learning objectives for a student. This may cause issues for the algorithm to identify the weakest learning objective resulting in a decrease in accuracy. However, this does not reduce the usefulness of the algorithm as identifying an equally weak arm will allow the student to continue learning.



**Fig. 3.** (a) Average accuracy of predicting the weakest learning objective for three aptitude students; (b): Average accuracy of predicting the weakest learning objective for four different stopping criteria in the **Adaptive-Feedback-Generation** algorithm; (c): Average accuracy of predicting the weakest learning objective with different weighting for student aptitude and prerequisite mastery when determining whether a student will answer a question for testing learning objectives correctly. (d): Number of questions required to identify the weakest learning objective of three aptitude students; (e): Number of questions required to identify the weakest learning objective with different stopping criteria in the **Adaptive-Feedback-Generation** algorithm; (f): Number of questions to identify the weakest learning objective when different weightings are associated with the student aptitude and prerequisite mastery when determining whether a simulated learner answers a question for testing learning objectives correctly.

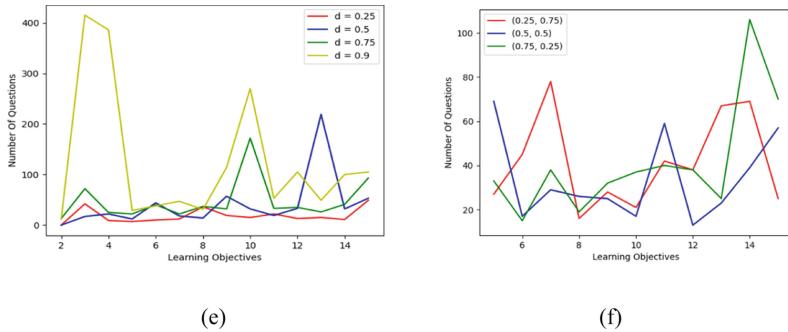
**Fig. 3. (continued)**

Figure 3 (a) shows that there is no significant difference in the algorithm's accuracy of predicting the weakest learning objective based on student aptitude. This suggests that this algorithm can be used with students of any aptitude. This is useful in a real classroom environment as we do not know a student's aptitude. Figure 3 (b) shows that the chosen stopping criteria did not significantly affect the algorithm's accuracy of predicting the weakest learning objective. This is unexpected as we would expect higher stopping criteria to have a greater accuracy as more questions will need to be asked to reach that stopping criteria. The stopping criteria of 0.25 performed weaker than the others when 12 or more learning objectives were used. It performed comparatively well when less than 12 learning objectives were used. However, a low stopping criterion such as 0.25 should be avoided as slips and guesses can easily have a negative impact of algorithm accuracy as fewer questions will be needed to reach the stopping criteria. This suggests that a very high or low stopping criteria should not be used as there will be a tradeoff between the achieved accuracy and the number of questions required (view Fig. 3 (e)). A stopping criterion of 0.5 or 0.75 would serve as a reasonable general stopping criteria. This will balance between the number of questions needed and the effect of slips and guesses on algorithm accuracy. Figure 3 (c) shows that an equal weighting or a weighting favoring a student's aptitude over prerequisite mastery improved the algorithm performance. An equal weighting between student aptitude and prerequisite mastery is a good general recommendation for future simulations of this algorithm with simulated learners.

Figure 3 (d) shows that medium aptitude students required more questions to detect the weakest learning objective when six or less learning objectives were used. Low aptitude students require significantly more questions than other learners when learning objectives seven through 12 were used. This could likely be attributed to not updating dependency information when a student answers a question incorrectly. If a student answers a question correctly, we can infer that they most likely understand the prerequisites and update prerequisite information. However, if a student answers a question incorrectly, we cannot update prerequisite information. A student may understand the prerequisite material and not understand the additional material needed to answer this learning objective. On the other hand, the student may not understand the prerequisite information. Due to either of these scenarios being likely, we cannot update prerequisite information when a student answers a question incorrectly. This may also be an artifact

of our random LO-Graph creation as characteristics such as number of edges affects the efficiency of weakest learning objective identification, although all LO-Graphs were generated randomly. Simulations on fixed LO-Graphs representing actual courses is warranted for further testing. High aptitude students required more questions when 13 or more learning objectives were used. It is best to run this simulation with a fixed LO-Graph to determine whether significantly more questions will be needed for students of different aptitudes. Figure 3 (e) shows that higher stopping criteria requires more questions to identify the weakest learning objective. This is expected as more questions will be needed to reach the higher stopping criteria. Therefore, 0.5 or 0.75 serves as good general stopping criteria to balance accuracy and the number of questions needed to identify the weakest learning objective. Figure 3 (f) showed no significant difference based on weighting. Equal weighting or weighting favoring student aptitude mostly outperformed the weighting favoring prerequisite mastery for formative assessments with 13 or fewer learning objectives. Combining these results with Fig. 3 (c), an equal weighting should be selected for future simulation.

## 5 Conclusions and Future Work

We have presented a novel approach to automating feedback generation by identifying the chain of weakest learning objectives to a learner working in a domain. We combine the domain knowledge representation about learning objectives and their pre-requisite relations with the Thompson sampling bandit based online and active machine learning algorithm to estimate the proficiency levels of students using the responses by them. We used the simulation approach to validate the feasibility and performance of the proposed approach.

We do not consider “partial credit” answers (e.g., based on the usage of hints or on the response time). Also, we did not model the soft or probabilistic pre-requisites among learning objectives in the proficiency model. By defining the opposite arm selection rule in the TTTS/T3C based **WLOI-List** algorithm, we can generate the strongest learning objectives as adaptive feedback to the student as encouragement.

For the future work, we will further develop methods for automated feedback generation to adapt to varied feedback strategies or students’ needs. Also, we will consider “partial credit” answers (e.g., based on the usage of hints or on the response time) and the soft or probabilistic pre-requisites among learning objectives in the proficiency model. Finally, we will test the proposed approach to measure its actual effectiveness in real-world learning settings.

**Acknowledgements.** We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovate, and Athabasca University, Canada.

## References

- Heritage, H.M.: Formative assessment: Making it happen in the classroom. Corwin, Thousand Oaks, CA (2010)

2. Krathwohl, D.R.: A revision of blooms taxonomy: an overview. *Theory Into Practice* **41**(4), 212–218 (2002)
3. Chaiklin, S.: The Zone of Proximal Development in Vygotsky’s analysis of learning and instruction. In: Kozulin, A., Gindis, B., Ageyev, V., Miller, S. (eds.) *Vygotsky’s educational theory and practice in cultural context*, pp. 39–64. Cambridge University, Cambridge (2003)
4. Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*, 2nd Edition. Prentice Hall (2010)
5. Lattimore, T., Szepesvári, C.: *Bandit Algorithms*, 1st ed., Cambridge University Press (2020)
6. Mui, J., Lin, F., Dewan, M.A.A.: Multi-armed Bandit Algorithms for Adaptive Learning: A Survey. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) *AIED 2021. LNCS (LNAI)*, vol. 12749, pp. 273–278. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78270-2\\_49](https://doi.org/10.1007/978-3-030-78270-2_49)
7. McCalla, G.I., Champaign, J.: Simulated Learners. *IEEE Intell. Syst.* **28**, 67–71 (2013)
8. Keuning, H., Jeuring, J., Bastiaan, H.: A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education* **19**(1), 1–43 (2019)
9. Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., Weerdt, J.D.: A review of automated feedback systems for learners: classification framework, challenges and opportunities. *Computers and Education* **162** (2021)
10. Gouli, E., Gogoulou, A., Papanikolaou, K.A., Grigoriadou, M.: An adaptive feedback framework to support reflection, guiding and tutoring. In: *Advances in Web-Based Education: Personalized Learning Environments*, pp. 178–202. IGI Global (2006). <https://doi.org/10.4018/978-1-59140-690-7.ch008>
11. Lu, C.: Automated Feedback Generation for Learner Modeling in Intelligent Tutoring Systems, PhD Dissertation. University of Alberta 2021). Available: <https://era.library.ualberta.ca/items/261291a8-c71f-4b3e-9cce-b43324c40a6a>
12. Russo, D.J., Roy, B.V., Kazerouni, A., Wen, Z.: A tutorial on thompson sampling. In: *Foundations and Trends® in Machine Learning* (2018). <https://doi.org/10.1561/2200000070>
13. Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in multi-armed bandits problems. In: *Algorithmic Learning Theory*, pp. 23–37. Springer (2009). [https://doi.org/10.1007/978-3-642-04414-4\\_7](https://doi.org/10.1007/978-3-642-04414-4_7)
14. Russo, D.: Simple bayesian algorithms for best-arm identification. *Operations Research* **68**(6), 1625–1647 (2020). <https://doi.org/10.1287/opre.2019.1911>
15. Shang, X., de Heide, R., Kaufmann, E., Ménard, P., Valko, M.: Fixed-confidence guarantees for bayesian best-arm identification. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR 108, vol. 108, pp. 1823–1832 (2020)
16. Herlihy, C., Dickerson, J.P.: Networked restless bandits with positive externalities. In: *AAAI 2023* (2023)



# How to Generate Early and Accurate Alerts of At-Risk of Failure Learners?

Amal Ben Soussia<sup>(✉)</sup>, Azim Roussanaly, and Anne Boyer

Université de Lorraine, CNRS, LORIA, Campus scientifique,  
54506 Vandoeuvre-lès-Nancy, France

{amal.ben-soussia,azim.roussanaly,anne.boyer}@loria.fr

**Abstract.** The high failure rate is a common serious issue among online learning institutions. In order to deal with this problem, Early Warning Systems (EWS) based on Machine Learning (ML) models are widely adopted in the literature as a solution to help teachers in their pedagogical monitoring. As the name implies, alert generation is intended to be the purpose of an EWS. However, the proposed systems don't go beyond the early prediction of at-risk of failure learners and don't suggest automatic methods to generate alerts. In this paper, we propose an algorithm that automatically generates early and accurate alerts for teachers of at-risk of failure learners. This algorithm uses both an original concept of alert rule to define the alerting method and temporal evaluation metrics to identify the reliable starting time for generating alerts. As a proof of concept, we apply this algorithm on four different EWS using real data of k-12 learners enrolled in online learning courses.

**Keywords:** online learning · early warning systems · machine learning · alert generation algorithm · alert rule · k-12 learners

## 1 Introduction

The use of online learning has rapidly evolved over the last decade, especially with the latest COVID-19 pandemic. It ensures the continuity of the learning process and the accessibility to educational programs from anywhere and at any time. However, this learning mode is facing many challenges, and the most widespread is the high failure rates among online learners. In order to meet this issue, Early Warning Systems (EWS) based on Machine Learning (ML) models are commonly proposed in the Learning Analytics (LA) literature to assist teachers in their pedagogical monitoring. EWS are used to better support learners based on their behavior and performance [3]. These systems are defined as a tool used by educationists to monitor learners' progress and identify those who are at-risk of either failing in a course or dropping out of a program [8]. Therefore, EWS rely on both numeric learning traces of online learners and analytical models of ML [3].

As the name implies, by using an EWS, teachers expect to be alerted of the possible learning risk cases. In other words, generating alerts after identifying

at-risk of failure learners is the purpose of EWS [9]. However, despite the popularity of EWS in LA, the literature doesn't propose particular methods and/or algorithms for an automatic alert generation. The existing EWS don't go beyond the risk prediction and the early identification of learners in critical learning situations. Given the importance of alerting teachers of at-risk of failure learners for a better monitoring and an effective follow-up, the main question is: **how to alert teachers accurately and at the earliest of at-risk of failure learners?**

To answer this question, we propose an algorithm for an automatic, early and accurate alert generation for teachers to alert them of at-risk of failure learners. In order to achieve this goal, this algorithm uses an original concept of *alert rule* ( $R_i$ ) where  $i$  is the number of successive predictions of a learner in a risk class [13]. The objective of an alert rule is to define the alerting method. Moreover, this algorithm uses time-dependent metrics, in particular *earliness* and the *Earliness-Stability Score (ESS)* introduced and defined in [11, 12]. The use of these metrics makes it possible to identify the earliest starting time point  $t_s$  from which the alert generation process following  $R_i$  can be triggered. To choose the best value of  $i$ , each  $R_i$  is evaluated using two metrics inspired from the usual ML Confusion Matrix (CM) and adapted to the alert problem: *precision* and the *False Positive Rate (FPR)* of generated alerts by  $R_i$  since  $t_s$ .

To validate the relevance of this algorithm, we applied it on four different EWS aiming to alert teachers of at-risk learners. These systems use real data of k-12 learners enrolled online within a French distance education center (CNED<sup>1</sup>

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 formalizes the problem and introduces the prerequisites of the alert Algorithm, which is described in Sect. 4. In Sect. 5 and 6, we present respectively the case study and the experimental results. The Sect. 7 concludes on the results and introduces the perspectives of this work.

## 2 Related Work

EWS are used in several fields: alerting about natural disasters, financial and economic slowdowns of companies and the stock market as well as alerting patients of the deterioration of their clinical signs [9]. Therefore, the main objective of EWS has always been alerting about possible future risks to intervene effectively. In fact, an alert is a signal that indicates a possible imminent risk. LA are also using EWS to solve education problems, such as the prediction of at-risk learners. Indeed, the EWS are able to anticipate potential risks based on present information and thus are able to send alerts to the person or group of people who may be affected by these risks and/or capable of countering them [9].

One of the most referenced EWS in the LA is the Course Signals (CS) system of Purdue University [2]. To provide meaningful feedback, this EWS introduces a traffic light system that allows students to see their pass/fail level based on a

---

<sup>1</sup> Centre National d'Enseignement à Distance, <https://www.cned.fr/>.

color code. [1] proposes an EWS to be used in programs relying on e-books as learning materials. This system uses reading data to weekly labeling students as low or high performing. Another EWS in [7] uses student activity traces on the LMS. This system follows a monthly approach to detect significant drops in student activity; these students will then be reported as at-risk. In order to identify students at-risk of failure, course withdrawal and dropout, [6] introduces a multi-view EWS following a weekly prediction incremental approach. [14] proposes an academic EWS in Hangzhou Normal University to early discover and identify the existing and potential students academic learning problems. [4] develops an EWS which purpose is to detect at-risk students in online high education. Thus, stakeholders are able to analyze information through the provided dashboards and teachers can intervene effectively. [5] proposes an EWS to early identify course dropout learners enrolled within an online university. The objective of such a system is to enable an early and effective intervention mechanism.

To summarize, most of the existing research projects working on proposing EWS stop at the prediction phase and don't go beyond the identification of at-risk of failure or dropping out learners. Despite the importance of alerting for an effective intervention, this notion is rarely mentioned and discussed in the literature especially in the LA field. Indeed, [10] highlights the importance of extending the work on EWS to answer more questions rather than the prediction of risk. Thus, in this work, we propose an early alert generation algorithm to overcome the state of the art limitations and design an EWS for online institutions teachers. This algorithm relies on the prediction results returned by ML models to identify the list of learners of whom teachers should be alerted.

### 3 Alert Prerequisites

In this section, we formally present the problem of alert generation. Then, we introduce the prerequisites needed by the alert generation algorithm.

#### 3.1 Problem Formalization

The main goal is to alert teachers of learners who are in a failing learning situation. This alert generation has to be accurate and as early as possible and targets a maximum number of at-risk of failure learners. Therefore, we rely on the prediction results returned by a classification approach.

Assume that  $Y = \{C_1, C_2, \dots, C_m\}$  is the set of predefined class labels and  $y_{risk} \subseteq Y$  is a subset of  $Y$  that groups together the risk class labels. Let  $S = \{S_1, S_2, \dots, S_q\}$  be the set of students in the test dataset and  $T = \{t_1, t_2, \dots, t_k\}$  be the set of prediction times. At each  $t_k \in T$ , each  $S_p \in S$  is represented by a vector  $X_{p_k} = \langle f_1, f_2, \dots, f_z, C_j \rangle_{p_k}$  where  $f_n \in \mathbb{R}$  represents the learning features of  $S_p$  and  $C_j \in Y$  her/his class label. The main objective of this work is to alert a teacher of learners in  $S$  who are at-risk of failure. In other words, the prediction of the learner  $S_p \in S$  in  $C_j \in y_{risk}$  triggers the alert generation for the teacher.

The Fig. 1 illustrates a system of  $S = \{S_1, S_2, S_3\}$ . The set of class labels is  $Y = \{C_1, C_2, C_3\}$  with  $y_{risk} = \{C_2, C_3\}$ . For each learner in  $S$ ,  $\{l_1, l_2, l_3, l_4, l_5\}$  and  $\{p_1, p_2, p_3, p_4, p_5\}$  represent respectively the true class labels and the predicted classes at each of prediction times in  $T = \{t_1, t_2, t_3, t_4, t_5\}$ . A box is colored red (Fig. 1) when a learner  $S_p$  is predicted in a risk class of  $y_{risk} = \{C_2, C_3\}$ ; this is how the teacher alert generation is launched. The Fig. 1 shows that EWS are characterized by the instability of their prediction results over time. Therefore, it would be pertinent to identify a time point from which the alert generation is reliable. This time point is called the *starting time point* ( $t_s$ ).

Furthermore, in order to define the alert method, a first assumption is the following: an alert is generated as soon as a learner is predicted as at-risk (*First assumption*). However, this is a strong assumption as EWS are characterized by prediction oscillation over time. Indeed, relying on the results given by the example of Fig. 1, we can distinguish two types of alert:

- **good alert:** alerting a teacher of an at-risk of failure learner
- **false alert:** alerting a teacher of a successful learner

Thus, the main objective is to generate a maximum number of good alerts starting from  $t_s$ . For this reason, we need to define an *alert rule* ( $R_i$ ).

	$t_1$		$t_2$		$t_3$		$t_4$		$t_5$	
	$l_1$	$p_1$	$l_2$	$p_2$	$l_3$	$p_3$	$l_4$	$p_4$	$l_5$	$p_5$
$S_1$	$C_1$	<span style="background-color: red;">C<sub>3</sub></span>	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$	$C_1$	<span style="background-color: red;">C<sub>3</sub></span>
$S_2$	$C_1$	$C_1$	$C_1$	$C_1$	<span style="background-color: green;">C<sub>2</sub></span>	<span style="background-color: red;">C<sub>2</sub></span>	$C_2$	$C_2$	<span style="background-color: red;">C<sub>1</sub></span>	$C_1$
$S_3$	$C_3$	$C_1$	<span style="background-color: yellow;">C<sub>3</sub></span>	<span style="background-color: red;">C<sub>2</sub></span>	$C_3$	$C_2$	$C_3$	$C_3$	$C_3$	<span style="background-color: red;">C<sub>3</sub></span>

**Fig. 1.** Example of an EWS

### 3.2 The Starting Time Point $t_s$

The alert generation is closely linked to the temporal dimension of EWS and their oscillations over time. Indeed, it is important to start generating alerts from a reliable time  $t_s$ . This starting time  $t_s$  is defined as both early enough to intervene effectively and ensuring the stability of correct predictions over time. For this aim, [12] proposes to calculate the ESS score to observe the compromise between the earliness of predictions of a system and its temporal stability. The temporal stability is defined as the longest sequence of correct predictions over time, while the earliness is the first prediction time when a class label is correctly predicted [11]. Once computed, the ESS is used to identify this earliness time point. Therefore, our algorithm uses the results of ESS to identify the starting time  $t_s$  from which the alert generation can be triggered reliably.

### 3.3 Alert Rule $R_i$

Since our goal is to alert teachers of at-risk learners, we intend to maximize the rate of good alerts and minimize false alerts. Although, producing a false alert is less crucial than missing out on a good alert. An example of a missing alert in Fig. 1 is the following: the learner  $S_3$  is predicted in  $C_1$  at  $t_1$  while she/he belongs to the risk class  $C_3$ . In addition, applying the assumption of Sect. 3.1 can generate a high rate of false alerts. Thus, it is irrelevant to generate early alerts without verifying their precision. In other words, the alert generation process should consider the earliness of alerts but also their precision and the temporal stability of the whole prediction system. Therefore, it is pertinent to propose an alert generation method that respects several parameters:

- early alerts: alerting teachers of at-risk learners at the earliest.
- precise alerts: maximizing the rate of good alerts compared to the total number of generated alerts.
- system stability: stable and successively correct predictions over time.

For all these reasons, we propose to follow the succession of predictions over time and define an alert rule as follows:

**Rule ( $R_i$ ): alert a teacher when a learner is predicted  $i$  times in a row in a risk class**

To summarize, we apply  $R_i$  on the prediction results of each learner. The value of  $i$  is defined depending on the context and the needs. In other words, to generate an alert of a learner  $S_p$ , this latter must have been predicted  $i$  times in a row in a risk class starting from  $t_s$ .

### 3.4 Evaluation of Generated Alerts

To assess the pertinence of the generated alerts following  $R_i$  since  $t_s$ , we adapt the usual ML confusion matrix to the alert evaluation as in Fig. 2. The objective of this evaluation is to identify the value of  $i$  of  $R_i$  that gives a maximum rate of good alerts and a minimum rate of false ones. Therefore, we propose to evaluate  $R_i$  based on two metrics which are *precision* and *FPR* defined as follows:

$$\begin{aligned} \text{precision} &= \text{good alerts rate}/\text{total number of generated alerts} \\ \text{FPR} &= \text{false alerts rate}/\text{total number of generated alerts} \end{aligned}$$

An alert is missing when a learner is actually at-risk but the alert rule does not detect it. Thus, the value  $i$  of  $R_i$  is optimal when the precision is high and both FPR and missed alerts are minimal and tend to 0.

## 4 Alert Generation Algorithm

In this section, we present the alert generation algorithm 1 that proposes an automatic method to generate alerts based on the results of the earliness and

Generated alert		
Real alerts	Alert	No alert
Alert	Good alert	Missing alert
No alert	False alert	No alert

**Fig. 2.** The ML confusion matrix adapted the alert evaluation

the ESS given by [12]. The Algorithm 1 takes as input the list of learners  $S$ , the subset  $y_{risk}$  of the risk class labels, the test dataset  $D_{test}$ , the number of prediction times given by  $T$  and the alert rule to apply given by the value of  $i$ . This algorithm returns  $A_{S_{alert}}$  which is the set of at-risk learners to alert the teachers of. This Algorithm starts by assigning the empty set to the variable  $E_{earliness}$  which will contain the values of the earliness points for each class of  $y_{risk}$  (Line 1). To identify the starting time point, the Algorithm 1 iterates over the set of risk class labels ( $C_k \in y_{risk}$ ) (Line 2). For each  $C_k$ , the variable  $early_{C_k}$  of type list receives the earliness time points (Line 3) [12]. At Line 4, the Algorithm assigns the empty list to the variable  $L_{C_k}$ . Then, it iterates over each earliness point of  $early_{C_k}$  (Line 5). For each earliness point  $j$ , the Algorithm proceeds to the calculation of its corresponding *ESS* score given by  $ess_{j_k}$  (Line 6). Then,  $\langle j_k, ess_{j_k} \rangle$  is added to  $L_{C_k}$  (Line 7). For each  $C_k$ , the Algorithm selects the best couple  $\langle earliness_{C_k}, ess_{C_k} \rangle$  ensuring the earliness of  $C_k$  prediction and the stability of the system (Line 9).  $earliness_{C_k}$  is added to  $E_{earliness}$  (Line 10). Once the  $E_{earliness}$  is established, the Algorithm extracts the minimum earliness value and assigns it to the  $earliness$  variable (Line 12). Then, the Algorithm iterates over the learners set (Line 13). For each  $S_p$ , the Algorithm initializes the prediction time  $t$  to  $earliness$ , the rule to apply  $r$  to 0 and  $alert$  to false. While  $t$  is inferior to  $T$  and the value of  $r$  is inferior to that of  $i$ , the Algorithm continues to run through the predictions (Line 17). If  $S_p$  is predicted in a risk class (Line 18) then the value of  $r$  is incremented by 1 (Line 19). Else, the value of  $r$  is again reset to 0 (Line 21). Then, the value of  $t$  is incremented by 1 (Line 23). For each  $S_p$ , if  $r$  is equal to  $i$  (Line 25), alert receives true (Line 26) and  $S_p$  is added to the  $A_{S_{alert}}$  (Line 27). Else, the teacher shouldn't be alerted of  $S_p$  (Line 29).

## 5 Case Study

In order to validate this approach, we apply the alert generation algorithm on four different EWS based on real data of k-12 learners enrolled in online modules.

### 5.1 Context Description

Our context is the k-12 online learners enrolled within CNED that offers multiple fully distance courses to a large number of heterogeneous and physically

**Algorithm 1.** Alert algorithm - Alert( $S, y_{risk}, D_{test}, T, i$ )**Require:**  $S, y_{risk}, D_{test}, T, i$ **Ensure:**  $A_{S_{alert}}$ 


---

```

1:  $E_{earliness} \leftarrow \emptyset$ 
2: for each  $C_k$  in  $y_{risk}$  do
3:    $earlyC_k \leftarrow \text{Earliness}(S, y_{risk}, x, D_{test})$ 
4:    $L_{C_k} \leftarrow \emptyset$ 
5:   for each j in  $earlyC_k$  do
6:      $ess_{j_k} \leftarrow \frac{2*(1-j)*stability}{(1-j)+stability}$ 
7:      $L_{C_k} \leftarrow put(L_{C_k}, <j_k, ess_{j_k}>)$ 
8:   end for
9:    $<earliness_{C_k}, ess_{C_k}> \leftarrow best(L_{C_k})$ 
10:   $E_{earliness} \leftarrow put(E_{earliness}, earliness_{C_k})$ 
11: end for
12:  $earliness \leftarrow minimum(E_{earliness})$ 
13: for each  $S_p$  in  $S$  do
14:    $t \leftarrow earliness$ 
15:    $r \leftarrow 0$ 
16:    $alert \leftarrow false$ 
17:   while ( $t \leq T$  and  $r < i$ ) do
18:     if (prediction( $S_p, y_{risk}$ ) == true) then
19:        $r \leftarrow r + 1$ 
20:     else
21:        $r \leftarrow 0$ 
22:     end if
23:      $t \leftarrow t + 1$ 
24:   end while
25:   if ( $r == i$ ) then
26:      $alert \leftarrow true$ 
27:      $A_{S_{alert}} \leftarrow put(S_p)$ 
28:   else
29:      $S_p$  should not be in  $A_{S_{alert}}$ 
30:   end if
31: end for

```

---

dispersed learners. In addition, learning in CNED is quite specific as the registration remains open during the school year. Therefore, the starting activity date  $t_0$  is different from one learner to another. Given these learning particularities, CNED reports yearly high failure rates among its learners. In order to resolve this issue, CNED intends to propose to its teachers an EWS to alert them of at-risk of failure learners. Based on the grades average and according to the French system where marks are out of 20, learners of each module are classified into 3 classes as follows:

- success ( $C_1$ ): when the average is higher than 12
- medium risk ( $C_2$ ): when the average is between 8 and 12
- high risk ( $C_3$ ): when the average is lower than 8

We track the activity of learners on weekly basis in order to generate alerts at the earliest. Thus, on each week, each learner is represented by a vector composed of learning features and the prediction class. In this context,  $y_{risk} = \{C_2, C_3\}$ .

## 5.2 EWS Description

As a proof of concept, we apply the alert generation algorithm on four different EWS exploiting different learning traces that occur regularly over the school year and using various ML models:

1.  $EWS_1$ : this system uses learning traces of 647 learners enrolled in the physics-chemistry module for 37 weeks during 2017–2018 school year.  $EWS_1$  uses the Random Forest (RF) model.
2.  $EWS_2$ : the learning traces of this system are of 647 learners enrolled in the physics-chemistry module for 35 weeks during 2018–2019 school year.  $EWS_2$  uses the Decision Tree (DT) model.
3.  $EWS_3$ : the learning traces of this system are of 679 learners enrolled in the mathematics module for 35 weeks during the 2017–2018 school year.  $EWS_3$  uses the RF model.
4.  $EWS_4$ : this system uses the learning traces of 729 learners enrolled in the history-geography module for 35 weeks during the 2018–2019 school year.  $EWS_4$  uses the DT model.

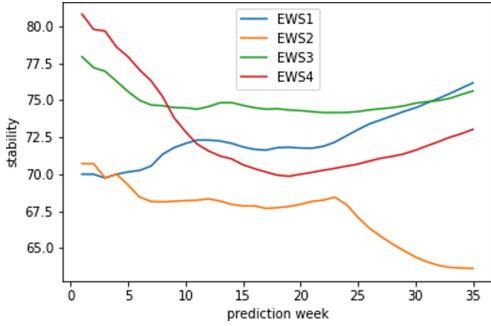
As the temporal stability is an important parameter to generate alerts (Sect. 3.3), the Fig. 3 shows the temporal stability evolution, measured based on the definition given in [12], of the four systems throughout the prediction weeks. The stability of  $EWS_1$  increases slightly over time. It is  $\approx 70\%$  and  $\approx 76\%$  respectively at the first and at the last prediction week. The stability of  $EWS_2$  decreases over time. It starts with  $\approx 70\%$  and is equal to  $\approx 63\%$  at the last prediction week. Both  $EWS_3$  and  $EWS_4$  systems start with a high stability which is respectively equal to  $\approx 78\%$  and  $\approx 80\%$ . However, these values decrease rapidly over the following weeks. The temporal stability values of  $EWS_3$  and  $EWS_4$  at the last prediction week are equal to  $\approx 75\%$  and  $\approx 73\%$ .

## 6 Experiments and Results

In this section, we interpret the results given by the application of the alert Algorithm 1 on each of the described EWS.

### 6.1 Identification of the Starting Time Point $t_s$

The first prerequisite of the alert generation algorithm is the identification of the starting time point  $t_s$  deduced by the calculation of ESS. By applying the ESS formula and all the protocol given in [12], the Table 1 presents, for both class labels  $C_2$  and  $C_3$  of each system, the ESS rate and its corresponding earliness week. Based on the Algorithm 1, for each system, the selected earliness week  $t_s$



**Fig. 3.**  $EWS_1$  VS  $EWS_2$  VS  $EWS_3$  VS  $EWS_4$  in terms of temporal stability

from which the alert generation could be triggered corresponds to the minimum week value between the earliness weeks of  $C_2$  and  $C_3$ . Thus, based on Table 1 the selected earliness for all systems is the week 8. In other words, starting from  $t_s = \text{week } 8$ , we can follow the learners' predictions to apply the alert rules. For other systems using other data and models and developed for other contexts, this starting time point  $t_s$  could be different from week 8.

**Table 1.** ESS of the four EWS and their corresponding earliness time point

	$EWS_1$		$EWS_2$		$EWS_3$		$EWS_4$	
	$C_2$	$C_3$	$C_2$	$C_3$	$C_2$	$C_3$	$C_2$	$C_3$
ESS (%)	68.04	74.73	65.79	73.25	62.88	76.33	75.18	77.31
Earliness (week)	13	8	13	8	16	8	9	8
Selected earliness (week)	8		8		8		8	

## 6.2 Application of Alert Rules Starting from $t_s$

The second part of the alert generation Algorithm 1 consists on applying alert rules on the predictions returned by an EWS. In order to identify the best  $R_i$  for each of our systems ( $EWS_1$ ,  $EWS_2$ ,  $EWS_3$  and  $EWS_4$ ), we study the results of the evaluation metrics (precision and FPR) of the following alerts rules:

- $R_1$ : alert a teacher as soon as a learner is predicted as at-risk
- $R_2$ : alert a teacher when a learner is predicted 2 times in a row as at-risk
- $R_3$ : alert a teacher when a learner is predicted 3 times in a row as at-risk
- $R_4$ : alert a teacher when a learner is predicted 4 times in a row as at-risk
- $R_5$ : alert a teacher when a learner is predicted 5 times in a row as at-risk

These 5 rules are applied starting from  $t_s = \text{week 8}$  as it is defined as the earliness time point of  $EWS_1$ ,  $EWS_2$ ,  $EWS_3$  and  $EWS_4$ . The Table 2 presents the evaluation of these rules on each of the four EWS. For each of these rules, this table shows the total number of generated alerts, the precision (%) of these alerts as well as their FPR (%).

According to the Table 2,  $R_1$  generates always a high number of alerts. However, given the results of FPR and precision, the vast majority of these alerts are false. For example, when applying  $R_1$  on  $EWS_2$  predictions, 89 alerts are generated and only 35.95% of which are correct. Given the instability of EWS results (especially at early prediction times), the application of  $R_1$  cannot ensure precise alerts and a minimum rate of false alerts.

The application of  $R_2$  on EWS generates fewer number of alerts compared to  $R_1$ . These alerts are more precise. This improvement is seen especially with the results of  $EWS_2$ : 62 alerts are generated and 51.61% of which are precise; this is an improvement of 16 points compared to the precision of  $R_1$  with this same system.  $R_2$  gives the best results with  $EWS_1$ : 45 alerts are generated of which 77.77% are correct.

The application of  $R_3$  allows a significant improvement of the precision of the generated alerts. The most important amelioration is observed with both  $EWS_3$  and  $EWS_4$  with precision rates of 79.41% and 51.16% respectively. The  $R_3$  rule considers more the temporal stability of the predictions which explains the significant improvement in the precision values and the decrease in the FPR rates of generated alerts.

The application of  $R_4$  allows a very important enhancement in the results of the generated alerts compared to all previous rules. This improvement is mainly observed with  $EWS_3$  and  $EWS_4$  which precision rates are respectively 85.71% and 65.62%. A slight improvement of 4% and  $\approx 3\%$  is observed with  $EWS_1$  and  $EWS_2$  respectively. The more the rule considers the importance of temporal stability of predictions, more accurate alerts are generated by the system.

The application of  $R_5$  slightly improves the precision of alerts compared to  $R_4$ . For example,  $R_4$  and  $R_5$  with  $EWS_4$  have the same number of generated alerts, precision and FPR. From a certain value of  $i$  of  $R_i$ , the difference in precision and FPR rates is no longer remarkable with those of the rule  $R_{i-1}$ . In such a situation, it is a trade-off to make between the earliness of the alerts and their precision. Indeed, with a high value of  $i$ , alerts are more precise; subsequently, the rate of false alerts is low. However, it is probable to miss out on an earlier generation of good alerts with a lower value of  $i$ .

### 6.3 Discussion

To summarize, the conducted experimentations yielded the following results.

The high number of generated alerts does not mean their effectiveness and precision. Commonly, the successful learners outnumber the failing ones, therefore, it is irrelevant that the EWS generates a very high number of alerts. Otherwise, these alerts are mainly false.

**Table 2.** Application of the five alert rules on each of the four EWS

	EWS <sub>1</sub>		EWS <sub>2</sub>			EWS <sub>3</sub>			EWS <sub>4</sub>			
	Generated alerts	Precision (%)	FPR (%)	Generated alerts	Precision (%)	FPR (%)	Generated alerts	Precision (%)	FPR (%)	Alerts number	Precision (%)	FPR (%)
<i>R</i> <sub>1</sub>	53	66.96	33.96	89	<b>35.95</b>	64.04	48	62.5	37.5	72	<b>31.94</b>	68.05
<i>R</i> <sub>2</sub>	45	<b>77.77</b>	22.22	62	<b>51.61</b>	48.38	42	69.04	30.95	57	38.59	61.40
<i>R</i> <sub>3</sub>	43	81.39	18.60	54	59.25	40.74	34	<b>79.41</b>	20.58	43	<b>51.16</b>	48.83
<i>R</i> <sub>4</sub>	41	85.36	14.63	51	62.74	37.25	28	<b>85.71</b>	14.28	32	<b>65.62</b>	34.375
<i>R</i> <sub>5</sub>	39	87.17	12.82	46	67.39	32.60	24	87.5	12.5	32	<b>65.62</b>	34.375

*R*<sub>1</sub> is always insufficient for a precise alert generation. Applying *R*<sub>1</sub> generates a large number of false alerts. Indeed, at early prediction times, EWS are likely to provide false predictions by identifying successful learners as belonging to risk classes and vice versa. In addition, *R*<sub>1</sub> does not consider the predictions stability which explains the high FPR rates. Starting from *R*<sub>2</sub>, the alerts are more precise and the FPR rates decrease.

The less precise generated alerts are observed with the least stable systems. For example, the temporal stability of EWS<sub>2</sub> and EWS<sub>4</sub> decreases over time (see Fig. 3). All the rules applied on these two systems are characterized by low precisions. Thus, the appropriate value of *i* of *R*<sub>*i*</sub> to apply is different from one system to another and closely depends on its temporal stability. In addition, the value of *i* should ensure the compromise between the importance and effectiveness of early alerts and their precision. Moreover, some systems don't satisfy the condition of *i* successive predictions in a risk class when *i* is high. Hence, we can probably miss out on good alerts. In addition, from a certain value of *i*, the precision of *R*<sub>*i*</sub> is slightly higher than *R*<sub>*i*-1</sub>. In such a case, it may be more relevant to consider the earliness of alerts rather than the succession of high number *i* of predictions and opt for *R*<sub>*i*-1</sub>.

## 7 Conclusion

In this paper, we proposed an algorithm for generating early and precise alerts to enrich the use of EWS by the teachers. To this end, this algorithm uses the temporal metrics and introduces the alert rule concept. Indeed, the algorithm starts by identifying the starting time point *t*<sub>*s*</sub> from which the application of an alert rule *R*<sub>*i*</sub> can be triggered. Then, the generated alerts are evaluated based on two metrics: precision and FPR. In order to validate the relevance of this algorithm, we applied it on four different EWS using real learning traces of k-12 learners enrolled within online modules.

The conducted experiments showed that a high number of generated alerts doesn't reflect their precision. Indeed, the higher *i* of *R*<sub>*i*</sub> is, the more precise are the generated alerts. However, this optimal value of *i* is different from one system to another and closely depending on the temporal stability of each EWS. Moreover, the higher *i* is, the later in the year the alert is generated. Thus, it is a compromise to make between the importance of early alerts and their precision.

As perspectives, we intend to extend this work by improving this alert generation algorithm to consider the prediction oscillations characterizing an EWS. In addition, we aim to work on suggesting a *trust indicator* that illustrates the trust level that a teacher could have toward the EWS generated alerts.

## References

1. Akçapinar, G., Hasnine, M.N., Majumdar, R., Flanagan, B., Ogata, H.: Developing an early-warning system for spotting at-risk students by using ebook interaction logs. *Smart Learn. Environ.* **6**(1), 1–15 (2019)
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 (2012)
3. Bañeres, D., Karadeniz, A., Guerrero-Roldán, A.E., Rodríguez-Gonzalez, M.E., Serra, M.: Analysis of the accuracy of an early warning system for learners at-risk: a case study. In: Proceedings of the 11th International Conference on Education and New Learning Technologies-EDULEARN19, Palma, Spain, pp. 1–3 (2019)
4. Bañeres, D., Rodríguez, M.E., Guerrero-Roldán, A.E., Karadeniz, A.: An early warning system to detect at-risk students in online higher education. *Appl. Sci.* **10**(13), 4427 (2020)
5. Bañeres, D., Rodríguez-González, M.E., Guerrero-Roldán, A.E., Cortadas, P.: An early warning system to identify and intervene online dropout learners. *Int. J. Educ. Technol. High. Educ.* **20**(1), 1–25 (2023)
6. Cano, A., Leonard, J.D.: Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Trans. Learn. Technol.* **12**(2), 198–211 (2019)
7. Cohen, A.: Analysis of student activity in web-supported courses as a tool for predicting dropout. *Educ. Tech. Res. Dev.* **65**(5), 1285–1304 (2017)
8. Jokhan, A., Sharma, B., Singh, S.: Early warning system as a predictor for student performance in higher education blended courses. *Stud. High. Educ.* **44**(11), 1900–1911 (2019)
9. Liz-Domínguez, M., Rodríguez, M.C., Nistal, M.L., Mikic-Fonte, F.A.: Predictors and early warning systems in higher education-a systematic literature review. In: LASI-SPAIN, pp. 84–99 (2019)
10. López Zambrano, J., Lara Torralbo, J.A., Romero Morales, C., et al.: Early prediction of student learning performance through data mining: a systematic review. *Psicothema* (2021)
11. Soussia, A.B., Labba, C., Roussanaly, A., Boyer, A.: Assess performance prediction systems: Beyond precision indicators. In: CSEDU (1), pp. 489–496 (2022)
12. Soussia, A.B., Labba, C., Roussanaly, A., Boyer, A.: Time-dependent metrics to assess performance prediction systems. *Int. J. Inf. Learn. Technol.* (ahead-of-print) (2022)
13. Soussia, A.B., Roussanaly, A., Boyer, A.: Toward an early risk alert in a distance learning context. In: 2022 International Conference on Advanced Learning Technologies (ICALT), pp. 206–208. IEEE (2022)
14. Wang, Z., et al.: Design and implementation of early warning system based on educational big data. In: 2018 5th International Conference on Systems and Informatics (ICSAI), pp. 549–553. IEEE (2018)



# Detecting Interlingual Errors: The Case of Prepositions

Natawut Monaikul<sup>(✉)</sup> and Barbara Di Eugenio<sup>ID</sup>

University of Illinois Chicago, Chicago, IL 60607, USA  
`{monaiku1,bdieugen}@uic.edu`

**Abstract.** Prepositions pose a particular challenge for many language learners, in part because of their seemingly arbitrary usage with certain verbs that does not necessarily translate directly across languages. As a consequence, many preposition errors in writing can be attributed to a direct transfer from a writer’s native language. While research in scalable tooling for second-language writing assistance has largely focused on automated error detection and correction, relatively little attention has been given to *explaining* why the errors may have occurred. A system that can distinguish *interlingual* errors – arising from an over-literal translation from a writer’s native language – from non-interlingual errors could serve as an insightful tool for language learners and teachers alike. In this work, we demonstrate the feasibility of classifying English preposition errors produced by native speakers of Spanish as interlingual. We propose a corpus-based method that exploits translation probabilities to estimate the likelihood that a writer has translated word-for-word from Spanish to English. We then show that this method correlates well with human judgments on the interlingual status of preposition errors and can be a basis for developing a tool for explaining one key source of errors in second-language writing.

**Keywords:** Second language learning · Grammatical error detection · ESL · Computer-assisted language learning · Preposition errors

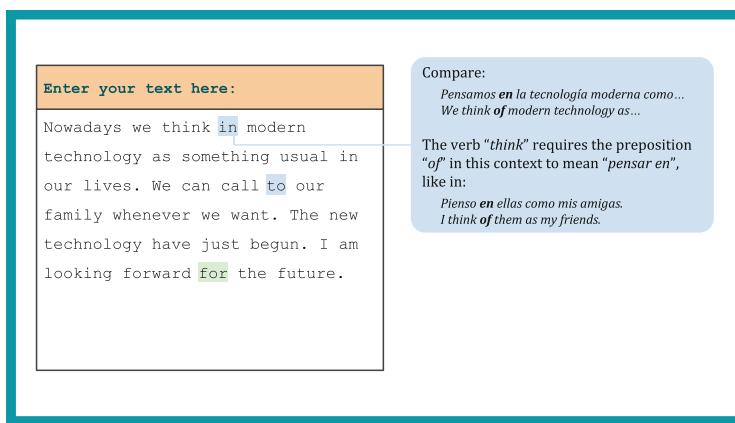
## 1 Introduction

Learning a language that uses prepositions comes with the challenge of learning to choose appropriate prepositions in various contexts, many of which are fixed expressions without generalizable rules and cannot be translated by considering only the words in isolation. For example, the English expression “dream **about** something” typically translates to *soñar con algo* in Spanish and *fan theung arai* in Thai, where *con* and *theung* typically translate to “with” and “to” in English, respectively. This mismatch in preposition usage can manifest as *interlingual* errors in second language writing, i.e., errors that arise from an “over-literal or word-for-word translation” from a native language (L1) to a target second language (L2) [9]. Indeed, the appearance of interlingual errors with

prepositions that co-occur with particular verbs has been documented in many analyses of L2 writing, in a variety of source and target languages [3, 17, 19].

Researchers in language pedagogy have advocated for language teachers to investigate differences in structure and usage between the native and target languages of their students so that targeted corrective feedback that references their students' native languages can be provided [7, 11]. This is further reinforced by studies that have demonstrated the effectiveness of delivering corrections accompanied by explanations contrasting the students' L1 and L2 on the target concept [10, 21]. However, a teacher unfamiliar with a student's L1 may not necessarily be able to provide these insights. A tool that can automatically detect when an interlingual error has been committed could then assist a teacher in understanding the source of the error when they are unfamiliar with a student's L1 and in formulating a plan for targeting the error. Such a tool could also serve students at scale if it is equipped with a feedback generation mechanism based on the interlingual status of an error.

Though the use of intelligent writing assistants for L2 writing such as *Grammarly* may be employed as a preliminary grammar or style checker, in the case of interlingual errors, these systems do not offer *contrastive* feedback that references the writer's L1 to explain a concept in the L2. Grammatical error correction in the field of natural language processing (NLP) has received a great deal of attention [4, 12], but we seek to complement corrections with more targeted feedback. In particular, we envision an intelligent writing assistant with an interface similar to that given in Fig. 1, where interlingual errors are distinguished from non-interlingual errors, and contrastive feedback is provided with examples specifically for interlingual errors.



**Fig. 1.** An example user interface for an intelligent writing assistant that provides contrastive feedback specifically for interlingual errors.

This work addresses a necessary step in delivering such feedback: automatically determining if a given error is interlingual. In this work, we target errors

concerning prepositions that co-occur with verbs. We propose a scalable method for collecting annotations on whether or not a given preposition error is interlingual. We then describe a statistical method for estimating the likelihood that a given preposition error was directly influenced by the writer’s L1 using information extracted from a large parallel corpus. We show that in an L1 Spanish-L2 English setting, we are able to generate predictions that are in agreement with human annotators and that can serve as a useful feature for classifying preposition errors as interlingual or not.

## 2 Related Work

Kim [10] studied different kinds of feedback in teaching about the passive voice in an intelligent computer-assisted language learning (ICALL) system for L1 Korean-L2 English students. Because Korean allows some verbs to be passivized that English does not, interlingual errors such as *are consisted of* instead of *consisted of* or *be happen* instead of *happen* are not uncommon. The study found that contrastive feedback that compared Korean and English sentences side-by-side was more effective than simple corrective feedback in promoting consistent improvement and long-term retention of the ability to identify and correct these types of errors. The ICALL system in the study contained pre-written exercises and manually-constructed feedback (delivered automatically); providing contrastive feedback on interlingual errors in free text, as in Fig. 1, would require machine learning-based components to handle arbitrary input. In particular, automatically classifying an error as interlingual or not is the first step towards enabling an intelligent writing assistant to deliver contrastive feedback in the appropriate contexts. To the best of our knowledge, our work is the first to explicitly address this classification task computationally.

Swan and Smith [18] developed a reference for English teachers that compares various languages with English to describe structures that may be particularly challenging for learners of English with those language backgrounds. In the same vein, Graën and Schneider [8] proposed a statistical method for automatically compiling lists of verb-preposition pairs that could be difficult for English language learners due to the mismatch in preposition usage between their L1s and English. In their work, a multilingual parallel corpus was mined for verb-preposition pairs and their translations to calculate a score that represented how likely an incorrect preposition would be chosen in a word-for-word translation. This value was then used to suggest preposition errors that are likely to be produced by an L2 English speaker with a given language background; however, these suggestions were not specifically evaluated for interlingual status.

Many previous studies that analyze L2 writing for the presence of interlingual errors [3, 17, 19] carry out analyses manually and provide examples of the interlingual errors that are found, but training a classifier to perform this task requires a sizable number of examples. This is also constrained by the L2 writing corpora available – annotating an interlingual error requires the error to be produced in the first place, and certain errors that are possible may not appear

in these corpora if they are not somehow elicited through the writing prompts used to gather the data. Alonso Ramos et al. [13] have made significant strides on this front by developing an annotation framework for collocation errors in L2 Spanish writing, which includes annotations at the “explanatory” level, where errors are categorized as interlingual or *intralingual* (i.e., errors that arise due to a particular difficulty in the L2 and not from over-literal translations). Our work explores a scalable method of data collection with an eye towards building an interlingual error classifier and, ultimately, a contrastive feedback generator.

### 3 Data Collection

#### 3.1 Validation Study

Studies in which interlingual errors in L2 writing are labeled and analyzed are typically carried out by researchers with backgrounds in linguistics or language teaching who are familiar with both the L2 and the L1. We posit that general native speakers of the L1 who are familiar with the L2 can also reliably distinguish interlingual from non-interlingual errors, drawing from their own or others’ experiences in learning the L2. To validate this, we conducted a study that elicited the intuitions of L1 Spanish-L2 English speakers as to why certain preposition errors may have occurred.

Six L1 Spanish-L2 English speakers were recruited, five of whom had no prior experience in teaching languages, and one of whom had experience in teaching languages and a background in linguistics. Participants read 20 English sentences that were written by L1 Spanish speakers and that contained an error involving a verb-preposition pair. Participants were then asked to explain, for each sentence, why they believe the marked preposition error had occurred, given that the writer is a native Spanish speaker. The 20 sentences were randomly selected from a dataset of error-annotated L2 English essays – written as part of the Cambridge ESOL First Certificate in English (FCE) exam [22] – such that each sentence contained a unique error involving preposition usage with a verb.

Because participants gave open-ended responses, we also labeled each response as either *interlingual* if the response suggested a word-for-word translation from Spanish, or *non-interlingual* otherwise. Two researchers independently labeled the 120 responses, and Cohen’s kappa coefficient for inter-annotator agreement was calculated to be  $\kappa = 0.87$ . This strong agreement indicated that the participants’ responses could be reliably interpreted for whether or not they suggested that a given error was interlingual. Examples of responses and their labels are given in Table 1. Disagreements were discussed and resolved to arrive at final labels for each response. Comparing the majority vote of the five participants (with no prior language teaching experience) on the preposition errors with the labels of the one experienced participant revealed agreement on 19 of the 20 errors. Notwithstanding the small sample size of errors and participants (considering the wide varieties of Spanish that are spoken), these results suggest that L1 Spanish-L2 English speakers, regardless of background in language teaching, can reliably annotate L2 English preposition errors for interlingual (when labels

**Table 1.** A sample of responses in our validation study. From left to right, we have the sentence participants read (with the target verb-preposition pair bolded), a response from a participant, and the label assigned to the response.

Sentence	Response	Label
“...so I <b>apologize</b> with her...”	“Translating word by word: ‘Me disculpo CON ella’”	interlingual
“... <b>calling</b> to radio stations...”	“Thinking in Spanish, the writer must specify who is calling to by starting with a preposition”	interlingual
“...when they have to <b>pay for</b> the bills!!”	“You normally learn that you pay for something...so they might have assumed that ‘for’ goes after pay all the time”	non-interlingual
“I need to <b>insist in</b> it...”	“Again, in Spanish ‘in’ and ‘on’ mean ‘en’. It is hard to choose between the two of these”	non-interlingual

are aggregated across multiple annotators) with respect to an L1 of Spanish. We note that these findings are likely influenced by the similarities in preposition usage between English and Spanish; further investigation is needed to determine if these findings can generalize to more dissimilar language pairs.

### 3.2 Annotation Crowdsourcing

With the annotation process validated, we would like to scale to a wider pool of annotators to obtain more annotations on larger datasets, for which we turn to crowdsourcing. Given the amount of time and cognitive effort it takes to provide free responses (about one hour for 20 sentences in the validation study), we converted the task to a multiple-choice annotation. Additionally, while the FCE dataset provides preposition errors that have actually been produced by L2 English speakers, it is only a subset of preposition errors that could be produced. Thus, we created a synthetic corpus of sentences with preposition errors. We selected 38 verbs of varying frequencies (according to the iWeb corpus of web text [1]) and created sentences with an intended preposition usage. For each verb, we selected three to five other prepositions (including *null*). This selection included any prepositions that could potentially result from a word-for-word translation from Spanish when consulting a bilingual dictionary, as well as other randomly selected prepositions that, if used in the corresponding sentence, would be ungrammatical or not reasonably similar semantically to the original sentence.

In total, 155 sentences containing preposition errors were created. Examples of sentences in this dataset are given in Table 2. Because the errors were semi-randomly generated, it is likely that some errors are not ones that would be naturally produced by an L1 Spanish-L2 English speaker. Annotators were therefore given three options for each sentence:

**Table 2.** A sample of sentences we created for our dataset. The first column is the intended sentence, and the second contains the prepositions that were instead used in the sentence (e.g., from the first sentence, five sentences were generated for annotation).

Sentence	Incorrect Prepositions
The child is <i>laughing at</i> the cartoon	of/for/to/on/ <i>null</i>
The customer <i>asked for</i> a drink	to/on/of/ <i>null</i>
Some people <i>think of</i> their pets as children	on/in/from/for/to
She <i>hugged her mother</i>	to/at/with

- “The writer was directly translating from Spanish”.
- “There is some other grammatical reason for this error”.
- “A native Spanish speaker would not make this error”.

The first option corresponded with labeling the error as *interlingual* and the second as *intralingual*. Examples of each option were also provided. Annotators could also optionally leave a comment after each sentence. The annotation task was deployed on the Amazon Mechanical Turk (AMT) platform in countries in which Spanish is an official language, and the task was deployed in four mutually-exclusive batches of 39 sentences<sup>1</sup>. Annotators were each compensated \$10 per completed batch. Nine annotators were recruited for each batch, though some annotators responded to more than one batch, resulting in 24 total annotators. If we again assign a label to each sentence based on the majority vote for that sentence, the annotated dataset consists of 43 interlingual errors, 25 intralingual errors, and 45 errors that seem unlikely to be produced (42 errors did not have a majority among the three options).

We note that the country filter does not necessarily guarantee L1 Spanish speakers, but it is the most reasonable filter provided (without cost) to try to recruit only L1 Spanish speakers. There was also no control for L2 English ability, other than presenting the task completely in English. We therefore chose a larger number of annotators per sentence to account for potential low-quality annotations – we assume that aggregating these responses will give more reliable labels. Half of the annotators did leave comments on multiple sentences, many of which were similar to the responses given in the validation study, and the average completion time was around 30 min, suggesting that at least half of the annotators were adequately engaged in the task and provided best-effort annotations. We plan to explore methods for controlling for these factors, such as including a questionnaire or comparing a subset of responses against a validated ground truth set, to ensure more reliable responses are collected in a scalable way.

---

<sup>1</sup> One sentence was discarded after responses were collected due to an error in the set-up of the annotation task.

## 4 Methods

In this section, we introduce a method for estimating the probability that an erroneous verb-preposition pair resulted from a word-for-word translation from a source L1 to a target L2. This method attempt to formalize an intuition in predicting if an error is interlingual: is there an expression in the writer's L1 that could explain the error, i.e., that the writer could have translated from to produce the error? We note that this method may generalize to other languages, but that the language pairs are assumed to use prepositions (or post-positions) that can generally be translated in isolation from one language to the other.

To estimate probabilities of translating words between the L1 and L2, we use a parallel corpus that has been aligned at the word level, i.e., the words of a sentence in one language are matched with the words of the corresponding sentence in the other language. Given a large enough corpus, this alignment can provide approximate translation probabilities. Suppose we have an erroneous verb-preposition pair  $v_{L2}, p_{L2}$  in the target language. Note that  $p_{L2}$  may be *null*, i.e., the error is that a preposition should have been used, but one was not used. Let  $p_{L2}^{corr}$  be the intended preposition (or, a preposition that could be suggested as a correction for  $p_{L2}$ ), which may also be *null*.

We would first like to find possible verb-preposition pairs in the source language that the writer could have been thinking of given that they intended to use  $v_{L2}, p_{L2}^{corr}$ . We consult the parallel corpus to find instances of  $v_{L2}, p_{L2}^{corr}$  and trace their word alignments to find potential corresponding verb-preposition pairs in the L1. We then estimate the probability that the writer was thinking of the verb-preposition pair  $v_{L1}, p_{L1}$  as

$$P(v_{L1}, p_{L1}|v_{L2}, p_{L2}^{corr}) = \frac{c(\{v_{L2}, p_{L2}^{corr}\}, \{v_{L1}, p_{L1}\})}{\sum_{v'_{L1}, p'_{L1} \in V} c(\{v_{L2}, p_{L2}^{corr}\}, \{v'_{L1}, p'_{L1}\})}, \quad (1)$$

where  $c(X, Y)$  denotes the frequency with which the set of words  $X$  aligns with the set of words  $Y$  in the corpus, and  $V$  is the set of verb-preposition pairs in the L1 that  $v_{L2}, p_{L2}^{corr}$  were found to align with in the corpus. In other words, we represent the probability that the writer was thinking of  $v_{L1}, p_{L1}$  given that they intended to use  $v_{L2}, p_{L2}^{corr}$  by the relative corpus-derived frequency with which  $v_{L1}, p_{L1}$  and  $v_{L2}, p_{L2}^{corr}$  align.

For each possible verb-preposition pair  $v_{L1}, p_{L1}$ , we also calculate the probability that  $p_{L1}$  translates back to the incorrect preposition  $p_{L2}$ . We similarly estimate this using the relative frequency with which  $p_{L1}$  aligns with  $p_{L2}$ , or

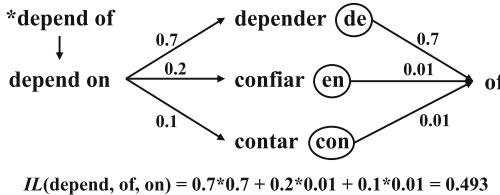
$$P(p_{L2}|p_{L1}) = \frac{c(\{p_{L1}\}, \{p_{L2}\})}{\sum_{w_i \in W} c(\{p_{L1}\}, \{w_i\})}, \quad (2)$$

where  $W$  is the set of all words  $p_{L1}$  is found to align with in the parallel corpus. We also define  $P(p_{L2}|p_{L1})$  to be 0 if exactly one of  $p_{L2}$  or  $p_{L1}$  is *null* and 1 if both  $p_{L2}$  and  $p_{L1}$  are *null*. In other words, *null* always translates to *null* – the absence of a preposition in one language never translates to using a preposition in the other when considering word-by-word translations.

Finally, we define the *interlingual likelihood* (IL) of the verb-preposition pair  $v_{L2}, p_{L2}$  given the correction  $p_{L2}^{corr}$  as

$$IL(v_{L2}, p_{L2}, p_{L2}^{corr}) = \sum_{v'_{L1}, p'_{L1} \in V} P(v'_{L1}, p'_{L1} | v_{L2}, p_{L2}^{corr}) P(p_{L2} | p'_{L1}), \quad (3)$$

i.e., the sum of the probabilities of all possible translation paths (from L2 to L1 and back to L2) from the intended verb-preposition pair to the erroneous verb-preposition pair. This calculation is depicted in Fig. 2 for the incorrect verb-preposition pair “depend of” produced by an L1 Spanish speaker.



**Fig. 2.** An example of how IL is calculated, where  $v_{L2}$  is “depend”,  $p_{L2}$  is “of”, and  $p_{L2}^{corr}$  is “on”. The probability of a path uses the relative frequency with which “depend on” translates to a Spanish verb-preposition pair and the Spanish preposition translates back to “of”. The interlingual likelihood is then the sum of these paths.

Ideally, preposition errors with higher values of IL indicate a stronger chance of being interlingual errors. Given enough data, a threshold on the IL of a verb-preposition pair could be empirically determined and set to classify the error as interlingual or non-interlingual.

## 5 Evaluation

### 5.1 Procedure

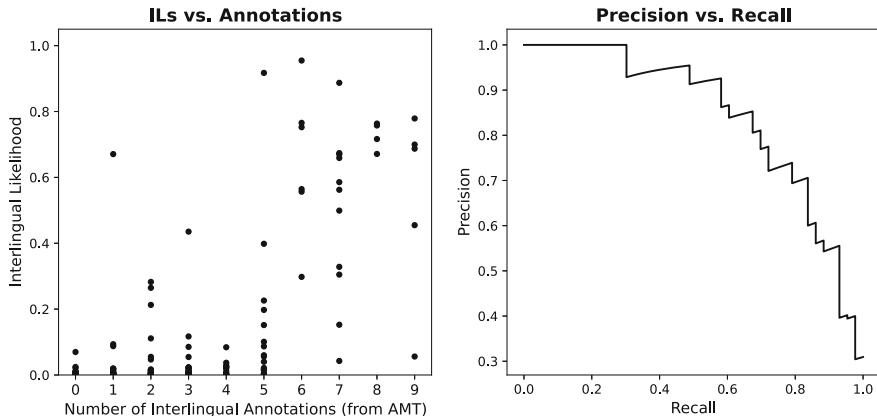
In order to measure the extent to which IL agrees with the labels collected on AMT, we implemented our proposed method of calculating the IL of each error in our dataset. We used the CCMatrix corpus of mined multilingual web text that has been automatically aligned for many language pairs [6, 16] as the parallel corpus, accessed via the OPUS repository [20]. The sentence pairs in the corpus are provided in decreasing order of LASER embedding similarity scores [14, 15], so the first 5M English-Spanish sentence pairs in which both sentences were between 4 and 20 tokens were extracted. The sentences were parsed and lemmatized using the English and Spanish models in the spaCy toolkit [2], and word alignment was performed using *fast\_align* [5]. Frequencies were tabulated only on the lemmas of each word.

We hypothesize that errors with more interlingual labels will have higher ILs than those with fewer interlingual labels, which can be tested with the Spearman

correlation coefficient. We also compute the precision and recall of a simple classifier that uses a threshold on the IL of an error to determine the label for the error. For this, we assign to each error the label *interlingual* if a majority of the annotations for that error are interlingual, and *non-interlingual* otherwise. Because only 43 of the 155 sentences are labeled as *interlingual*, we do not have a large enough dataset to split the data and empirically determine a threshold, so we record the classifier's performance at various thresholds, plot a precision-recall curve (which shows the precision of a classifier at each possible recall), and measure the area under the curve (AUC), where an AUC closer to 1 indicates better performance over all thresholds tested.

## 5.2 Results

Figure 3 on the left shows the plot of the computed IL of each error against the number of annotators who labeled the error as *interlingual*. We found that most of the errors with fewer than five *interlingual* labels have ILs below 0.2, while most of the errors with at least five have ILs above 0.2, which can be observed in the plot. The correlation was calculated to be  $\rho = 0.67$ , indicating a reasonably strong correlation between IL and human judgments.



**Fig. 3.** Computed IL for all errors vs. the number of annotators who labeled the error as *interlingual* (left) and the precision-recall curve of classifying errors as interlingual based on IL (right).

Figure 3 on the right shows the precision-recall curves of a simple classifier whose only feature is the IL of each error, with an AUC of 0.84. These results taken together suggest that our definition of interlingual likelihood roughly agrees with L1 Spanish-L2 English annotators on preposition errors and can be used to distinguish between interlingual and non-interlingual errors. In particular, setting the threshold on ILs for assigning labels to errors at 0.2, we

achieve a precision of 83.9% at a recall of 60.5%. We plan to collect more data, in this and other language pairs, to be able to train a classifier with IL as a feature to corroborate these results.

We investigated the preposition errors that exhibited low IL but received a majority *interlingual* label. One example is from the sentence “I decided *for* a new book” (where the intended usage is “decided *on*”). All nine annotators for this sentence labeled the error as *interlingual*, but the IL of the error was only 0.056. We found that in the corpus, there were many instances of “decide *on*” aligning with Spanish *decidir por*, but many more with *decidir* (without a preposition). Both usages are valid in Spanish, notwithstanding dialectal differences. If an annotator is in the mindset of first seeing if an interlingual explanation is possible for an error before ruling it out, then they may decide that an error is interlingual if *any* valid Spanish sentence can explain it, even if it is not how they themselves would necessarily say it. IL, on the other hand, is strongly influenced by verb-preposition usage frequency and will be low if there are not enough supporting examples in the corpus, even if those examples are valid.

Another example is from the sentence “Their kids play *to* basketball” (where the intended usage is “play” without a preposition), for which the IL is low despite a majority label of *interlingual*. One contributor we found is Spanish *jugar a* (in the sense of “play a game”) getting overshadowed by Spanish *tocar* without a preposition (in the sense of “play an instrument”). The current calculation of IL takes into consideration all possible translations of a verb-preposition pair in a corpus, but it does not consider how the context in the given sentence disambiguates, which can make some preposition usages more likely than others. We plan to experiment with different definitions of IL and other features to account for these cases.

One important limitation of this evaluation is with the “in/on” distinction that is notoriously difficult for L1 Spanish speakers learning English. While choosing the incorrect one was not marked as an interlingual error in the validation study, the IL would be high if the Spanish translation of the verb-preposition pair uses *en*, which can translate as either “in” or “on”. No such instances appeared in the dataset annotated on AMT. Our future work seeks to address this type of intralingual error, potentially by first mining for pairs of prepositions in the target language that may easily be confused due to similarities in usage and form without a clear mapping to the source language.

## 6 Conclusions

An intelligent writing assistant that provides more targeted feedback based on whether or not a detected error is interlingual would be a useful tool for language learners and language teachers alike. A necessary component of this is predicting if a given error is likely to be interlingual, something that has not yet been addressed in NLP or language pedagogy research. In this paper, we demonstrated the possibility of building datasets of L2 errors with annotations of interlingual status with respect to an L1 from native speakers of that L1. We

proposed a definition of the *interlingual likelihood* of errors involving prepositions that co-occur with verbs, by using statistical properties of a parallel corpus. We presented evidence that the defined IL in practice correlates with the judgments of L1 Spanish-L2 English speakers and can accurately predict when preposition errors resulted from a word-for-word translation. Extensions of this work include gathering and evaluating on larger datasets and in different language pairs. We may also explore how this definition of interlingual likelihood can be extended to adjective-preposition pairs (e.g., “interested *in*” or “happy *with*”) and even verb-noun collocations.

## References

1. iWeb: The 14 billion word web corpus. <https://www.english-corpora.org/iweb/>. Accessed 31 Dec 2021
2. spaCy: Industrial-strength natural language processing. <https://spacy.io/>. Accessed 19 Jan 2022
3. Alonso, M.R.A.: Language transfer: interlingual errors in Spanish students of English as a foreign language. *Rev. Alicantina Estud. Ingleses* **10**, 7–14 (1997)
4. Bryant, C., Felice, M., Andersen, Ø.E., Briscoe, T.: The BEA-2019 shared task on grammatical error correction. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 52–75. Association for Computational Linguistics, Florence (2019). <https://doi.org/10.18653/v1/W19-4406>
5. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 644–648. Association for Computational Linguistics, Atlanta (2013)
6. Fan, A., et al.: Beyond English-centric multilingual machine translation. *J. Mach. Learn. Res.* **22**(107), 1–48 (2021)
7. Ferris, D.R.: Treatment of Error in Second Language Student Writing. The Michigan Series on Teaching Multilingual Writers, 2nd edn. The University of Michigan Press, Ann Arbor (2011)
8. Graën, J., Schneider, G.: Crossing the border twice: reimporting prepositions to alleviate L1-specific transfer errors. In: Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, pp. 18–26. LiU Electronic Press, Gothenburg (2017)
9. James, C.: Errors in Language Learning and Use. Applied Linguistics and Language Study, Addison Wesley Longman, New York (1998)
10. Kim, D.H.: Explicitness in CALL feedback for enhancing advanced esl learners' grammar skills. Ph.D. thesis, University of Illinois at Urbana-Champaign (2009)
11. Koban, D.: A case study of Turkish ESL learners at LaGuardia Community College, NYC error analysis. In: Dan, C. (ed.) Languages, Literature, and Linguistics. International Proceedings of Economics Development and Research, vol. 26. IAC-SIT Press (2011)
12. Ng, H.T., Wu, S.M., Wu, Y., Hadiwinoto, C., Tetreault, J.: The CoNLL-2013 shared task on grammatical error correction. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–12. Association for Computational Linguistics, Sofia (2013)

13. Ramos, M.A., et al.: Towards a motivated annotation schema of collocation errors in learner corpora. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta (2010)
14. Schwenk, H.: Filtering and mining parallel data in a joint multilingual space. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp. 228–234. Association for Computational Linguistics, Melbourne (2018). <https://doi.org/10.18653/v1/P18-2037>
15. Schwenk, H., Li, X.: A corpus for multilingual document classification in eight languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki (2018)
16. Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A., Fan, A.: CCMATRIX: mining billions of high-quality parallel sentences on the web. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 6490–6500. Association for Computational Linguistics (2021). <https://doi.org/10.18653/v1/2021.acl-long.507>
17. Sumonsriwakun, P., Pongpairoj, N.: Systematicity of L1 Thai learners' English interlanguage of dependent prepositions. *Indon. J. Appl. Linguist.* **6**(2), 246–259 (2017)
18. Swan, M., Smith, B.: Learner English: A Teacher's Guide to Interference and Other Problems. Cambridge Handbooks for Language Teachers, 2nd edn. Cambridge University Press, Cambridge (2001). <https://doi.org/10.1017/CBO9780511667121>
19. Tarnaoui, M.M.: Analyse contrastive FLE/ Tachelhit: le cas des prépositions diagnostique des difficultés et remédiations didactiques. *Stud. Gramatică Contrastivă* **30**, 69–81 (2018)
20. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 2214–2218. European Language Resources Association (ELRA), Istanbul (2012)
21. Tomasello, M., Herron, C.: Feedback for language transfer errors: the garden path technique. *Stud. Second. Lang. Acquis.* **11**(4), 385–395 (1989)
22. Yannakoudakis, H., Briscoe, T., Medlock, B.: A new dataset and method for automatically grading ESOL texts. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 180–189 (2011)



# Automated Extraction of Domain Models from Textbook Indexes for Developing Intelligent Tutoring Systems

Rabin Banjade<sup>(✉)</sup>, Priti Oli, and Vasile Rus

University of Memphis, Memphis, USA  
`{rbanjade1, poli, vrus}@memphis.edu`

**Abstract.** Domain modeling is an important task in designing, developing, and deploying intelligent tutoring systems and other adaptive instructional systems. We focus here on the more specific task of automatically extracting a domain model from textbooks. In particular, this paper explores using multiple textbook indexes to extract a domain model for computer programming. Our approach is based on the observation that different experts, i.e., authors of intro-to-programming textbooks in our case, break down a domain in slightly different ways, and identifying the commonalities and differences can be very revealing. To this end, we present automated approaches to extracting domain models from multiple textbooks and compare the resulting common domain model with a domain model created by experts. Specifically, we use approximate string-matching approaches to increase coverage of the resulting domain model and majority voting across different textbooks to discover common domain terms related to computer programming. Our results indicate that using approximate string matching gives more accurate domain models for computer programming with increased precision and recall. By automating our approach, we can significantly reduce the time and effort required to construct high-quality domain models, making it easy to develop and deploy tutoring systems. Furthermore, we obtain a common domain model that can serve as a benchmark or skeleton that can be used broadly and adapted to specific needs by others.

**Keywords:** domain modeling · computer programming · intelligent tutors system · computer science education · Java programming

## 1 Introduction

When developing an intelligent tutoring system (ITS), it is imperative to understand what students need to master in the target domain. This requires an effective representation of the domain, which includes the key concepts or knowledge components (KCs) and a structure that specifies a prerequisite structure over those KCs. This representation of the domain is the domain model, and it plays a key role in ITSs. Indeed, these intelligent tutors rely on the domain model to

select what tasks the student should work on and to provide relevant support, e.g., scaffolding in the form of hints or correcting a misconception when students work on a task. While our main goal was to develop an ITS for code comprehension, we learned that there is no off-the-shelf, ready-to-use domain model for intro-to-programming topics typically covered in CS1 and CS2 courses. To the best of our knowledge, there is no domain model for intro-to-programming that covers the vast majority of topics and is at the right level of granularity for our purposes. Furthermore, there is no widely accepted domain model that anyone can reference and use as a baseline [13, 21].

In sum, the work in this paper was driven by two main reasons: (i) our broader goal of developing an Intelligent Tutoring System (ITS) for intro-to-programming and (ii) the need for a domain model for intro-to-programming that is publicly available and ready to use, i.e., it can serve as a benchmark or skeleton that can be used broadly and adapted to specific needs by others.

Typically, the design and evaluation of the domain model for a given domain is expert driven, which is both tedious and expensive. Such manual approaches hinder the development of ITSs for new domains and, therefore cross-domain scalability of ITSs. Automating the extraction of domain models will address these challenges. To extract domain models, various sources can be used, such as textbooks that represent authors' perspectives of the targeted domain [17]. It should be noted that the authors are both experts in the domain and pedagogical experts, i.e., they do have experience teaching about the target domain, which makes them familiar with, for instance, the typical misconceptions and more effective ways to teach about a particular concept. Thus, extracting domain models from textbooks is based on solid expertise.

Textbooks are mostly free text but contain other, more structured elements such as table of contents, chapters, sections, etc. Of particular interest to our work is the index at the back of a typical textbook which serves as an important navigational tool that links the important domain terms curated by the expert to specific parts of the textbook. That is, entries in such textbook indexes are a rich source of domain terminology. The index at the end of a textbook is created mostly manually, even though some automated methods have been studied [24]. Index terms are not just a collection of words but a reference model based on predefined rules. Although index terms can be a source of curated domain terms, from an epistemological perspective, an index of any document reflects not only the expertise and efforts of its creator but also the needs of the group of users for whom the index is created and the task that these users engage in [9]. This can lead to index terms that might not be fully representative of the domain. Hence, extracting the domain model from the textbooks using the indexes may not necessarily lead to a clean domain model. Also, not all the index terms are useful for learning or teaching and may contain various useful entries about the history of a domain. Such examples, in our case, ranged from method names to names of persons (e.g., Gosling James, the creator of Java). If the goal is to create a reference domain model that includes concepts that everyone agrees on, a good approach would be to use the intersection of concepts from multiple

books i.e. including concepts that all the authors specify and excluding concepts unique to a single textbook.

Indeed, our approach is to use the intersection of index terms from multiple textbooks to extract a common domain model for intro-to-programming. This intersection will create a common skeleton which can then be augmented with author/expert-specific concepts. A union approach, i.e., the union of all the concepts specified in the index terms lists of various textbooks, will lead to a more comprehensive domain model that would include many author-specific concepts. We plan to explore such an approach in the future. Here we focus on discovering the common domain model across many textbooks.

It should be noted that performing the intersection or reunion of index terms lists in an automated way is not a straightforward task, given the various ways authors may specify the same underlying concept. For example, the concept of ‘two-dimensional arrays’ could be represented in a hierarchical way, with level 1 specifying ‘array’ and level 2 ‘two-dimensional’, which can be conflated together to get ‘arrays- two-dimensional’. To address this issue, we investigate the use of various approaches based on concept similarity methods (perfect string matching, approximate string with word-to-word semantic similarity) and majority voting, i.e., a concept should be present in the majority of the textbooks used, e.g., in two textbooks out of three textbooks considered. To identify common concepts across three books, we use TFIDF (term frequency-inverse document frequency at character level as detailed later), and transformer-based pre-trained language models for approximate string matching [7].

The major contributions of this paper are 1) automated approaches to extract a common domain model using index terms at the end of textbooks and 2) evaluation of these approaches with respect to a gold standard domain model generated by experts from the same textbook indexes.

The outline of the paper is as follows. The next section, *Related Work*, briefly highlights key prior efforts in the automated extraction of domain models and automatic domain model refinement. It also covers previous efforts on general techniques for approximate string matching. The *Approach* section outlines the key steps of the proposed approach to domain modeling extraction from intro-to-programming textbooks. Then, we present details about our experiments and the results obtained. The *discussion and Future Work* section highlights the important aspects of the proposed approach, contributions, and plans for future work.

## 2 Related Work

### 2.1 Domain Modeling

Previous work in domain modeling extraction has explored various approaches and extracted domain models from various sources. Many such previous works rely on three main information sources: experts in the domain, learners’ data, and textbooks. Many previously proposed approaches are based on keyphrase extraction, and content analysis [1–4]. One previous effort similar to our work

was proposed by Zechmeister and colleagues [25] who manually analyzed ten psychology textbooks to create a set of core concepts. Our work is a step towards automating this approach. In the context of the domain model for computer programming, although works such as [21] provide guidelines for conceptual content to be covered in introductory computing courses, it does not provide a fine-grained list of concepts needed for ITSs. Efforts to extract fine-grained concepts using multiple textbooks were reported [22]; however, these efforts were manual and difficult to scale. Some automatic efforts such as [11, 20] relied on Java parsers or Abstract Syntax Tree (AST) to extract concepts related to a particular computer program. Our work which is done in the context of CS1 and CS2 computer programming courses can provide a fine-grained domain model based on textbooks and can be used in other domains as well.

## 2.2 String Matching Algorithms and Pretrained Transformers

Using exact string matching for our purposes is too constraining, as illustrated later because authors tend to refer to the same concept using different linguistic forms. Therefore, to identify commonalities, we must use more flexible matching methods.

Approximate string matching has been explored in various applications including address matching, name matching, biomedical abbreviation matching as well as spelling correction [5, 19]. The use of such string-matching techniques and algorithms has not been explored previously for domain modeling to the best of our knowledge. Approximate string matching [16] has been explored in many different variants; for example, Cohen et al. [5] experimented with edit-distance, token-based distance and hybrid methods, Ji and colleagues used BERT for biomedical entity normalization [12]. Despite extensive work on fuzzy or approximate string matching, there is no consensus on what method works best and under what circumstances. For our task of matching similar index terms across different textbooks, we consider two main approaches based on TF-IDF and pretrained transformers. The former approach takes into account surface form similarity of index terms based on a character-level analysis whereas the latter takes into account semantic similarity based on pretrained neural embedding models. We experimented with BERT and its variation PhraseBERT [23] which is specifically fine-tuned to capture lexically diverse phrases.

## 3 Dataset

### 3.1 Sample of Textbooks and Index Terms

We used three textbooks to develop and evaluate the proposed automated methods for domain modeling: Introduction to Java [6], Bigjava [10], and Java-How to Program [18]. These textbooks are used as introductory Java programming textbooks but also contain advanced concepts such as networking, java database programming, etc. We focused on index terms from the chapters commonly taught

in CS1 and CS2 courses [22]. The indexes of these textbooks generally include entries or index terms, followed by locators, which can be page numbers, page number ranges, or section numbers. Some indexes contain semantic relationships with other index terms, such as subcategories, synonyms, or cross-references. While there are guidelines for creating an index, there can be organizational and content variations among publishers, domains, books, and authors.

From the perspective of our work here, we took advantage of the hierarchical structure for the indexes of the three textbooks, i.e., indexes in those textbooks describe entries using two hierarchical levels: first/top level and second level. Examples of this hierarchical structure in the three textbooks are shown in Fig. 1.

Floating-point numbers	<code>abs</code> method, <code>Math</code> class, 121–122, 524	argument list 1398
assigning to integer variables, 134	Absolute file name, 473	argument promotion 207
comparing, 188	Abstract classes	argument to a method 41, 76
converting to integer, 142–143.	<code>AbstractCollection</code> class, 762	arithmetic and logic unit
<i>See also</i> cast operator.	<code>AbstractGraph</code> , 1025–1026	(ALU) 9
description, A-72	<code>AbstractGraph.java</code> example, 1028–1033	arithmetic calculation 53
mixing with integer, 139	<code>AbstractMap</code> class, 812	
precision, 133–134	<code>AbstractSet</code> class, 798	
	<code>AbstractTree</code> class, 935–936	

**Fig. 1.** A snippet of index terms in three different books

The second-level index terms are related to the first/top level, but if the second-level index terms were important, they were also present as first-level entries. For example, for the first-level index term “Array,” the second-level index terms might be “sorting,” “for each loop,” “declaring,” and “accessing elements.” Out of these second-level index terms, “for each loop” and “sorting” were also present as separate first-level entries, but “declaring” and “accessing elements” were not. In this paper, we focused on the first-level entries as it seems to be a good start to create a common domain model. We obtained 1495, 2090, and 875 first-level index terms from three books.

### 3.2 Data Annotation

We created a gold standard domain model for computer programming by annotating a sample of index terms that we considered for our analysis. To achieve this goal, we recruited three graduate students well-versed in computer programming to perform the annotation process. The index terms were annotated based on their importance from a pedagogical perspective as judged by the graduate student. They thus represented a computer programming concept worth learning for novices. Our annotators annotated the union of index terms from three Java programming books. As stated earlier, our annotators annotated 4460 first-level index terms. To ensure that our annotations were not specific to these books, we only marked those terms as concepts that were not specific to the book, e.g., we excluded terms that represented method names, classes, or Java-specific terms and only marked those terms as concepts that were generalizable to other programming languages.

We evaluated the inter-rater agreement of our annotation process and found that it was considerable with a pairwise agreement of 0.82, 0.83, and 0.72. However, to ensure the accuracy of our domain model, we only considered index terms as domain terms to which all annotators agreed after the first round of annotation. This scheme enabled us to create a consensus domain model for computer programming. We obtained 465 terms out of 4,204 concepts in the union of index terms from all three textbooks.

The next step in creating the gold standard domain model was concept normalization or canonization, which means identifying different phrasing of the same concept and marking them as referring to a single concept for which a canonical form should be adopted. We call such groups of terms representing a single concept a synset (synonymous set) similar to the set of synonymous words in WordNet [15] that describe the same concept. For example, “binary digits” and “bits” represent a synset. We grouped terms together only if they represented exactly the same concept at the same level of generality or specificity. For example, “for loop” and “while loop” represented two different synsets. As a result, we obtained 263 concepts out of 465 terms. This gold standard domain model will be used to guide our further analysis of computer programming concepts. We use this gold standard domain model to evaluate our results.

## 4 Approach

We experimented with different approaches for extracting domain models based on index terms and using automated approaches. The first approach, *Consensus voting with perfect string matching*, identifies index terms that are common across all the three books in exactly the same linguistic form. For example, concepts such as ‘recursion’, ‘inheritance’, and ‘infinite recursion’ were present in all three books. The second approach, *majority voting with perfect string matching*, relaxes a bit the prior approach and only requires exact terms be present in at least two textbooks, one such example is ‘infinite loops’ which is mentioned in two indexes of two of the textbooks but not all three. Furthermore, our third approach, *majority voting with approximate string matching*, tries to identify similar concepts such as (‘abstract data type (adt)’, ‘adt (abstract data type)’) expressed linguistically in different ways in different books. Another example of a concept expressed in different ways is ‘bits’ and ‘binary digits’. Even though these two phrases refer to the same concept, a string-based method cannot identify them as similar. However, models such as BERT can capture the semantic similarity between these terms. We explain each method in this section.

### 4.1 Consensus Voting with Perfect String Matching

Let A, B, and C be the lists of index terms representing index terms from each of the three textbooks and  $t$  a specific string describing a concept. Let D be the list of identical terms in all three textbooks. Considering D as the domain

model, we can represent domain model  $D$  obtained using consensus voting with perfect string matching as:

$$D = t \mid (t \in A) \wedge (t \in B) \wedge (t \in C) \quad (1)$$

## 4.2 Majority Voting with Perfect String Matching

Majority voting with perfect string matching represents a set of terms present in at least two sets of index terms. As in Eq. 1, let  $D$  be the list of terms present in at least two of the three indexes:  $A$ ,  $B$ , and  $C$ . We can then represent domain model  $D$  as:

$$D = t \mid (t \in A \wedge t \in B) \vee (t \in A \wedge t \in C) \vee (t \in B \wedge t \in C) \quad (2)$$

The equation ensures that a term must be present in the same linguistic form (identical string) in at least two out of the three indexes to be included in  $D$ . This criterion represents majority voting, as it ensures that a term is only included in  $D$  if it has majority support from the textbook indexes.

## 4.3 Majority Voting with Approximate String Matching

Considering the list of index terms from each book  $A$ ,  $B$ , and  $C$  and  $D$  as the resulting domain model, Let  $t$  and  $t'$  be two terms in  $A$ ,  $B$ , and  $C$  such that they belong to different indexes and  $\text{sim}(t, t')$  be the similarity between the terms  $t$  and  $t'$  based on matching approaches. Let  $\tau$  be a threshold value for similarity. The domain model is thus the following:

$$D = t \mid \text{sim}(t, t') > \tau \wedge ((t \in A \wedge t' \in B) \vee (t \in A \wedge t' \in C) \vee (t \in B \wedge t' \in C)) \quad (3)$$

This equation states that a term  $t$  is in the domain model  $D$  if its similarity with any other term  $t'$  in another list is higher than the threshold  $\tau$ . This represents the majority voting, as a term must be similar to terms in the majority of lists to be included in the domain model  $D$ . We experimented with two term similarity methods: TF-IDF at the character level, which is more of an approximate string matching method capturing morphological variations of two index entries, and pre-trained transformer-based models from which we extract embeddings of index terms that can capture the semantic similarity between two index entries. Approximate string matching using TF-IDF and pre-trained transformer models were implemented using freely available python library [8].

**Approximate String Matching with TF-IDF:** We opted for TF-IDF-based string matching to calculate the string similarity between two index entries based on the cosine similarity of TF-IDF of character level n-grams ( $n = 3$ ) of index terms. We select the index term from another book with the highest similarity score for each index term in a book. A similarity score of 1 indicates that the terms match perfectly, whereas 0 indicates the index terms do not have any n-grams in common. Even though 0 and 1 thresholds on similarity give minimum

and maximum bounds, we need a data-driven threshold value that maximizes the accuracy of detecting index terms that represent the same concept. Our threshold selection approach is explained in further detail in Sect. 4.3. After thresholding, we obtain pairs of index terms representing similar concepts that may have different forms in different books. TF-IDF-based approximate matching captures index terms that vary on character level n-grams, such as: ‘string concatenation’ and ‘concatenate strings.’

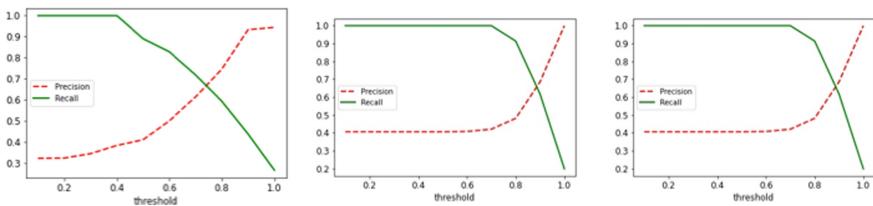
**String Matching with Pretrained Transformer Models:** Even though token-based string matching can capture the differences in the token level of the index terms, such as inflections or different forms, i.e., orthographic variations, they are limited in expressiveness and fail to capture similar concepts based on their semantic similarity, i.e., index terms that use different wordings to represent the same concept. To overcome this limitation and accurately measure the semantic similarity of index terms, we conducted experiments with two transformer-based models: BERT and PhraseBERT.

For instance, the index terms “bound checking” and “bound errors in arrays” refer to the same concept of bound errors in arrays, but a traditional n-gram method like TF-IDF might treat them as different due to the lack of shared words. In contrast, transformer-based models like BERT can consider contextual information and identify that these terms are semantically similar. However, research has shown that BERT relies more on lexical overlap to determine semantic similarity [14, 26], which can result in less diverse phrases in phrase-level semantic-relatedness tasks. To address this issue, we also experimented with PhraseBERT, which fine-tunes BERT using contrastive learning to produce more robust and diverse phrase embeddings.

We extract the embedding of each index term by averaging the embeddings of tokens obtained from pretrained transformer models representing the index terms. We calculate the cosine similarity of obtained index term embeddings to find the best match.

**Selecting Approximate Threshold:** Our string matching techniques estimate the similarity between two index entries based on the cosine similarity values and a threshold value above which the two index entries are considered similar, i.e., referring to the same underlying concept. We need the optimal threshold value: a higher threshold will result in fewer matches being identified but likely to be more precise at the expense of lower recall, whereas a lower threshold will result in more matches but with less precision. Since we are using cosine similarity, a score of 1 would give the exact matching index terms but would ignore similar terms using different words. To select the appropriate threshold for each of our methods for approximate string matching, we searched over the set of possible threshold values (between 0 and 1) using an increment step of 0.1. The goal was to identify the threshold value that maximizes the F-score, which is the balance of precision and recall. To calculate precision and recall and the F-measure, we manually annotated a sample of concepts as annotating all

concepts was too prohibitive. We used stratified sampling with proportionality to ensure that our validation set consisted of a representative sample of each threshold value based on its proportion to the overall sample size, which in our case represents similarity scores of each pair of indexes. This allowed us to obtain accurate precision and recall values without annotating all the samples. Based on this search maximizing the F-measure, we selected 0.8 as the threshold for TF-IDF-based methods and 0.9 for BERT-based methods. This allowed us to choose appropriate thresholds that balanced precision and recall and maximized the F-score. Overall, this approach enabled us to select appropriate thresholds for string-matching techniques efficiently and effectively by maximizing the F-score for the selected thresholds. The plot for precision and recall for various threshold values for TF-IDF is shown in Fig. 2.



**Fig. 2.** Precision-Recall against various thresholds for a) TF-IDF b) BERT c) Phrase-BERT based string matching

#### 4.4 Grouping Similar Terms Together

From the approach explained in the previous sections, we obtained pairs of index terms that are similar based on the optimal thresholds. Consider a pair of index terms ('less than operator', '<operator') and (less than operator, logical operator (<)) where the first and second terms in each pair are from different textbooks. As these pairs represent the same concept, which we refer to in canonical form as 'less than operator,' we combine them into a single group representing a concept, i.e., a synset as in our gold standard domain model.

It is to be noted that this grouping depends on the threshold value and string matching algorithm used; in our case, grouping based on TF-IDF consists of terms that are similar in surface forms, whereas pretrained transformer-based models consist of semantically similar terms. For example, *boolean data type* and *boolean expression* were grouped together by pretrained transformer-based models but not by TF-IDF-based model.

### 5 Results and Evaluation

For each method, we first evaluate the number of terms identified and then compare that to the gold standard. As seen in Table 1 extracting common index terms using the exact string matching methods with consensus leads to only 16 terms.

Even though some preprocessing steps, such as stemming, might considerably increase this number by removing morphological variations among words, this small number of exact common terms highlights the challenge of finding commonalities using index terms lists and the fact that different authors/textbooks use different ways to identify and describe the concepts of a domain. When we consider exact matching index terms that are present in at least two glossaries out of three (majority voting approach with exact string matching), we end up with 163 common terms. One might argue that this could be a very good common skeleton of a domain model that many may agree upon, which is indeed a good argument. We intend to release this common skeleton as a good starting point for an intro-to-programming domain model. Using the approximate matching methods (TFIDF, BERT, and Phrase-BERT) to capture common concepts in combination with majority voting across three glossaries, we ended up with 852 terms (using TFIDF), 362 terms (BERT), and 739 (Phrase-BERT). We evaluated those domain models with respect to our gold standard and reported precision, recall, and F-measure.

**Table 1.** Number of terms obtained for domain model representation using different approaches

Approach	No. of terms
Consensus voting with Perfect String Matching	16
Majority voting with Perfect String Matching	163
Approximate String Matching with TF-IDF	852
Approximate String Matching with BERT	362
Approximate String Matching with PhraseBERT	739

Our gold standard domain model consists of terms grouped into synsets, i.e., terms representing the same concept represent a group. A term in the extracted domain models has a match if it has an exact string match with any term in any synset in the gold standard.

**Table 2.** Precision, Recall and F1 score for different approaches for domain model extraction

Approach	Precision	Recall	F-Score
Consensus voting with Perfect String Matching	0.5	0.029	0.054
Majority voting for Perfect String Matching	0.34	0.205	0.25
Approximate String Matching with TFIDF	0.26	0.71	0.38
Approximate String Matching with BERT	0.22	0.36	0.27
Approximate String Matching with PhraseBERT	0.22	0.63	0.33

A summary of the results is shown in Table 2. As seen in the table, the highest precision is obtained with consensus voting and perfect string matching, which is understandable since common identical terms across three textbooks are likely to be domain terms. However, not all the terms that were common represented important domain terms. For example, the method identifies ‘*Gosling James*’ as a common term across all three books, which is not a core programming concept. Most of the terms identified by the method refer to syntactic aspects of Java programming, such as keywords in java. Our second approach, majority voting with perfect string matching, leads to higher recall as expected. The TFIDF-based approach had the highest precision and recall of domain terms indicating that most of the domain terms only varied morphologically rather than semantically as indicated by results in BERT and Phrase-BERT. We also noticed that BERT-based methods grouped terms like ‘bitwise shift operator’ and ‘bitwise operators’ together although they are different in the sense that one is a more specific concept than the other which explains the lower recall of those methods.

## 6 Discussion and Future Work

This paper presents the implementation details and evaluation of different approaches for automated extraction of domain model from the glossary of textbooks. In our case, approximate string matching with TF-IDF gave better results compared to other approaches. This suggests that string-matching techniques that compare lexical differences can capture domain terms better for computer programming, one of the reasons being, the use of specialized jargon and technical terms in programming. Methods that rely on semantic similarity might perform better in domains where a single concept can be represented in many different ways. Nevertheless, considering string matching techniques greatly improves the F1 score. Overall our approaches are promising. One of the concerns during our experiments was grouping the terms while creating a gold-standard domain model. While we resorted to grouping in the most strict sense, grouping only if two terms represent the same concept, this assumption can be relaxed based on the downstream task. Another concern stems from what granularity to consider for domain representation. For example, ‘pass by value’ is an important concept, however, this could also be considered as an underlying concept of ‘function’. As a remark, while our work is mainly based on the first level of index terms, which we sample for core domain model creation, we can extend it further to different levels of index terms to create a hierarchical model showing the relationship among index terms.

**Acknowledgments.** This work has been supported by the following grants awarded to Dr. Vasile Rus: the Learner Data Institute (NSF award 1934745); CSEdPad: Investigating and Scaffolding Students’ Mental Models during Computer Programming Tasks to Improve Learning, Engagement, and Retention (NSF award 1822816), and Department of Education, Institute for Education Sciences (IES award R305A220385). The opinions, findings, and results are solely the authors’ and do not reflect those of NSF or IES.

## References

1. Alpizar-Chacon, I., Sosnovsky, S.: Knowledge models from pdf textbooks. *New Rev. Hypermedia Multimed.* **27**(1–2), 128–176 (2021)
2. Banjade, R.: Domain model discovery from textbooks for computer programming intelligent tutors. In: FLAIRS Conference Proceedings, p. 34 (2021)
3. Banjade, R., Oli, P., Tamang, L.J., Rus, V.: Preliminary experiments with transformer based approaches to automatically inferring domain models from textbooks. In: Proceedings of the 15th International Conference on Educational Data Mining, p. 667 (2022)
4. Chau, H., Labutov, I., Thaker, K., He, D., Brusilovsky, P.: Automatic concept extraction for domain and student modeling in adaptive textbooks. *Int. J. Artif. Intell. Educ.* **31**(4), 820–846 (2021)
5. Cohen, W.W., Ravikumar, P., Fienberg, S.E., et al.: A comparison of string distance metrics for name-matching tasks. In: IIWeb, vol. 3, pp. 73–78 (2003)
6. Daniel Liang, Y.: Introduction to java programming (2007)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
8. Grootendorst, M.: Polyfuzz: fuzzy string matching, grouping, and evaluation (2020). <https://doi.org/10.5281/zenodo.4461050>
9. Hjørland, B.: What is knowledge organization (KO)? *Knowl. Organ.* **35** (2008). <https://doi.org/10.5771/0943-7444-2008-2-3-86>
10. Horstmann, C.S.: Big Java: Early Objects. John Wiley & Sons, Hoboken (2016)
11. Hosseini, R., Brusilovsky, P.: Javaparser: a fine-grain concept indexing tool for java problems. In: CEUR Workshop Proceedings, vol. 1009, pp. 60–63 (2013)
12. Ji, Z., Wei, Q., Xu, H.: Bert-based ranking for biomedical entity normalization. AMIA Summits Transl. Sci. Proc. **2020**, 269 (2020)
13. Kumar, A.N.: Model-based reasoning for domain modeling in a web-based intelligent tutoring system to help students learn to debug C++ programs. In: Cerri, S.A., Gouardères, G., Paraguaçu, F. (eds.) ITS 2002. LNCS, vol. 2363, pp. 792–801. Springer, Heidelberg (2002). <https://doi.org/10.1007/3-540-47987-2-79>
14. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9119–9130 (2020)
15. Miller, G.A.: Wordnet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
16. Navarro, G.: A guided tour to approximate string matching. *ACM Comput. Surv. (CSUR)* **33**(1), 31–88 (2001)
17. Olney, A.M.: Extraction of concept maps from textbooks for domain modeling. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6095, pp. 390–392. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13437-1\\_80](https://doi.org/10.1007/978-3-642-13437-1_80)
18. Paul, D., Harvey, D.: Java-how to program (2012)
19. Recchia, G., Louwerse, M.M.: A comparison of string similarity measures for toponym matching. *Comp. J. South. Comp. Lit. Assoc.* (2013)
20. Rivers, K., Harpstead, E., Koedinger, K.R.: Learning curve analysis for programming: which concepts do students struggle with? In: ICER, vol. 16, pp. 143–151 (2016)

21. Seehorn, D., et al.: CSTA k-12 computer science standards: Revised 2011. Technical report, New York, NY, USA (2011)
22. Tew, A.E., Guzdial, M.: Developing a validated assessment of fundamental cs1 concepts. In: Proceedings of the 41st ACM Technical Symposium on Computer Science Education, pp. 97–101 (2010)
23. Wang, S., Thompson, L., Iyyer, M.: Phrase-bert: improved phrase embeddings from bert with an application to corpus exploration. In: EMNLP, pp. 10837–10851 (2021)
24. Wu, Z., Li, Z., Mitra, P., Giles, C.L.: Can back-of-the-book indexes be automatically created? In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 1745–1750 (2013)
25. Zechmeister, J.S., Zechmeister, E.B.: Introductory textbooks and psychology's core concepts. Teach. Psychol. **27**(1), 6–11 (2000)
26. Zhang, Y., Baldridge, J., He, L.: Paws: Paraphrase adversaries from word scrambling. arXiv preprint [arXiv:1904.01130](https://arxiv.org/abs/1904.01130) (2019)



# Teaching Cardiovascular Physiology with a Minigame-Based ITS

Reva Freedman<sup>(✉)</sup>, Lucas Edwards, Ian Sullivan, Annalise Brockmann, Dean LaBarbera, and Virginia Naples

Northern Illinois University, DeKalb, IL 60115, USA  
[r freedman,vlnaples}@niu.edu](mailto:{r freedman,vlnaples}@niu.edu), [z1861935@students.niu.edu](mailto:z1861935@students.niu.edu)

**Abstract.** This paper describes the use of minigames to teach cardiovascular physiology in the CRAM system. We describe two new intermediate-level games we have written to improve the teaching of the physiology of the heart and lung to advanced undergraduates. Each of these games builds on the knowledge students have acquired in an elementary game and teaches new material using a higher level of Bloom's taxonomy. We also describe a third game where students simulate a medical procedure and the taxonomy behind the games. The purpose of these games is to improve motivation by making the games fun while at the same time improving student learning by reducing cognitive load. The games have been designed with reusable design and gamification elements for efficiently expanding the system to other topics. Our goal is to increase student success by teaching higher-level concepts and increasing time on task while reducing student frustration.

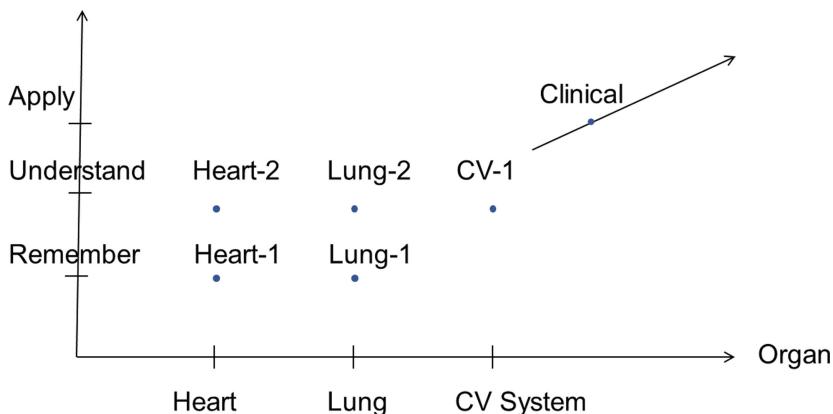
**Keywords:** intelligent tutoring system · anatomy education · minigames

## 1 Introduction

Many students who are not biology majors have difficulties with the upper-level course in human anatomy and physiology required for health sciences programs in our university. Since gaming is increasingly popular among all types of college students, we are studying whether a game-based ITS would be successful for this population of learners.

To study this question, we are developing an ITS based on a set of minigames. The CRAM system (Cardiovascular and Respiratory Anatomy) is designed as a three-dimensional matrix of minigames (Fig. 1), where the horizontal axis identifies the organ system involved, the vertical axis identifies the learning objective of the minigame according to Bloom's taxonomy, and the third axis identifies the point of view that the user takes in the game. In most of our games so far, our goal is to teach the facts of anatomy and the causal relations involved in physiology. We know that students are motivated by applied problems such as diagnosis, treatment and patient followup, so we have designed one game in

the applied realm and are working on more. Students can progress through the matrix at their own speed and following their own needs and interests.



**Fig. 1.** Conceptual structure of the CRAM system

We are using Bloom's taxonomy [1] as a guide to the type of game play elements appropriate for different kinds of learning. Bloom's taxonomy provides an approach to categorizing learning tasks with respect to the level of abstraction required. Students have difficulty achieving the learning objectives for the course since the available textbooks for anatomy and physiology are too complex for many students to master the material at the different and interlocking levels of detail required.

Although there are many recent survey articles on games in both medical and other types of education [2,5,6,9] there is no domain-independent answer yet as to the circumstances in which games are worthwhile. In addition to the great number of variables involved, not all of the experiments to date have met the gold standard: some have not included a control group, and some that have shown learning have not been able to explain the mechanism.

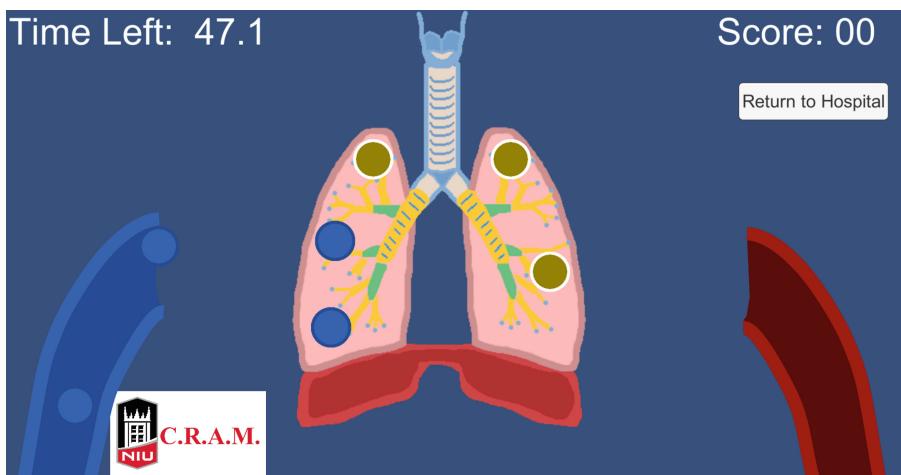
In this paper we illustrate the taxonomy by showing three games from different sections of the matrix. Section 2 gives a brief discussion of the software powering the games. Section 3 shows Lung Stage 2, which teaches students about the physiology of the lung. Students use this game after completing Lung Stage 1 [4], which teaches them the anatomy, as they need to learn the names of the parts of the lung before they can understand how lungs work. To illustrate the functioning of the game, we show it at startup, at an intermediate point, and after it has concluded.

Section 4 demonstrates Heart Stage 2. Each game is conceptualized individually depending on students' needs. Thus Lung Stage 1 and Lung Stage 2 teach anatomy and physiology, respectively, while the order is reversed for the heart. Heart Stage 1 [4] teaches the basic functioning of the heart while Heart Stage 2

teaches the names of the main heart vessels. Although these two topics are generally combined in discussions of the heart in most textbooks, we feel that separating them reduces the cognitive load on students. A future game, Cardiovascular Stage 1, will help the student master the function of the cardiovascular system in total. To illustrate the use of the clinical axis, Sect. 5 demonstrates a game that helps the student master the concept of cardiac arrhythmia. Students are always interested in clinical material that applies the concepts learned to a real-world situation. Finally, Sect. 6 describes our future plans.

## 2 Methodology

The games are implemented in C# using the Unity game engine [11]. Unity has several advantages for game development: deployment on multiple platforms, a built-in event loop, handling of networking and concurrency issues, artistic elements such as graphics and animation, and specialty programming such as a physics engine. Although the system consists of a series of minigames, reusing design elements and game mechanics where possible not only reuses the cognitive load for students, it reduces the development effort as well.



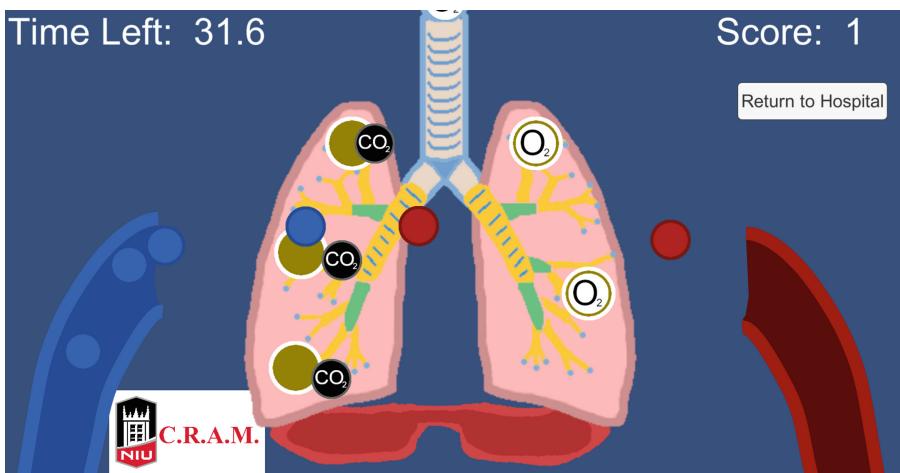
**Fig. 2.** Startup state for Lung Stage 2: Physiology (Color figure online)

## 3 Learning the Physiology of the Lung

Lung Stage 2 is a follow-on to Lung Stage 1, a game that teaches the names of parts of the lung. As a result, Lung Stage 2 is the first game that students meet that will teach them about the physiology of the lung. As part of our goal to reduce the cognitive load on students, we separate these topics into two separate

games. While Lung Stage 1 is at the “remember” stage of Bloom’s taxonomy, Lung Stage 2 includes aspects of the “understand” and “apply” levels.

The pedagogical goal of Lung Stage 2 is for the student to understand how oxygen enters the lungs, enters the alveoli to oxygenate the blood, and then CO<sub>2</sub> is exhaled. While we could use a basic memorization package like Quizlet [8] to teach these facts, that has disadvantages from both motivational and pedagogical points of view. With regard to motivation, the distinct creative concepts we have used as the basis for each individual game provide more fun than yet another set of flash cards. From the pedagogical point of view, each game has been designed to make the student think through the issues at the second and third levels of Bloom’s taxonomy rather than just memorize facts at the first level. This level of understanding will be helpful to the student in two ways. First, it will help them do better on exams and in the lab component of the course. Second, it will help them with more advanced games, either at the same level for different topics, or for games on the application axis that will be at level 4 of Bloom’s taxonomy. For example, Lung Stage 2 is a prerequisite for Cardiovascular Stage 1, which integrates knowledge about the functioning of the heart and the lung with the goal of ensuring that students understand the role of the cardiovascular system in human functioning.

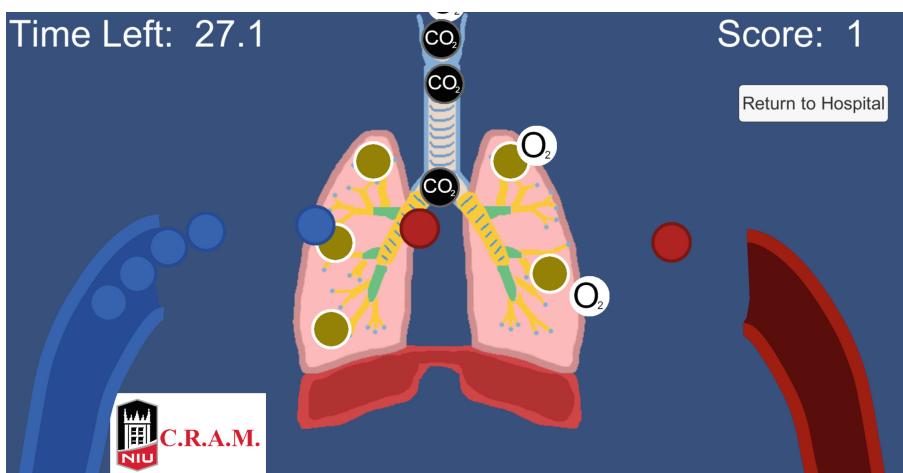


**Fig. 3.** Lung Stage 2 in play

Figure 2 shows the game in its startup state. The game loop is divided into four parts. First, the user clicks on the diaphragm to expand the lungs and draw in oxygen molecules. Second, they drag oxygen molecules and deoxygenated blue blood cells together in the alveoli (represented by the yellow and white circles), which results in oxygenated red blood cells. Third, the user drags the oxygenated blood cells to the right, for each of which they receive points added to their score.

Fourth, they press the diaphragm again to expel the CO<sub>2</sub> out via the trachea and make the lungs contract. The user repeats the loop until they run out of time. The goal is to get as high a score as possible. There are three alveoli on the left and only two on the right to represent the fact that the right lung of the body has three lobes and the left has only two. The rate at which the deoxygenated blue blood cells spawn can be adjusted to the player's skill level.

Figure 3 shows the game in an intermediate state. The user has pressed on the diaphragm and the diaphragm has expanded downward. As the lungs expand, oxygen cells move down the trachea. The player will combine the oxygen molecules with the blue blood cells by dragging them. Then the oxygen becomes a carbon dioxide molecule and the blue blood cell becomes an oxygenated red blood cell. The user can then deposit the oxygenated blood cell in the red artery on the right hand side where the oxygenated cell re-enters the blood stream.

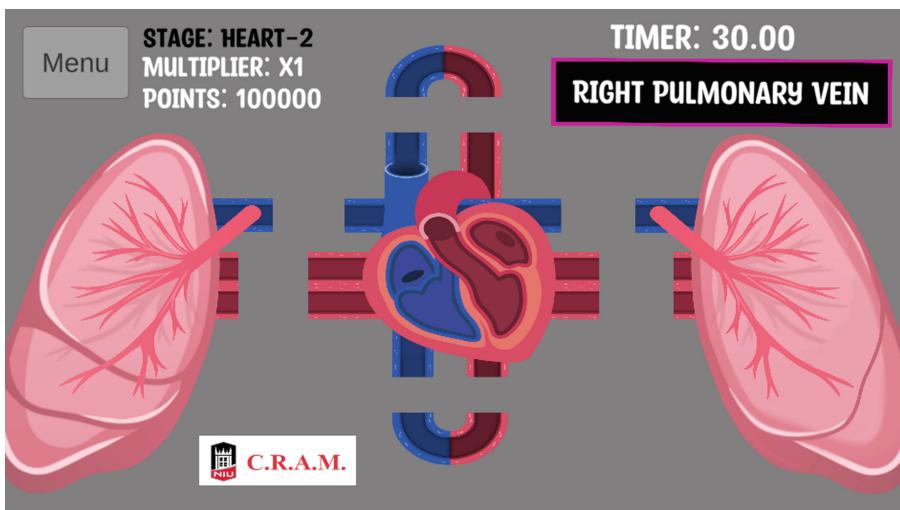


**Fig. 4.** Lung Stage 2 after completion

Finally, Fig. 4 shows the completion of the game. The user has pressed the diaphragm again, which causes carbon dioxide molecules to return to the trachea and move up it. At the same time, the lungs shrink back to their original size and the diaphragm returns to its original position. Users can then press the diaphragm again to obtain more oxygen molecules, and repeat this process until they run out of time.

Gamers and game designers know that games are more fun when game play is intuitive and does not need to be explicitly taught. Thus students can play Lung Stage 2 without using a single word in the main gameplay loop. In addition to increasing the time-on-task of this game, this can be a benefit to students who prefer to learn visually or through actual play rather than mediated by words.

Additionally, the concept behind this game could easily be extended to the physiology of any bodily organ that involves a free floating substance like blood, such as nutrient exchange or interaction inside a cell. What makes this possible is the use of advances in game development technology in the Unity platform. In exchange for not being able to control the physics completely, we do not have to develop our own folk physics. This game uses Unity's built-in physics engine to handle molecular movement with little modification or configuration. Therefore developing additional games that teach physiology at the second and third level of Bloom's taxonomy could prove to be both cost effective and effective for student learning.



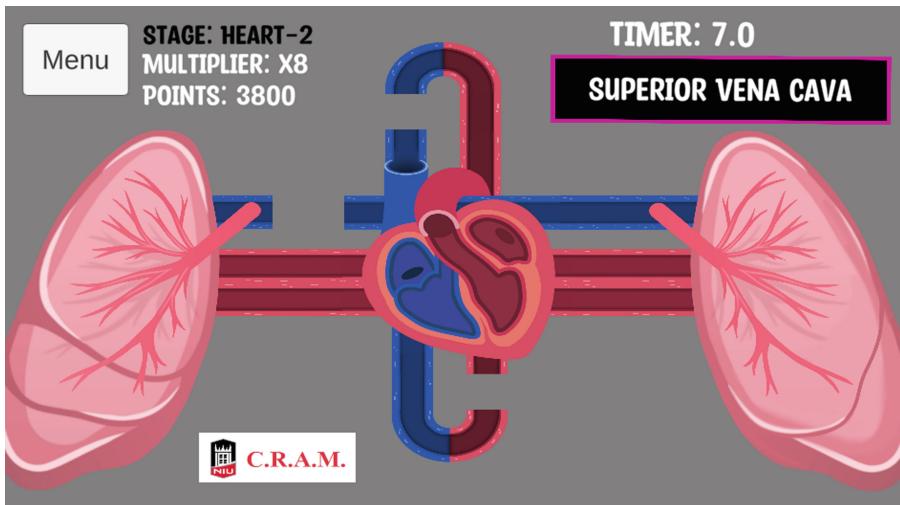
**Fig. 5.** Heart Stage 2 at startup

#### 4 Learning the Anatomy of the Heart

The heart games have been conceptualized in a slightly different fashion than the lung games. Our approach to teaching the anatomy and physiology of the heart differs from conventional textbooks [7, 10]. In most textbooks, students learn the names of the major vessels at the same time as they are trying to understand blood circulation to and from the heart. We reduce the cognitive load on the student by splitting these topics into two stages. In stage 1, students learn how oxygenated and deoxygenated blood moves between the heart, the lungs and the rest of the body. Then in stage 2, students connect this understanding with the names of the major vessels. Thus while we teach anatomy first for the lung, we teach physiology first for the heart, followed by details of anatomy.

Figure 5 shows the stage 2 game for the heart in its startup state. In this game, students are tasked with observing the blood vessel named in the top

right of the screen. They then must drag the ends of the correct blood vessel together until they connect, which visually “completes” the blood vessel. Figure 6 shows Heart Stage 2 in an intermediate state, where some but not all of the open vessels have been completed.



**Fig. 6.** Heart Stage 2 in play

The game randomly precompletes a number of blood vessels, allowing for each run of the game to be distinct. Each instance of the game always has at least one unconnected vessel for a student to complete. If students have chosen a timed mode and finish the game before the timer runs out, the game will automatically restart with a different random set of blood vessels to complete.

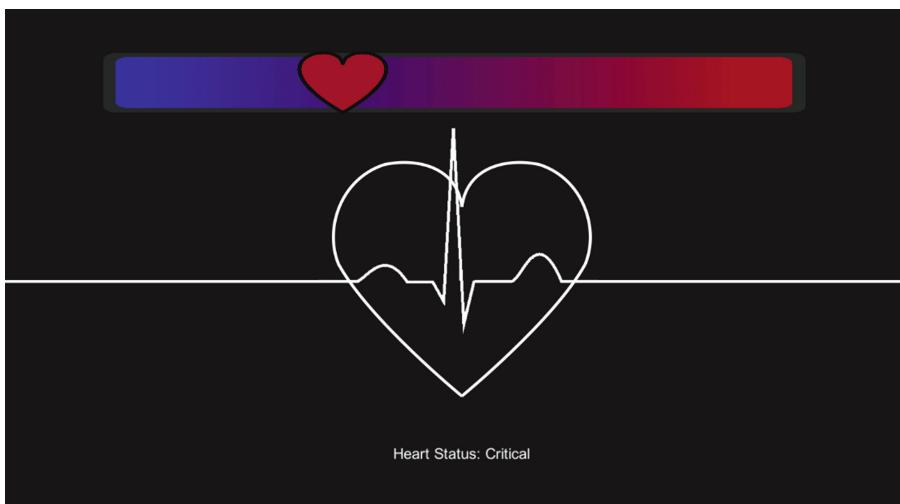
Students get points for every correctly connected blood vessel. Since we want to emphasize accuracy over speed, students get additional points for multiple correct answers in a row. Without this correct answer streak bonus, students might be more inclined to rush than to think it through. Although that may be appropriate in a game played purely for enjoyment, our games are primarily intended for learning. Fun is important primarily for maintaining student attention and motivation.

The game has a leaderboard that can keep track of repeated runs. It tracks statistics such as the percentage of correct answers and total points. The use of the leaderboard is intended to push students to beat their scores in a fashion akin to arcade games. The intent is to use the scores of other students to generate friendly competition and give the students a way to indirectly push each other forward.

There are several timing options available to the player, including an untimed option. The timed options are for those who want a more arcade-like experience,

while the untimed option is for those who prefer a less stressful experience and those who want to use the game to review but cannot physically handle the speed required in the other modes. Options like these allow students to customize the game to their wants and needs.

The game attempts to correct players who miss answers by wiggling the blood vessel representing the correct answer. This non-verbal hint allows the student to then actively choose the correct answer. In this fashion the student gets active practice to learn the material without wasting time guessing. The goal is to get the student to learn the names of the vessels in the most efficient manner possible.



**Fig. 7.** Cardiac arrhythmia game

## 5 Cardiac Arrhythmia: An Applied Clinical Game

Figure 7 shows a simple game that teaches the student what to do about cardiac arrhythmia at a conceptual level. The student is presented with the following scenario: “Mr. B has an arrhythmia with an irregular heart rate.” Instructions are shown on how to engage with the minigame, for example, “Tap the heart on the beat.” A heart appears in the center of the screen followed by a line from an electrocardiogram which approaches the heart from the left. The QRS complex occurs in the center of the heart. The student taps the heart as this occurs. The line then exits from the right side of the heart and reappears on the left. The loop continues until the student correctly taps the heart a specified number of times. A congratulatory message appears indicating that “Mr. B is feeling better,” and the student sees an animation. At this point cardiac output

will return to normal, proper blood circulation in the lungs will resume, and the arterial system can again fulfill the oxygen demand requirement of the body.

This game is different from the others shown because the task is different: Students are simulating a concrete action to help a patient rather than just demonstrating their knowledge. Although the details of medical treatment are beyond the scope of the course, students planning on a career in the health professions are motivated by learning about medical applications. In addition, learning about how to correct a heart rhythm problem is a good way for them to understand both what a normal rhythm is and how it can go awry.

## 6 Conclusions and Future Work

In this paper we have shown how we can increase fun and reduce cognitive load in the teaching of cardiovascular physiology through the use of minigames with reusable design elements and gamification mechanics.

This research is an advance over earlier games for teaching anatomy, e.g., [3], because the games are explicitly based on levels 2 and 3 of Bloom's taxonomy. In these games students learn some of the key causal mechanisms underlying the behavior of the heart and lung rather than just memorizing facts or concepts. In addition, interactive learning using mechanisms known to appeal to college students helps to keep them on task.

We have attempted to reuse popular game design elements wherever possible, both to reduce development cost and to provide a smoother experience for the student. Art and other screen design elements are also repeated where possible. We have also attempted to insure the consistency of various design elements, such as element sizes, logos and borders.

All of our games can be made more challenging via time pressure and other game mechanics, which allows students to use the games at the level they need for learning and also makes the game more fun for experienced players.

In future work, the games will send information about student performance to a shared database so that we can analyze student behavior and provide an open learner model. In addition to studying the relationship of game play to student success in the course, we will attempt to identify which gamification features have contributed to student learning.

## References

1. Bloom, B.: Taxonomy of Educational Objectives. Handbook I: The Cognitive Domain. David McKay, New York (1956)
2. Donkin, R., Rasmussen, R.: Student perception and the effectiveness of Kahoot!: A scoping review in histology, anatomy, and medical education. *Anat. Sci. Educ.* **14**(5), 572–585 (2021). <https://doi.org/10.1002/ase.2094>
3. Educational Technology and Mobile Learning: 11 free tools to teach human anatomy in 3D (2012). <https://www.educatorstechnology.com/2012/04/11-free-tools-to-teach-human-anatomy-in.html>

4. Freedman, R., Naples, V., Sullivan, I., Edwards, L., LaBarbera, D.: Gamification, user-centered design and learning objectives as the basis for a minigame-based cardiovascular anatomy ITS. In: Proceedings of the 18th International Conference on Intelligent Tutoring Systems, pp. 336–342 (2022)
5. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work? – A literature review of empirical studies on gamification. In: Proceedings of the Forty-Seventh Annual Hawaii International Conference on System Sciences, pp. 3025–3034. IEEE, Piscataway, NJ (2014). <https://doi.org/10.1109/HICSS.2014.377>
6. McCoy, L., Lewis, J.H., Dalton, D.: Gamification and multimedia for medical education: A landscape review. *J. Am. Osteopath. Assoc.* **116**(1), 22–34 (2016). <https://doi.org/10.7556/jaoa.2016.003>
7. Moore, K., Agur, A., Dalley, A., II.: Clinically Oriented Anatomy, 8th edn. Wolters Kluwer, Philadelphia (2017)
8. Quizlet: (2005). <http://quizlet.com>
9. Rutledge, C., et al.: Gamification in action: Theoretical and practical considerations for medical educators. *Acad. Med.* **93**(7), 1014–1020 (2018). <https://doi.org/10.1097/ACM.0000000000002183>
10. Tortora, G., Nielsen, M.: Principles of Human Anatomy, 14th edn. Wiley, Hoboken, NJ (2016)
11. Unity: (2021). <http://unity.com>



# Expert-Centered Design Recommendations to Promote Engagement in Authoring Activities of Adaptive Learning Technologies

Emmanuel G. Blanchard<sup>(✉)</sup>

IDÛ Interactive Inc., Montréal, Canada

emmanuel.g.blanchard@gmail.com

**Abstract.** Despite a growing interest in adaptive educational technologies in recent years, research findings are rarely integrated into popular mainstream technologies. We believe that a possible explanation is the lack of consideration for engaging domain experts. Based on observations from two multi-year projects aimed at developing adaptive technology for knowledge transfer (one in a business context, the other one in a research context), we present design recommendations for creating authoring tools that are more centered on domain experts.

**Keywords:** authoring tools · domain expert · expert-centered design · adaptive learning technologies · design recommendations

## 1 Introduction

Currently, educational technologies deployed in the *real world*<sup>1</sup> generally implement a very classic eLearning approach while integrating only limited “adaptive” techniques and methodologies such as those of interest for the ITS/AIED community. This is paradoxical given the spotlight that “Intelligent” Educational Technologies have received during the COVID era. This lack of transfer between our community’s research output and real-world applications affects not only technologies for typical educational contexts (teaching in schools, universities, etc.) but also many other situations where knowledge transmission technologies can be deployed (e.g., assisting museum visits, lifelong learning, reskilling). Therefore, it is a topic frequently discussed at conferences in our field such as ITS, AIED, ICALT, ECTEL, in a more or less formal ways.

One possible explanation lies in difficulties that real-world actors have in appropriating learning ecosystems resulting from research initiatives. Indeed, researchers generally

<sup>1</sup> In this paper, the term “**real world**” is used in contrast to the “academic research world”. Real-world systems are primarily intended to provide a service in a real-world context, with a commercial objective in some cases. In such a situation, innovation and/or contribution to scientific research is a possible but not necessary by-product. On the contrary, it is the main driver of a research initiative, whether it takes the form of technical proofs of concept or technopedagogical hypothesis validations. Therefore, this term serves to differentiate the primary objectives of a development initiative and should not be seen as a qualitative judgment.

aim to develop technical proofs of concept and/or to validate specific techno-pedagogical hypotheses, which frequently results in the development of relatively complex authoring tools in order to adequately consider phenomena of interest for research purpose.

Most of these phenomena are related to different aspects of the computer-based learning process. In this context, ensuring that the complexity of authoring tools remains low is rarely a priority, and the availability of sufficient adaptive learning materials may not be questioned at all. However, a good “domain expert” is well known to be a scarce resource, and we believe that too little consideration for domain experts and their involvement is a risk that can have a major negative effect on the impact of our community on Society for at least two reasons:

- Adding complexity to any process reduces the quantity or quality that it generates in a given period of time. If we consider domain experts as resources whose availability is limited, having them use complex authoring tools will thus limit the quantity or quality of contents they produce. In both cases, the effect will be negative: a lower quality of content increases the risks of introducing noise (affective, emotional, motivational), while having less content limits adaptation capacities of educational systems that need diversity to allow variations in teaching approaches and strategies.
- In the real world, authoring activities of domain experts frequently “compete” with their other responsibilities. Therefore, the complexity of a tool is a primary criterion when choosing a platform, and not giving it enough importance may eventually hinder the diffusion of other assets of a technological ecosystem for knowledge transfer.

To address this topic and stress possible solutions, the following two research questions are considered:

- What characteristics are found in domain experts that can impact the process of authoring adaptive educational content?
- What authoring tool design guidelines could facilitate domain expert engagement in adaptive content creation activities?

The rest of this paper is organized as follows. The next section provides a brief state of the art on authoring systems. We then present two adaptive technology projects on which our reflection is based. The next section reports characteristics commonly observed on experts involved in these respective projects. We then present a set of design guidelines inspired by these observations, as well as how we translated them into implementation features in our authoring tools. Finally, in conclusion, we provide additional comments recently collected from four domain experts.

## 2 Brief State of the Art of Authoring Tools

Developing authoring tools for adaptive educational technologies has been a major interest of the ITS community for many years [1, 2]. Recently, Demerval et al. surveyed 4633 articles, of which 33 were selected to illustrate different research directions [3]. Results show that the ITS community has been interested in facilitating authoring activities for

all major ITS components, but especially for the curriculum [4, 5] and the planner [6, 7]. As in other ITS research subdomains, work on authoring tools has incorporated many advanced techniques associated with the broad field of artificial intelligence, beginning with the use of machine learning techniques to enable advanced features into ITS such as emotion recognition [8], learning style recognition [9, 10], or cognitive tutor integration [11].

The recent and well-documented literature review of Demerval and colleagues has considered several other foci of research on authoring tools such as listing elements that facilitate the authoring process, characterizing the temporality of such a process, and listing demonstrated benefits of using authoring tools. Yet, one element stands out. No key section of this literature review investigates authoring tool users. Is this because the authors did not consider this element important? Is it because work on the topic is limited or nonexistent? In either case, this highlights that authoring tool users do not seem to be a major research topic, which is surprising for a very user-centered community.

### 3 Presentation of Two Adaptive Learning Technology Projects

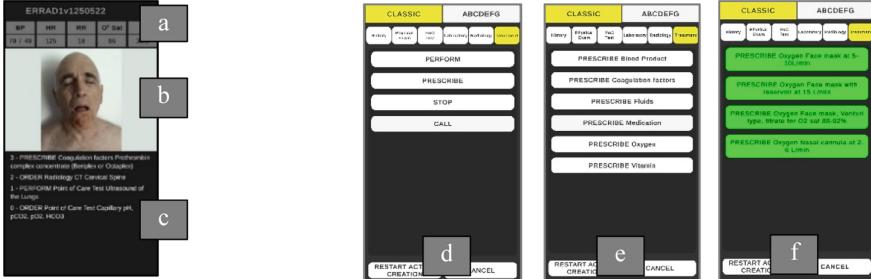
The work reported in this paper stems from two projects on adaptive technologies for knowledge transmission. Both were conducted in parallel over several years, using an iterative participatory design approach involving domain experts. In what follows, we give a brief presentation of these projects.

#### 3.1 The Deteriorating Patient Application (DPA)

This research-focused project aims to digitize a real-world activity for training medical students, the *Deteriorating Patient*, in order to make it scalable (more learning opportunities without affecting the time required by experts).

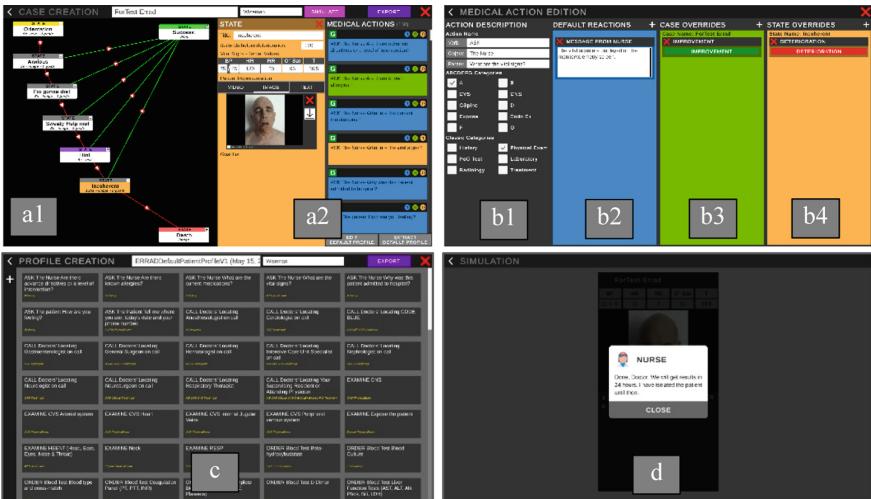
**User Experience.** A medical student must stabilize a virtual patient by applying an appropriate set of virtual medical actions. The medical student will therefore try to identify clues to diagnose the problem of the virtual patient in order to apply an appropriate treatment. Some of these cues are provided directly by the patient's vital signs and multimedia representation, which evolve dynamically, while others are provided in response to the actions the student decides to apply to the virtual patient. These actions also impact the state of the virtual patient. Figure 1 presents screenshots of the DPA application.

The GUI of the DPA client app is organized as follow. A bar presenting five dynamically evolving vital signs is on top (a). Below, the state-evolving representation of the virtual patient is displayed (b). Finally, a list of actions already performed on the patient is shown at the bottom (c). Clicking on the patient representation will bring forth a popup allowing the student to progressively construct a medical action to be performed on the virtual patient. First an action verb is selected e.g., PRESCRIBE (d). The user then chose to combine it with one of the available "objects" e.g., PRESCRIBE Oxygen Face Mask (e), and possibly with parameters e.g., PRESCRIBE Oxygen Face Mask at 5–10 L/min (f). When the constructed action is complete and can be performed, its item appears in green.



**Fig. 1.** On the left, main screenshot of the DPA client application. On the right, screenshots of the 3-step medical action construction popup.

**Authoring Experience.** An expert first defines a list of medical actions and associates each of them with default reactions (i.e., how a healthy patient would react to these actions). To create a specific medical case simulation, the expert overrides some of the default reactions with others specific to the medical condition being simulated in the case. The expert also creates a graph to appropriately describe the possible states of the virtual patient and transitions between them. Figure 2 presents screenshots of the DPA authoring tool.

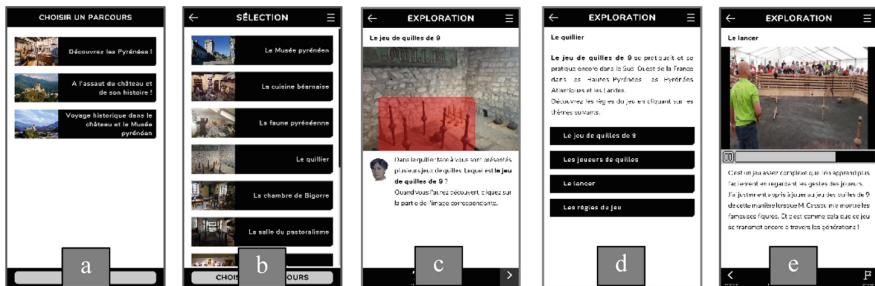


**Fig. 2.** From left to right, top to bottom. Main panel of the DPA authoring tool with a graph representation of a medical case (a1) and editable information about the currently selected state (a2). Medical action panel of the DPA authoring tool with areas to define the action (b1), its default reaction (b2), as well as possible reaction overrides at case (b3) or state (b4) levels. Default patient panel (c) where all defined medical actions are listed. Clicking on one of them will display the previously mentioned action panel (b). Simulation panel (d) currently showing a “nurse message” reaction to a previously performed medical action.

### 3.2 ADaptive COmpanion for Museum and Cultural Heritage Sites (ADACO)

This commercial project aims to provide visitors of museums and cultural heritage sites with a virtual tour guide with which they can customize an interactive storytelling experience based on various factors such as their personal interests, the time they have available for their visit, or the presence of children. It allows for both indoor and outdoor experiences, primarily on smartphones but was also adapted for use on interactive tables. It has been successfully deployed in cultural institutions and interpretation centers in Canada, France and Belgium.

**User Experience.** Once the visitor has decided to receive a support for a point of interest, a short introduction is presented. At the end of this introduction, he/she can move on to the next piece of information by clicking on a “Next” button. If there are several possible follow-up experiences, a dialog panel allows the visitor to make a choice. Each information screen can use different multimedia or interaction technologies (text, images, sound, video, 3D, NFC, BLE, AR...), and will sometimes require the visitor to act in order to unblock access to the “Next” button. Consequently, as he/she makes choices, the visitor progressively build a personalized visit support according to his/her level of interest in what is presented. Figure 3 presents screenshots of the ADACO app.



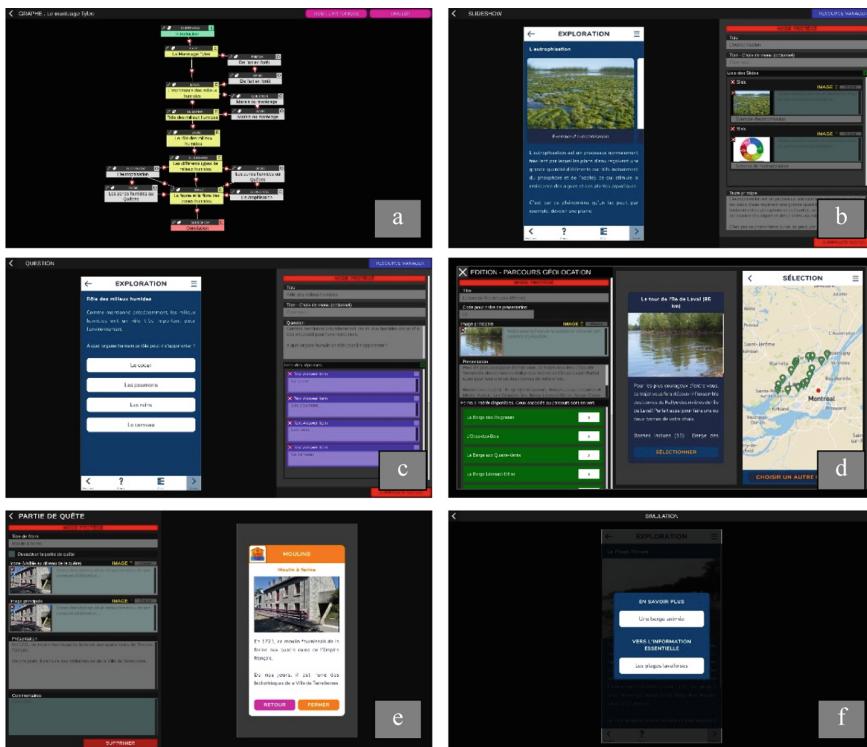
**Fig. 3.** Screenshots of the ADACO client app: choose a tour (a); select the point of interest to explore -list trigger- (b); enigma micro-interaction (c); multiple choices questionnaire micro-interaction (d); video micro-interaction (e)

The GUI of the client app first allows visitors to choose one of the various available tours for the site. Depending on the tour trigger strategy, the interface adapts to allow visitors to choose which point of interest (unit) to explore. During the exploration, the interface adapts depending on the type of the micro-interaction being displayed.

**Authoring Experience.** “Unit” is the name given to an interactive storytelling associated with a point of interest of a museum or a cultural heritage site. For each such unit, an expert creates a narrative graph, whose nodes are called “micro-interactions”. They are the association of a knowledge chunk with a transmission strategy that can use different interactive or multimedia technologies (see above). Fifteen micro-interaction types have been defined so far and are organized in three categories according to their main interest

for the visitor: knowledge transmission, interactive involvement, navigation help. When creating the narrative graph, an expert determines a type for each node. He can then “enter” the node. The next interface of the authoring tool asks him/her to provide the necessary information and resources according to the micro-interaction type.

The expert also has to group units in order to create tours (a unit can be in several separate tours) based on a common theme or a specific visitor target (e.g., adults, children). Each tour is associated with a trigger type that determines how visitors access the interactive storytelling associated to units (e.g., a list of unit items, a geolocated map, contact with NFC chips, etc.). More advanced features are also available for experts to enrich visitor experiences (e.g., creating quests for virtual objects and inserting them in narrative graphs) or personalize the look and options available in the app (theming). Figure 4 presents screenshots of the ADACO authoring tool.



**Fig. 4.** Screenshots of the ADACO authoring tool: storytelling graph for a unit (a); edition panels for a slideshow micro-interaction (b) and a multiple choices questionnaire (c); edition panel for a tour with a geolocation trigger (d); edition panel for a virtual object for a quest (e); simulation panel (f) with a choice dialog activated.

### 3.3 DPA and ADACO in the Context of Its and Adaptive Learning Technologies Research

These two projects are examples of interactive learning platforms. However, the direct influence of AI techniques in what is presented is intentionally limited in the context of this paper. Considering the traditional ITS architecture, we focus our presentation on the curriculum and GUI modules. However, these platforms also incorporate means for collecting user events as well as simulation/experience events that can be used in the planner and student model modules. As mid-term goals, we are developing processes to i) improve the quality of content production by informing experts of successful or unsuccessful experiences, ii) automatically propose the next learning experience for a specific learner or group of learners, iii) assist medical experts with pattern discoveries and data visualizations in hybrid debriefing sessions.

## 4 Characteristics of Domain Experts

Both projects have been ongoing for several years now. As a result, many experts from different domains have been involved (more than 15 for the ADACO project and 2 for the DPA project), from Canada, France, and Belgium. Despite their different cultural backgrounds, and areas of expertise, we frequently noticed common characteristics. In this section, we report on some of these characteristics that strongly influenced the development of our authoring tools. Although these observations were made in an uncontrolled manner and on a limited number of experts, explicitly stating them is important because we believe they are applicable and relevant to other experts in real-world contexts.

**Authoring is a secondary objective.** For all the experts, authoring interactive contents is an activity aimed at support larger goals, from providing a satisfying visitor experience to ensuring that medical students are as well trained as possible.

**Irregular involvement in authoring.** An underlying consequence of this non-priority of authoring is that experts' involvement in this task is irregular. They all have other, more pressing obligations, and sometimes weeks or months pass between periods of meaningful interactive content creation.

**Passionate about their field of expertise.** We observed a great passion from most of the experts we interacted with. As a result, their motivation to get involved in authoring activities was strong if they were convinced of the added value that technology would bring to their main domain-related mission.

**Positive impact of feeling considered.** Because of their desire to produce a quality result, experts frequently made suggestions for improvement. Whether these ideas were accepted or not, discussing ways for the tools to better meet their needs always resulted in a boost in motivation.

**Limited ability for abstract reasoning.** Most experts we interacted with had difficulty discussing proposals in an abstract manner. They needed to see illustrations of possible solutions.

**Limited computer skills.** Technical skills of experts were limited, ranging from a satisfactory knowledge of computers as a classic user to a very weak interest for technology in general. None had basic skills in computer programming or similar topics.

**Limited techno-pedagogical expertise.** Stemming from the previous point, while many experts had expertise in pedagogy, all had difficulty fully considering the potential of technology (its opportunities and challenges) in knowledge transmission activities. Many also expressed positive or negative biases about it.

**Need to share progress with other experts not involved in authoring activities.** All experts needed to share their creative efforts with other not involved experts. Sometimes this was to convince a superior or keep him/her informed. In other cases, it was to share progress with colleagues to confirm that the approach made sense or to find new ideas for improving quality.

## 5 Design Recommendations for Expert-Centered Design of Authoring Tools

The common characteristics of domain experts mentioned above may seem obvious to many researchers accustomed to interacting with such individuals. However, we feel it is important to list them explicitly because they explain the genesis of some of the design guidelines we present in this section. Most of these guidelines eventually appear as common sense, which is preferable in usability research. However, as mentioned in the introduction, they are far from being mainstream in authoring tools resulting from research activities and, as stated earlier, in the real world, non expert-friendly designs may have a negative impact on the involvement of experts, on the quality/quantity of their output, and on their willingness to choose and/or persist in using the tool itself.

### 5.1 Keeping Cognitive Load Under Control

As mentioned, authoring activities are secondary objectives for experts. Therefore, the creation of overly complex or intellectually demanding authoring tools appeared to be a risk for the medium and long-term motivation of experts. Indeed, all initiatives we adopted to control the cognitive load during authoring activities have been positively received.

The main one is to adopt a multi-layer tree-based navigation. In other words, the authoring experience starts with a general screen (structuring a graph) that gives access to layers for more detailed authoring (micro-interactions in ADACO, medical action in DPA). However, we also wanted to control the depth of our navigation tree, because the deeper it is, the more information an expert needs to keep in mind. Hence, screens were carefully designed following participatory discussions in order to find a manageable amount of information for experts to change at each level.

Avoiding extension of the graph area was another decision to manage cognitive load. To not frustrate experts with arbitrary limits, we wanted them to be able to create as many nodes as they wanted for their graph. However, we observed early on that overly complex graphs did not lead to better end-user experiences. Prohibiting the graph area from changing dimensions quickly led experts to realize that their own experience became less enjoyable when they overloaded it with nodes, and they naturally self-regulated their creative effort to create simpler graphs that were easier to read/manage, while providing a more enjoyable experience for the end users.

## 5.2 Making Usage Simple and Obvious

Not all experts are comfortable or familiar with information technology. Moreover, they have irregular authoring activities. Considering these two observations, we realized that interfaces had to be as self-explanatory as possible. Thus, most buttons are labelled with a text stressing their function rather than more aesthetic icons whose interpretation is sometimes uncertain.

Evolution of our graph creation mechanism was also an interesting lesson. Initially, the same mouse event (double click) allowed any creation action, sometimes implying the appearance of a contextual menu (e.g., to choose the type of edge or micro-interaction to be created). This was a simple yet effective and intuitive creation approach. At some point, we thought making a better use of mouse possibilities (1 vs 2 clicks, left vs right) would make the authoring experience more productive. Yet, experts whose authoring activities are commonly separated by large periods of time strongly disliked it because they tend to forget how to properly do all actions and there was no visual cues to support them. We eventually retrofitted our graph creation mechanism to the original one.

## 5.3 Visualizing the Process

End-user experiences are centered around dynamic screen changes that illustrate DPA patient state changes or evolutions of ADACO interactive storytelling. However, domain experts are not good at abstraction, and a graph representation is not enough to give them a precise-enough idea of what they are building. Therefore, both authoring tools integrate a simulation screen that allows experts to visualize and test their experience at any time, as if they were using a smartphone.

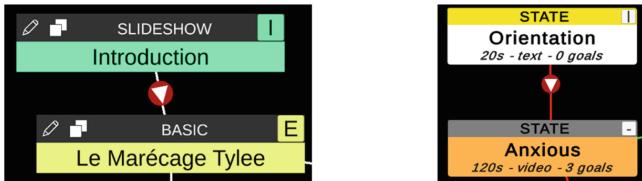
The very structure of the ADACO interface evolves a lot because of its seldom use of many interactive and multimedia technologies in micro-interactions, quests, and tours. In this context and for experts to be aware of what they are creating, a *What You See Is What You Get* approach is adopted. That is to say that each creation screen of the authoring tool includes both an area for editing required information and another one that displays what the resulting interface looks like while incorporating changes in real time.

## 5.4 Facilitating Recollection

As already mentioned, expert activity is irregular. While it is important that authoring tool usability remains as intuitive as possible, the tool should ideally help experts to remember the interactive experience that he/she has been producing so far. Two already discussed features also contribute to this goal.

First, the graph-based representation allows expert to have a rough idea of the experience they are creating, in a few seconds. Beyond the graph structure, nodes also include additional useful information without overly impacting cognitive load as illustrated in Fig. 5.

Node designs similarities exist in both systems. The node title is central, while a button at the top right allows experts to quickly determine the role of each node in its



**Fig. 5.** Closeups on graph nodes in the ADACO (left) and DPA (right) authoring tools.

respective graph (and the corresponding user experience). The role options vary by platform: **Introduction**, **Conclusion**, **Essential**, **Optional** in ADACO; **Introduction**, **Success**, **Failure**, **Pause** for hints, and – (for neutral) in DPA. Distinct colors are associated with each role: in ADACO, a role color impacts all the node background, whereas in DPA, it manifests only in the top bar (an orange background being used to represent the currently selected node).

Differences in node design are also observable. ADACO is characterized by a variety of micro-interactions (aka node types) whose name appear in the top bar whereas all DPA nodes represent states of a virtual patient. Therefore, in DPA, each node includes a brief snapshot describing the maximum duration of the state, the type of multimedia resource used to represent the patient, and the number of goals (i.e., medical actions that need to be performed for the patient to stabilize).

Consequently, the graph structure along with information carried out by each node quickly provide experts with an overview of the experience, and he/she can complement his/her recollection by launching the simulation screen directly from the graph screen.

## 5.5 Facilitating External Sharing

As mentioned, experts commonly need to interact with colleagues and superiors not involved in the authoring process. In these situations, this external individual usually has no expertise with the authoring tool. The interactive content creator must then rely on the intuitive design of our tools to make his/her efforts understood.

In this context, elements we designed to allow an expert to remember work he/she has already undertaken have proven to be very appropriate. Similarly, efforts we made to facilitate visualization of processes were also well received. All the experts with whom we interacted reported the ease with which they can give a first overview of interactive experiences to external collaborators by means of graph representations of our tools, before using the simulation mode to provide a more dynamic and precise evaluation. This approach to communicating with external experts works equally well in a face-to-face setting as it does in a videoconference context.

## 6 Discussions and Conclusion

The five design recommendations we have presented in this paper are the result of two projects dealing with the development of interactive technologies for knowledge transmission. One of them has a commercial purpose while the other one is closer to the

research domain. Moreover, they are in different fields and therefore involved experts with very different profiles.

In preparation for this article, four of them agreed to talk to us. One of them is still using the DPA authoring tool while the other three have used the ADACO tool in the past but are no longer in the organizations where they had access to it. The following is a summary of feedbacks they provided.

All were extremely positive about the tool they had access to, and its overall ease of use, although some pointed out areas for improvement. They also all praised the intuitiveness of the tool and especially the use of graph representations that, for some experts, were close to work strategies they had already adopted on paper. Some stressed that the different layers of authoring imposed by the tool (a form of implicit scaffolding) helped them to better structure their thinking. They pinpointed the importance of being able to simulate experiments they were designing and also emphasized how easy it was for them to remember how to use the tool, even after a long period of interruption. They also confirmed that they felt that suggestions for improvement they made had been considered. Three of these experts expressed real enjoyment in designing interactive experiences with the tool they had access to, with one mentioning that she sometimes missed these creative experiences now that she was working for another organization.

These recent comments confirm the success of the participatory design process that was put in place several years ago to involve domain experts in interactive content creation activities. We are aware that a limitation of this study is that it is less controlled than other research contributions, but it has the parallel advantage of stemming from the Real World, and of reporting incremental improvement over a period of several years.

Although we are very satisfied with the current performance of our tools, there is room for improvement. Their main weakness is that none of them currently allows for synchronous and collaborative authoring of interactive contents. This would be a real improvement since many experts are integrated in teams. Certainly, the addition of such capabilities could open the door to additional design recommendations.

**Acknowledgements.** We would like to thank all the experts who contributed with their comments to the improvement of our two authoring tools over the past few years. We are especially grateful to Clémence Aguéra, Samuel Courtemanche, Marie Toulouse and Jeffrey Wiseman for their recent comments.

## References

1. Murray, T.: Authoring intelligent tutoring systems: an analysis of the state of the art. *Int. J. Artif. Intell. Educ.* **10**, 98–129 (1999)
2. Murray, T.: An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: *Authoring Tools for Advanced Technology Learning Environments*, pp. 491–544. Springer, New York (2003)
3. Dermeval, D., Paiva, R., Bittencourt, I.I., Vassileva, J., Borges, D.: Authoring tools for designing intelligent tutoring systems: a systematic review of the literature. *Int. J. Artif. Intell. Educ.* 1–49 (2018)
4. Aleven, V., et al.: Example-tracing tutors: intelligent tutor development for non-programmers. *Int. J. Artif. Intell. Educ.* **26**(1), 224–269 (2016)

5. Heffernan, N.T., Heffernan, C.L.: The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* **24**(4), 470–497 (2014)
6. Mitrovic, A., et al.: ASPIRE: An authoring system and deployment environment for constraint-based tutors. *Int. J. Artif. Intell. Educ.* **19**, 155–188 (2009)
7. Olney, A.M., Cade, W.L.: Authoring intelligent tutoring systems using human computation: designing for intrinsic motivation. In: International Conference on Augmented Cognition, pp. 628–639. Springer (2015)
8. Barron-Estrada, M., Zatarain-Cabada, R., Tamayo, P., Tamayo, S., Perez-Espinoza, H.: A learning social network with multi-modal affect. In: Proceedings of the 10th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence and Applications (MICAI), pp 163–168 (2011)
9. Barron-Estrada, M., Zatarain-Cabada, R., Zatarain-Cabada, R., Barbosa-Leon, H., Reyes-Garcia, C.A.: Building and assessing intelligent tutoring systems with an e-learning 2.0 authoring system. In: Proceedings of the Ibero-American Conference on Artificial Intelligence (IBERAMIA), pp 1–9 (2010)
10. Zatarian-Cabada, R., Barron-Estrada, M., Reyes Garcia, C.A.: EDUCA: a web 2.0 authoring tool for developing adaptive and intelligent tutoring systems using a Kohonen network. *Expert Systems with Applications*. **38**(8), 9522–9529 (2011)
11. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Teaching the teacher: tutoring simstudent leads to more effective cognitive tutor authoring. *Int. J. Artif. Intell. Educ.* **25**(1), 1–34 (2015)



# Towards a Chatbot-Based Learning Object Recommendation: A Comparative Experiment

Wijdane Kaiss<sup>1,2(✉)</sup>, Khalifa Mansouri<sup>1</sup>, and Franck Poirier<sup>2</sup>

<sup>1</sup> Laboratory Signals, Distributed Systems and Artificial Intelligence, ENSET Mohammedia, University Hassan II of Casablanca, Casablanca, Morocco  
wijdane.kaiss@univ-ubs.fr

<sup>2</sup> Lab-STICC, University Bretagne Sud, Vannes, France

**Abstract.** An adaptive learning system aims to provide learning that is adapted to a learner's current status, different from the traditional classroom experience. A key element of an adaptive learning system is the recommendation system, which provides the most suitable resources based on learner profiles. Recommending the most appropriate learning resources to learners has always been a challenge in the field of e-learning. Thus, learners may have difficulties in choosing the appropriate material when faced with a large volume of recommended material during their learning process. This challenge led us to implement a chatbot to help learners improve their learning experience and knowledge. New solutions use artificial intelligence (AI) techniques such as machine learning (ML) and natural language processing (NLP). The use of our chatbot integrated in Moodle, named *LearningPartnerBot*, provides learners a personalized recommendation of learning objects according to two strategies, one based on their knowledge level (KL) and the other based on their learning style (LS). The objective of this article is to compare the learning outcomes obtained after the realization of the two experiments based on these two approaches centered mainly on the use of the *LearningPartnerBot*. Consequently, the approach of recommending learning objects based on the knowledge level gave promising results by guaranteeing a more adapted learning to the learners.

**Keywords:** E-learning · Learning Object Recommendation · Experimental Design · Adaptive Learning · Chatbot

## 1 Introduction

In recent years, educational technology has increased at a rapid rate. Once learning experiences are personalized, e-learning content is getting richer and more diversified [1, 2]. E-learning provides constructive learning outcomes as it allows learners to actively participate in learning at any time and place [3, 4]. Recently, adaptive e-learning has become a commonly implemented approach by higher education institutions. Adaptive e-learning is a learning process in which content is adapted based on student profiles [5, 6].

An adaptive learning system is a recommendation strategy that sequentially makes decisions about what to learn based on currently available information. According to [7], a recommender system in an e-learning environment has the potential to help learners discover relevant learning actions that match an individual learner's profile, and thus, provide learners with personalized content “at the right time, in the right context, and in the right way.”

By providing personalized content, adaptive e-learning environments enhance the quality of e-learning. Adaptive e-learning dynamically changes the level of instruction based on students' knowledge levels and learning styles and personalizes instruction to increase or accelerate a student's success.

However, in e-learning environments, learners may have difficulty in choosing appropriate materials when faced with a large volume of materials during their learning process [8]. Although, the teacher is not able to provide personalized recommendations in real time, either due to availability or due to the number of learners, the more number of learners and resources, the more difficult and complicated it becomes. This is the reason for building our chatbot, named *LearningPartnerBot* that we integrated into Moodle, to have an adaptive learning capable of handling complicated queries and providing relevant learning objects to learners in need. It has been shown in [9] that a chatbot can assist learning with the same benefits as those obtained from a “real” interview. In addition to making the learning process helpful, the conversational chatbot could make students more likely to improve their level because they do not feel judged.

We conducted our experimentation on the C programming techniques module and 71 learners used the *LearningPartnerBot*. At the beginning, we programmed our chatbot to recommend learning objects according to the learners' knowledge level (approach 1: chatbot-based LO recommendation based KL). After the learners used it and obtained their learning outcomes, we reprogrammed it to provide learning object recommendations based on learning style (approach 2: chatbot-based LO recommendation based LS), in order to determine which approach helps to improve the learning experience. This paper presents the comparison between these two approaches, answering our research question: “Which approach has an impact on learners' learning outcomes?”.

The rest of this article is organized as follows: In Sect. 2, discusses previous research related to our topic. In Sect. 3 describes the methodology (the participants of our experimentation, the experimental setup, and the procedure). The experimental results are discussed in Sect. 4, and Sect. 5 concludes this comparative experiment.

## 2 Theoretical Background

### 2.1 Adaptive E-Learning Environments

Adaptive systems are defined as systems that modify their output and behavior based on different characteristics of the user, such as preferences, skills, and personality. Adaptive systems can be described as “the technological component of joint human-machine systems that can modify their behavior to meet the changing needs of their users, often without explicit user instructions” [10].

In the field of human-computer interaction (HCI), adaptability involves adjusting a system, graphical user interface, or content to meet the needs of a user [11–13]. Learning

strategies can be matched and adapted to learners' profiles (their learning styles, knowledge level). The term personalization is also relevant to adaptability; personalizing means designing an object following the needs of a specific user.

Adaptive technologies can be applied to a wide variety of different domains such as e-health, e-commerce, and e-learning. For instance, AEADS provides adverts based on a user's preferences and behavior in the e-commerce field [14]. The CHAIN approach helps users accomplish tasks by incorporating adaptive help and assistance within a user interface [15].

Adaptivity is an essential component of modern e-learning systems [16]. It facilitates student learning by recommending appropriate learning strategies, providing relevant learning materials, and guiding navigation through the material [17]. Adaptive and personalized learning based on knowledge level, preferences, and learning style is still an important educational consideration [18].

Student modeling represents a more focused stream that involves the storage, representation, and maintenance of student attributes such as motivation, knowledge level, and learning style [5, 19].

The authors in [20], proposed an approach for automatic domain modeling in a way that allows the system to take into account different modes of adaptation that provide personalized and adaptive learning material.

Many attempts have been made to build and evaluate adaptive e-learning systems (AESs). However, there is a dearth of studies that consider detailed, carefully designed and controlled experimental evaluations that assess learning effectiveness [5, 16, 18, 21–23].

Many models of learning style exist, the Felder-Silverman learning style model is commonly used, especially in online learning research [21, 24]. The dimensions of the model are comprehensively detailed, and each dimension is associated with one or more learning strategies [25]. The model includes four dimensions, including Processing, Reception, Understanding, and Perception. The Index of Learning Style (ILS) tool that can be used to identify learning styles is also based on this model [26], which consists of 44 questions that have been shown to be effective in identifying the learning style of each learner.

According to the Felder-Silverman model, the Processing dimension (active-reflective) details the technique students use to process information. Active students learn by interacting with and manipulating something and by communicating with their peers. Reflective students think deeply about something before they act.

The Reception dimension (visual-verbal) is concerned with the presentation of information. For example, visual students' learning can be enhanced by using pictures, videos, graphs, and diagrams. Verbal students can be supported by offering oral information and written details. In the article [27], the authors examined the effect of learning with this dimension in mind.

The Understanding dimension (sequential-global) addresses the desired structure of information. Sequential students understand learning material better if it is presented in a linear and logical manner, with each step of the learning process described in detail. Global students learn best when they are provided with an overview of the information before being given details.

The Perception dimension (sensory-intuitive) concerns the types of information preferred. Concrete learning resources are more likely to benefit sensory students, while abstract resources help intuitive students develop their understanding of the concept being studied. Examples of concrete information include facts, examples, simulations, and interactive lessons. Examples of abstract information include mathematical notations, definitions, and theories.

## 2.2 Recommendation Through Chatbot in E-Learning

According to [28], academic advising in China is an essential support service for students in higher education, but it is not always accessible to learners due to the lack of confidential information, availability of the advisor, or other financial constraints. To overcome these problems, the authors proposed an interactive chatbot called “EASE-selective”, dedicated to serve many learners at the same time and at any time of the day and to complement existing academic advising. The implemented chatbot provides the same expected advising service for all students.

Other research has confirmed that chatbots can be virtual companions for users intended to resolve availability issues, provide support and customer assistance [29, 30] powered by artificial intelligence.

In this field of education, chatbot as a learning tool for teaching is still in its infancy [31]. However, to our knowledge, there is no chatbot that has been used in the recommendation of educational resources on the Moodle platform. Recommending appropriate educational resources has become a current challenge for educators and researchers to develop new ideas to help learners improve their learning process.

This is the reason for building our chatbot that we named *LearningPartnerBot*, and integrated it into the Moodle platform. The chatbot we have implemented is dedicated to recommending learning objects to learners according to two approaches, one based on their knowledge level (KL) and the other based on their learning style (LS). This is the objective of this article, which is to compare the learning outcomes obtained after the realization of the experiments based on these two approaches.

The following section describes the methodology such as the participants of our experiment, the experimental setup, and the research experimental design.

## 3 Methodology

### 3.1 Experimental Setup

The e-learning platform we used for our experiment is Moodle. It is a flexible and secure platform that can be adapted and extended for a variety of possibilities to create personalized learning environments. It can even run on mobile devices.

We implemented the *LearningPartnerBot* chatbot and integrated it into Moodle, which aims to provide recommendations of learning objects adapted to learners' profiles. The tool we used to implement it is Google DialogFlow, which allows users to develop human-computer interaction technologies capable of handling natural language understanding (NLU).

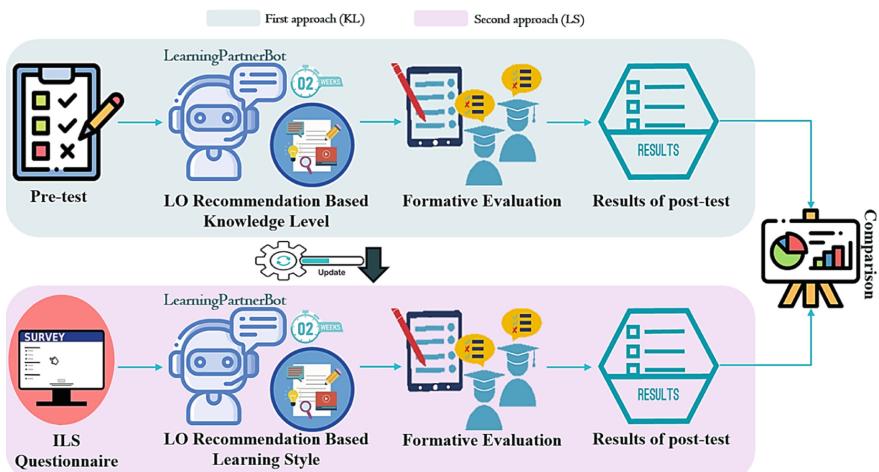
### 3.2 Research Participants

We programmed our chatbot on the C programming techniques module to provide the recommendation of learning objects. We have chosen the Ecole Normale Supérieure de l'Enseignement Technique de Mohammedia (ENSET), an engineering school in Morocco to conduct our experiment. The sample included 71 students (52 males and 19 females), who participated in this study. The participants were first-year engineering students in the Software Engineering and Distributed Computing Systems (GLSID) and Big Data and Cloud Computing Engineering (BDCC) fields, and ranged in age from 20 to 21. Participants were allowed to use our *LearningPartnerBot* chatbot on any device of their choice.

It would have been preferable to form two groups, the first receiving recommendations according to their knowledge level and then according to their learning style, the second receiving recommendations according to their learning style and then according to their knowledge level. But since we had less than hundred learners, in order to have more meaningful results, we have chosen to make only one group.

### 3.3 Procedure

As shown in Fig. 1, our experiment is divided into two. The first one is the approach of recommending learning objects according to the knowledge level (KL) of the learners, and the second is the one based on their learning style, in order to compare the results obtained from each approach.



**Fig. 1.** Research Experimental design.

Participants will pass a pre-test in C programming to determine their knowledge level and identify the concepts they have not mastered and on which the *LearningPartnerBot* chatbot will rely to provide personalized learning object recommendations based on

their level. The use of our chatbot will be for two weeks for the first approach (the one based on KL).

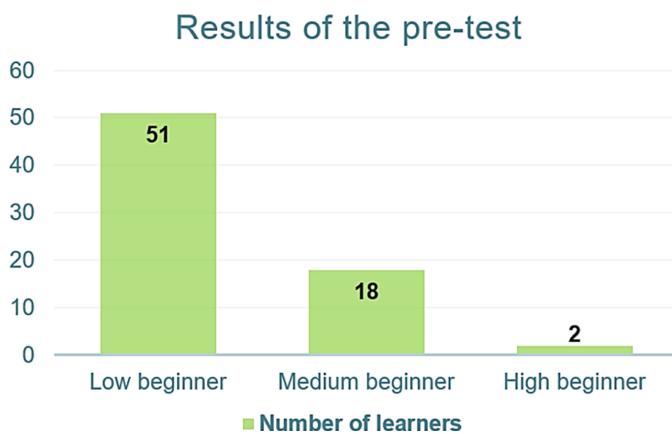
After this first experience, participants will have to answer the ILS questionnaire [32] to determine their preferred learning style. Thus, we reprogram our chatbot to be based on the learners' learning style characteristics in order to provide them learning object recommendations. After two weeks of participants learning, using our chatbot, they will pass a formative evaluation as a second post-test.

The learning results from these two post-tests will be compared to determine in which approach the learners have progressed the most.

## 4 Results and Discussion

### 4.1 Recommendations Based on the Learners' Knowledge Level

As presented in the previous section, at the beginning, we programmed our chatbot to provide personalized recommendations of learning objects based on the learners' knowledge level. In their initial login into Moodle, they passed a pre-test in C programming using *QuizCb0t* [33] to determine their knowledge level in all C concepts. Figure 2 shows the results obtained in this pre-test. Out of 71 learners, 72% of learners (51 learners) are at the low beginner level, 25% of learners (18 learners) are at the medium beginner level, and 3% of learners (2 learners) are at the high beginner level (as we had a limited number of participants, we give the results as a percentage without further statistical analysis).

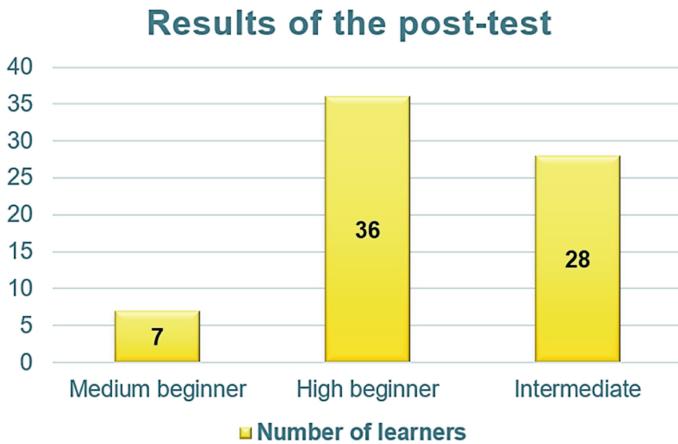


**Fig. 2.** Results of the pre-test.

Following the pre-test, *LearningPartnerBot* provided learning objects to the learners according to the level they belong to (on the concepts they did not master). During the two weeks of the first experiment, that of the approach chatbot-based Learning Object recommendation-based Knowledge Level, the learners were given to use the learning

resources on the different concepts of C programming techniques that were added in Moodle, such as exercises, lessons, quizzes.

After these two weeks of learning with support of our chatbot (using the recommendations it provides to them), they were given a formative evaluation considered as a post-test. The result obtained is shown in Fig. 3.



**Fig. 3.** Results of the post-test (approach 1).

It shows that there is an improvement in the results, the low beginner level became nonexistent and the medium beginner level decreased from 25% (18 learners) in the pre-test to 10% (7 learners) in the post-test. For the high beginner level, there was an improvement from 3% (2 learners) to 51% (36 learners). For the intermediate level, there is an increase from 0% (no learners) to 39% (28 learners).

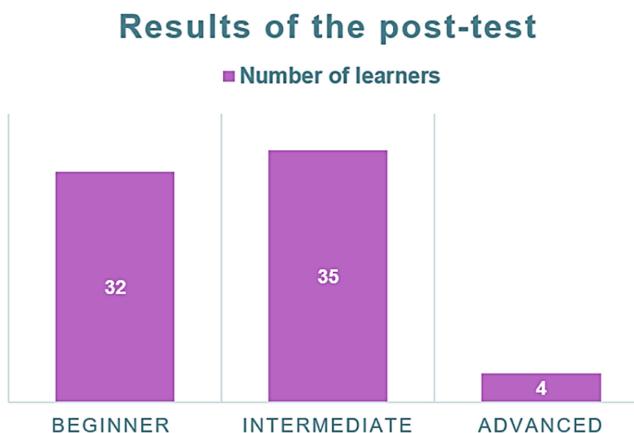
#### 4.2 Recommendations Based on the Learning Style

Then, in the third week, we moved on to the second experiment. To determine the preferred learning style of the learners, we used the ILS questionnaire because it is the most widely used in educational systems and in technology-enhanced learning and some researchers even argue that it is the most appropriate learning style model for use in adaptive learning systems such as [34, 35], as well as being easy to implement [36, 37].

After the participants answered this questionnaire, we found that the learning style to which most of the learners belong is the “visual learning style”, followed by the “verbal learning style”, those of the Reception dimension (explanations of different dimensions and learning styles are presented in Sect. 2.1). The learning objects are provided in different formats and media to suit the learning styles of each learner. These can be text documents (e.g., pdfs), presentations (e.g., Powerpoint slides), videos, etc. For example, a visual learner will prefer to watch a video rather than read a pdf document, while a verbal learner will choose the opposite.

Our chatbot relies on these questionnaire results to provide recommendations of learning objects according to the preferred learning style for each learner. The pre-test for this second approach (chatbot-based LO recommendation-based LS) is the result of the last formative evaluation done which is the result of the post-test (Fig. 3) of the first approach (chatbot-based LO recommendation-based KL).

After using the chatbot for two weeks, the learners are submitted to another post-test to evaluate their learning result. The results are shown in Fig. 4.

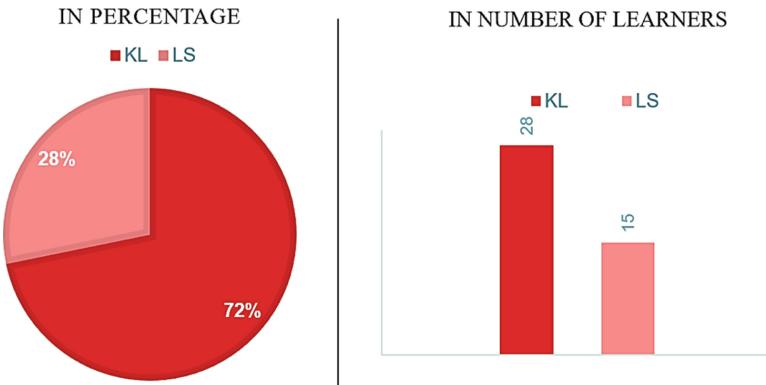


**Fig. 4.** Results of the post-test (approach 2).

Out of 71 learners, there are no more learners classified at the medium beginner level, so the number of learners who were classified at both levels (medium beginner, high beginner) in the pre-test became only classified at the high level in the post-test, which is why I renamed the high beginner class to beginner class. The beginner level decreased from 61% in the pre-test (10% medium beginner and 51% high beginner) to 45% (32 learners) in the post-test. For the intermediate level, there was an improvement from 39% (28 learners) to 49% (35 learners). For the advanced level, there is an increase from 0% (no learners) to 6% (4 learners).

Figure 5 shows a comparison of the learners who improved in learning outcomes after using our chatbot in the two approaches, 72% of the learners (28 learners) improved in the first approach, while just 28% of the learners (15 learners) improved in the second approach, which indicates that the chatbot-based LO recommendation based KL approach improved the learners' learning outcomes compared to the learning style-based approach.

We asked learners after each post-test the following questions: "Q1: Were the recommendations provided by our *LearningPartnerBot* chatbot useful in your learning process?"; "Q2: How would you rate the overall experience with *LearningPartnerBot*?", to find out their satisfaction towards the recommendations and their satisfaction towards the chatbot for each of the two approaches. Table 1 shows the distribution of responses to each question.



**Fig. 5.** Comparison of learners who progressed in learning outcomes between KL and LS.

**Table 1.** Evaluation questions.

	Recommendation satisfaction		Chatbot satisfaction	
	Approach 1 (based on KL)	Approach 2 (based on LS)	Approach 1 (based on KL)	Approach 2 (based on LS)
Very interesting	92%	73%	94%	91%
Interesting	8%	25%	6%	9%
Not at all interesting	0%	2%	0%	0%

It is shown that learners were satisfied with recommendation in the first case (chatbot-based LO recommendation-based KL) more than in the second case (chatbot-based LO recommendation-based LS). Thus, based on these results, we conclude that our *LearningPartnerBot* chatbot was perceived as interesting and helpful in providing learners personalized recommendations of learning objects.

## 5 Conclusion

In this paper, we compared the learning outcomes obtained after the realization of two experiments based on two approaches, one concerns the use of the chatbot-based learning object recommendation according to the knowledge level (KL) of the learners, and the other concerns the use of the chatbot-based learning object recommendation according to their learning style (LS). The experimentation was conducted on 71 learners, on the C programming techniques module, and the results obtained show that the one based on the knowledge level (KL) improved the learning outcomes of the learners compared to the one based on the learning style (LS). Based on the results obtained from questions asked to the learners after each post-test, which were conducted to assess their satisfaction

towards the recommendations and the chatbot, we concluded that our *LearningPartner-Bot* chatbot is perceived as interesting and useful and shows a positive attitude of the learners. Future work will focus on personalizing learning path recommendations by combining these two approaches together (KL&LS), also based on the use of the chatbot integrated in Moodle, in order to improve the learning experience of learners and evaluate also the specific role of each type of recommendation by doing the experiment at the same time.

## References

1. El-Sabagh, H.A.: The relationship between learning-styles and learning motivation of students at Umm Al-Qura university. Egypt. Assoc. Educ. Comput. J. **8**(1) (2020)
2. Yang, J., Huang, R., Li, Y.: Optimizing classroom environment to support technology enhanced learning. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 275–284. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39146-0\\_24](https://doi.org/10.1007/978-3-642-39146-0_24)
3. Chen, P.S.D., Lambert, A.D., Guidry, K.R.: Engaging online learners: the impact of Web-based learning technology on college student engagement. Comput. Educ. **54**(4), 1222–1232 (2010). <https://doi.org/10.1016/J.COMPEDU.2009.11.008>
4. Lee, J., Song, H.D., Hong, A.J.: Exploring factors, and indicators for measuring students' sustainable engagement in e-learning. Sustainability (Switzerland) **11**(4), 985 (2019). <https://doi.org/10.3390/SU11040985>
5. Afifi Normadhi, N.B., Shuib, L., Md Nasir, H.N., Bimba, A., Idris, N., Balakrishnan, V.: Identification of personal traits in adaptive learning environment: systematic literature review. Comput. Educ. **130**, 168–190 (2019). <https://doi.org/10.1016/J.COMPEDU.2018.11.005>
6. Kolekar, S.V., Pai, R.M., Manohara Pai, M.M.: Prediction of learner's profile based on learning styles in adaptive e-learning system. Int. J. Emerg. Technol. Learn. (iJET) **12**(6), 31 (2017). <https://doi.org/10.3991/ijet.v12i06.6579>
7. Maravanyika, M., Dlodlo, N., Jere, N.: An adaptive recommender-system based framework for personalised teaching and learning on e-learning platforms. In: 2017 IST-Africa Week Conference, IST-Africa 2017 (2017). <https://doi.org/10.23919/ISTAFRICA.2017.8102297>
8. Udupi, P.K., Malali, P., Noronha, H.: Big data integration for transition from e-learning to smart learning framework. In: 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pp. 268–271 (2016). <https://doi.org/10.1109/ICBDSC.2016.7460379>
9. Pérez, J.Q., Daradoumis, T., Puig, J.M.M.: Rediscovering the use of chatbots in education: a systematic literature review. Comput. Appl. Eng. Educ. **28**(3), 1549–1565 (2020). <https://doi.org/10.1002/CAE.22326>
10. Feigh, K.M., Dorneich, M.C., Hayes, C.C.: Toward a characterization of adaptive systems: a framework for researchers and system designers. Hum. Factors **54**(6), 1008–1024 (2012). <https://doi.org/10.1177/0018720812443983>
11. Brusilovsky, P.: Adaptive hypermedia. User Model User-Adapt Interact. **11**(1–2), 87–110 (2001). <https://doi.org/10.1023/A:1011143116306/METRICS>
12. Hauger, D., Augstein, M.: State of the art of adaptivity in e-learning platforms. In: Workshop on Adaptivity and User Modeling in Interactive Systems ABIS (2007)
13. Klašnja-Milićević, A., Ivanović, M., Nanopoulos, A.: Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. Artif. Intell. Rev. **44**(4), 571–604 (2015). <https://doi.org/10.1007/s10462-015-9440-z>
14. Qaffas, A.A., Cristea, A., Mead, M.A.: Lightweight adaptive e-advertising model. J. Univers. Comput. Sci. (2018). <https://doi.org/10.3217/JUCS-024-07-0935>

15. Akiiki, P.A.: CHAIN: developing model-driven contextual help for adaptive user interfaces. *J. Syst. Softw.* **135**, 165–190 (2018). <https://doi.org/10.1016/J.JSS.2017.10.017>
16. Rodrigues, H., Almeida, F., Figueiredo, V., Lopes, S.L.: Tracking e-learning through published papers: a systematic review. *Comput. Educ.* **136**, 87–98 (2019). <https://doi.org/10.1016/J.COMPEDU.2019.03.007>
17. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. *User Model User-Adapt Interact.* **6**(2–3), 87–129 (1996). <https://doi.org/10.1007/bf00143964>
18. Xie, H., Chu, H.C., Hwang, G.J., Wang, C.C.: Trends and development in technology-enhanced adaptive/personalized learning: a systematic review of journal publications from 2007 to 2017. *Comput. Educ.* **140**, 103599 (2019). <https://doi.org/10.1016/J.COMPEDU.2019.103599>
19. Chrysafiadi, K., Virvou, M.: Student modeling approaches: a literature review for the last decade. *Expert Syst. Appl.* **40**(11), 4715–4729 (2013). <https://doi.org/10.1016/J.ESWA.2013.02.007>
20. Simko, M., Bielikova, M.: Lightweight domain modeling for adaptive web-based educational system. *J. Intell. Inf. Syst.* **52**(1), 165–190 (2019). <https://doi.org/10.1007/S10844-018-0518-3/METRICS>
21. Akbulut, Y., Cardak, C.S.: Adaptive educational hypermedia accommodating learning styles: a content analysis of publications from 2000 to 2011. *Comput. Educ.* **58**(2), 835–842 (2012). <https://doi.org/10.1016/J.COMPEDU.2011.10.008>
22. Özyurt, Ö., Özyurt, H.: Learning style based individualized adaptive e-learning environments: content analysis of the articles published from 2005 to 2014. *Comput. Hum. Behav.* **52**, 349–358 (2015). <https://doi.org/10.1016/J.CHB.2015.06.020>
23. Truong, H.M.: Integrating learning styles and adaptive e-learning system: current developments, problems and opportunities. *Comput. Hum. Behav.* **55**, 1185–1193 (2016)
24. Alshammari, M., Anane, R., Hendley, R.J.: Adaptivity in e-learning systems. In: Proceedings – 2014 8th International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2014, pp. 79–86 (2014). <https://doi.org/10.1109/CISIS.2014.12>
25. Felder, R., Silverman, L.: Learning and teaching styles in engineering education. *Engr. Educ.* **78**(7), 674–681 (1988)
26. Felder, R., Spurlin, J.: Applications, reliability and validity of the index of learning styles. *Int. J. Contin. Eng. Educ. Life-Long Learn.* **21**(1), 103–112 (2005)
27. Kollöffel, B.: Exploring the relation between visualizer-verbalizer cognitive styles and performance with visual or verbal learning material. *Comput. Educ.* **58**(2), 697–706 (2012). <https://doi.org/10.1016/J.COMPEDU.2011.09.016>
28. Chun Ho, C., Lee, H.L., Lo, W.K., Lui, K.F.A.: Developing a chatbot for college student programme advisement. In: International Symposium on Educational Technology (ISET), pp. 52–56 (2018). <https://doi.org/10.1109/ISET.2018.00021>
29. Wu, E.H.-K., Lin, C.-H., Ou, Y.-Y., Liu, C.-Z., Wang, W.-K., Chao, C.-Y.: Advantages and constraints of a hybrid model k-12 e-learning assistant chatbot. *IEEE Access* **8**, 77788–77801 (2020)
30. Yang, S., Evans, C.: Opportunities and challenges in using AI chatbots in higher education. In: 3rd International Conference on Education and E-Learning, pp. 79–83 (2019). <https://doi.org/10.1145/3371647.3371659>
31. Gonda, D.E., Luo, J., Wong, Y.L., Lei, C.U.: Evaluation of developing educational chatbots based on the seven principles for good teaching. In: IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), pp. 446–453 (2018). <https://doi.org/10.1109/TALE.2018.8615175>
32. Soloman, B.A., Carolina, N., Felder, R.M.: Index of learning styles questionnaire. *Learning* 1–5 (2012)

33. Kaiss, W., Mansouri, K., Poirier, F.: Pre-Evaluation with a personalized feedback conversational agent integrated in Moodle. *Int. J. Emerg. Technol. Learn.* **18**(6), 177–189 (2023). <https://doi.org/10.3991/ijet.v18i06.36783>
34. Carver, C.A., Howard, R.A., Lane, W.D.: Enhancing student learning through hypermedia courseware and incorporation of student learning styles. *IEEE Trans. Educ.* **42**, 33–38 (1999)
35. García, P., Amandi, A., Schiaffino, S., Campo, M.R.: Evaluating Bayesian networks' precision for detecting students' learning styles. *Comput. Educ.* **49**, 794–808 (2007)
36. Lu, J.: A personalized e-learning material recommender system. *Icita* **2004**, 374–379 (2007)
37. Hwang, G.-J., Sung, H.-Y., Hung, C.-M., Huang, I.: A learning style perspective to investigate the necessity of developing adaptive learning systems. *Educ. Technol. Soc.* **16**, 188–197 (2013)



# Development of a Conversational Agent for Tutoring Nursing Students to Interact with Patients

Tomasz Sosnowski<sup>1</sup>(✉), Moh'd Abuazizeh<sup>1</sup>, Thomas Kirste<sup>1,2</sup>,  
and Kristina Yordanova<sup>1</sup>

<sup>1</sup> University of Rostock, Rostock, Germany

{tomasz.sosnowski,mohd.abuazizeh,thomas.kirste}@uni-rostock.de

<sup>2</sup> University of Greifswald, Greifswald, Germany

kristina.yordanova@uni-greifswald.de

**Abstract.** Conversational Intelligent Tutoring Systems can provide a more interactive and engaging learning experience compared to classical teaching methods. As such, conversational tutoring agents have their place as an additional learning resource for students, especially when a need arises for a student to work through a learning scenario which would be difficult or dangerous to reproduce in real life. Such is the case with the training of caretakers for nursing homes and hospitals, needing to train nurse-patient interaction. In this work we present our approach to the development of a conversational agent intended to simulate a patient. The conversational agent is flexible enough to handle large number of possible conversations with the help of automatic story generation methods. This approach enables us to deploy new conversational agents with no need for manual writing of every possible storyline, while remaining compatible with available commercial off-the-shelf chatbot frameworks. Our work was concluded with a user study validating positive usefulness of this type of a system.

**Keywords:** Artificial Intelligence · E-Learning · Education · Tutoring System · Intelligent Tutoring System · Chatbot · Conversational Agent

## 1 Introduction

Nursing staff shortage is a concerning issue especially during a pandemic as shown in the 2022 World Health Organisation (WHO) report [18]. This problem arises from factors such as population growth, population ageing and lack of qualified nursing faculty [16] An effective teaching strategy that can reduce the load on teaching staff would help in reducing that shortage.

We have developed an Intelligent Tutoring System (ITS) that supports teaching of nursing topics. It provides quiz-like questions to test the students' hard skills such as checking and monitoring vital signs of patients. To test the students' soft skills such as communication, critical thinking and problem solving,

system allows the students to interact with a conversational agent that simulates a patient. This paper focuses on the conversational agent developed for the ITS.

A major part of a nurse's job involves interacting with patients. As such, developing social interaction skills is essential part of the staff training, including communication and empathy skills [3]. The chatbot (*conversational agent*) enables the student to interact with a simulated patient in a controlled environment. Moreover, nurses are often faced with situations that they never had the chance to experience in their course of studies. Simulation of case scenarios with the help of the chatbot can help prepare students for such situations.

Despite the progress of chatbot technologies over the course of the last few years, developing a chatbot that simulates a patient to help student achieve his learning goals is a challenging task. A typical use case for off-the-shelf chatbot frameworks is a relatively simple chatbot for a specific task, such as flight booking agent, or customer service support agent. Available chatbot frameworks are not intended for long, free-flowing conversations, conversational agent able to handle this kind of a conversation would require defining multiple paths of a dialog flow.

In this paper we describe the conversational agent for the ITS, demonstrating how to develop a conversational agent for complex teaching scenarios using off-the-shelf chatbot framework (RASA). In particular, we demonstrate the need for automatic story (interaction scenario) generation as well as how we generate them.

## 2 Related Work

### 2.1 Intelligent Tutoring Systems

An Intelligent Tutoring System (ITS) is a software designed for training and education. Typically, it is *goal-oriented* [15] in sense that it helps the user (in our case, the student) to reach certain educational goal, acquisition of a particular knowledge as opposed to free, unguided exploration. Because of that, the ITS often has a capability of detecting student's mistakes and helping to correct them by providing hints or feedback.

Conversational Agents, known also as chatbots, have found their use in multiple tutoring systems [1, 8], where human-like interaction was considered beneficial for the learning process. However, the existing Conversational Intelligent Tutoring Systems tend to use rather rudimentary approach to conversational agents (typically, pattern-matching user inputs against a set of predefined rules) and do not utilise latest advances in the field of chatbots. For example, LANA [1] uses pattern matching approach with wildcards, to match student's written (in Arabic) answers to rules. When the answer does not match any pattern, cosine similarity is used to find the pattern being the closest match.

Moreover, Conversational Intelligent Tutoring Systems typically simulate the role of a tutor and focus on exam-like questions and answers rather than a natural conversation. In contrast, we aim to simulate a patient, and the student's task is to successfully navigate through entire interaction scenario rather than answer specific knowledge questions.

## 2.2 Chatbots

Chatbot technology has shown considerable progress in recent years, moving from pattern-matching based on regular expressions, like in the original ELIZA [17] to complex systems utilising Machine Learning for better recognition of the user's input [13].

Nevertheless, while there was a considerable progress in Natural Language Understanding, existing commercial chatbot frameworks typically follow strict, rule-based interaction scenarios suitable for question-answering, but not for simulating a complex interaction. They are retrieval-based chatbots, using predefined responses [10, 13].

Finally, there is a recent spur of GPT-3 based chatbots, with ChatGPT<sup>1</sup> being the most famous example. While they offer unparalleled ability to carry on human-like conversation, all their knowledge comes from a language model and they have no means of ensuring correctness of information or looking it up in a database. This leads to some spectacular mistakes, where chatbot can not only provide an information that is factually wrong, but even make up ("hallucinate") facts to suit the dialog<sup>2,3</sup>. Also, their interactions are not directed towards conveying a particular information, but rather social interaction (small talk), what makes this approach (as of now) not particularly useful for the domain of tutoring systems.

This is insufficient for our use case, where we need to construct a flexible Conversational Agent allowing student for a relatively free conversation, but still guiding said student towards certain educational goal, ensuring the successful completion of the exercise.

## 2.3 Serious Games

Serious game can be generally defined as a video game-like software used for purposes beyond entertainment, often education [6], but also for dealing with anxiety, PTSD and other disorders. However, it is not meant to be an explicit exam, thanks to the "game" aspect. As a game, it is typically goal-oriented, with user needing to complete some defined task within a simulated environment. Like other types of games, serious games depend on the user's immersion in the game environment [7, 11].

On the other hand, serious games are often not meant as faithful simulations of the world in its entirety, rather focusing on elements relevant for the task. For example, interactiveness of NPCs is not a primary concern [7] – they are not meant to simulate living persons, but just to create a believable background.

While also goal-oriented and intended for education, our tutoring system is markedly different from serious games in the fact, that it does not employ a simulated game-like world, instead relying only on textual descriptions.

---

<sup>1</sup> <https://openai.com/blog/chatgpt/>.

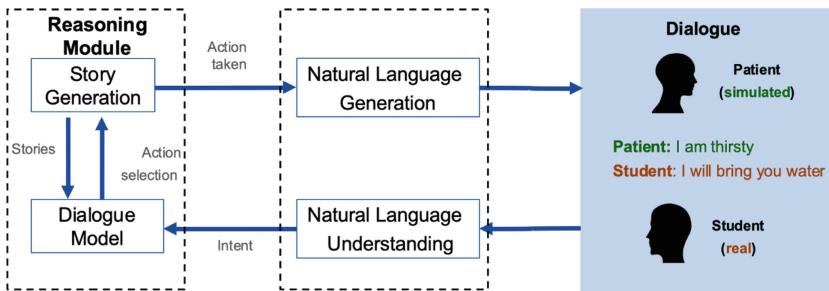
<sup>2</sup> <https://github.com/giuvenc95/chatgpt-failures>.

<sup>3</sup> <https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/>.

### 3 Development

#### 3.1 System Design

A dialogue arises through multiple interactions between a student and a simulated patient for simulating a proper patient (agent) response. The chatbot should understand the intention of a student, reason what would be a valid response and provide the student with a proper response. The conversational agent consists of two main components: the natural language processing and the reasoning modules (see Fig. 1). In the NLP module, the natural language understanding (NLU) takes the *user's input* and maps it to a set of predefined *intents*. Intents are classes, to which input utterances can be assigned based on their meaning (i.e. greeting). Intents along with their confidence values are then used to update the state of the probabilistic model. Based on them as well as generated stories (see Sect. 3.3), the *action* most appropriate for the current situation is selected and delivered. Action is the chatbot's reaction to the user's input – either a textual reply, or terminating the (successful or failed) exercise. Because the dialogue model follows stories, inputs given in the previous steps of the dialog have influence on the action taken. This model is also responsible for handling exceptions - when intent classification fails (input does not match any intent) or when intent does not fit any of the stories at this point, fallback mechanism described in Sect. 3.4 is triggered.



**Fig. 1.** Conceptual schema of the Conversational Agent

#### 3.2 Case Acquisition

Our work started with case descriptions provided by project partners from the German Center for Neurodegenerative Diseases and Hochschule Neubrandenburg, being domain experts in areas of dementia and teaching. Each case had a textual description detailing setting, the patient's background, and the sample dialog (in German). An example of such dialog, translated from German, is provided in Table 1. The aim of this case scenario was to bring the patient to a doctor's appointment and the dialog represents a successful completion of that task. In addition to dialogs with their variants, we have also procured about 20 alternative wordings for each of the possible user's inputs (*intents*).

**Table 1.** Conversation between a nurse (user) and simulated patient

<b>Nurse</b>	Good morning
<b>Patient</b>	Good morning
<b>Nurse</b>	My name is Geraldine and I am a nurse intern
<b>Patient</b>	Nice to meet you, I am Hermann
<b>Nurse</b>	You have doctor's appointment today, please come with me
<b>Patient</b>	Wait a minute, I need to dress my shirt
<b>Nurse</b>	Do you need help?
<b>Patient</b>	Thanks, I am ready
<b>Nurse</b>	Let's go outside, taxi is waiting

### 3.3 Story Generation

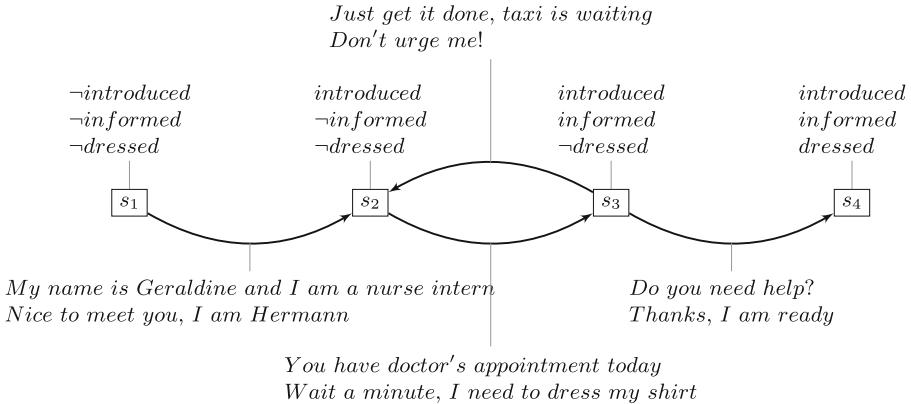
An immediate problem was the linearity of the provided dialogs. Even though we were provided with some alternative dialog variants, they obviously could not cover all possible scenarios of interaction. This was an issue, because in order to provide the users with an interactive simulation of a patient, we had to allow for a lot of freedom in terms of the free dialog flow, rather than just follow a short and simple script. Moreover, this issue is reinforced by the existing off-the-shelf commercial chatbot frameworks, which are primarily designed for short, linear conversations rather than simulating actual free talk with a person. For example, in case the user gets impatient with the patient, we would like the patient to react negatively - such negative development of the previous conversation has been depicted in Table 2. Such alternative scenario could be provided by our project partners, but it is infeasible to manually write down all possible permutations of the dialog with differing moments of urging.

**Table 2.** Example of a dialog, where nurse (user) failed to handle situation appropriately due to unnecessary urging the patient, resulting in failed task

<b>Nurse</b>	You have doctor's appointment today, please come with me
<b>Patient</b>	Wait a minute, I need to dress my shirt
<b>Nurse</b>	Ok, but hurry up
<b>Patient</b>	Don't urge me
<b>Nurse</b>	Just get it done, taxi is waiting
<b>Patient</b>	No! (patient reacts violently)

However, it is possible to generate storylines automatically, if only we define the state space and the transition rules. A state space is the set of all valid states for a certain problem domain. It can be visualised as a graph (see Fig. 2),

where transitions between states correspond to user-patient interactions (here visualised patient's responses). Each state is a set of predicates, which can be either true or false. The actual story generation is then a result of the depth-first traversal of the graph (with constraints on max depth and lack of cycles), resulting in stories being equivalent to paths through state space graph.



**Fig. 2.** State space of a conversation; user needs first to introduce herself to the patient, then inform about doctor's appointment

These transitions are defined as rules, where left-hand-side is a conjunction of preconditions, and right-hand side defines effect of applying the transitions:

$$\begin{aligned} \text{informed} \wedge \text{irritated} &\rightarrow \text{refused} \\ \text{informed} \wedge \neg\text{irritated} \wedge \text{dressed} &\rightarrow \text{outside} \\ \text{informed} \wedge \neg\text{irritated} \wedge \neg\text{dressed} &\rightarrow \text{dressing} \wedge \text{irritated} \end{aligned}$$

In addition, each rule corresponds to a particular *intent* and has a corresponding natural-language response (*action*). This way, a story is generated as an alternating series of intents and responses. Due to the relatively low number of *intents* (on the order of 10 to 20 per scenario), defining rules manually is not a problem. Even this small number of rules allows for a complex dialogue to be generated thanks to the large number of possible paths through the state space graph.

Thus, we had to encode each interaction scenario as a set of rules to be followed by the story generator. This approach is not unlike a plan generation task with planning languages such as PDDL.

Automatic story generation allowed us to significantly increase the number of storylines. For one of the scenarios our initial, manually prepared prototype consisted of 17 different storylines, whereas automatic story generation allowed us to increase this number to 3108.

### 3.4 Fallback Mechanism and State Tracking

Chatbot users tend to end conversation after not receiving the expected answer, due to the frustration stemming from perceived counterproductivity of their actions [14]. In order to avoid the situation where the user is getting irritated due to being stuck, our fallback mechanism provides hints. It is triggered, when the user provides an utterance (input) that either is not recognisable as an intent for the system at all, or if it is classified as an intent that has no use in the currently available context of the conversation. In such case, the system looks into the intents that are relevant for the current context, selects the one that, if provided by the user, would progress the exercise towards the end, and provides a hint based on that intent. The hint depends on how many times the user has provided an input not resulting in any progress in a dialog, as shown in Table 3.

1. There is no hint, just a generic “I don’t understand” reply after the first user’s utterance;
2. If the system has failed to recognise a meaningful intent for the second time, the user is provided with a hint corresponding to the selected best intent – there is a predefined hint text for each existing intent, which should give the user a clue about what to write;
3. After a third attempt, the system provides verbatim input utterance text corresponding to the intent selected as the best by the system, an utterance is chosen at random from a set matching particular intent in the training data.

**Table 3.** 3-step-fallback mechanism

<b>Nurse</b>	Why are you not dressed
<b>Patient</b>	I don’t understand ...
<b>Assistant</b>	Patient doesn’t seem to know the purpose of your visit
<b>Nurse</b>	You should get dressed now
<b>Patient</b>	What do you want?
<b>Assistant</b>	Tell the patient about doctor’s appointment
<b>Nurse</b>	Come on, get dressed now
<b>Patient</b>	What do you want?
<b>Assistant</b>	Try "you have a doctor’s appointment"

Tracking our conversation state is necessary to provide hints when a fallback is triggered. The fact that the story generation is automated enables us to implement state tracker in a form of a **finite state machine**. It also allows us to deploy different heuristics such as goal distance to choose the next best action.

### 3.5 Data Augmentation

Classification of input utterances to intents is a functionality provided by the chatbot framework we have used to implement our conversational agent (see Sect. 4.1). Nevertheless, we had to train the provided classifier using training data. More precisely, we had to provide several example sentences for each *intent*, where all sentence examples for a given *intent* are different wordings expressing the same meaning. While the number of *intents* for each scenario was relatively low (on the order of 10 to 20 per scenario), it nevertheless has proven difficult to collect a larger number of sentence examples for training.

Therefore, we have resorted to data augmentation, that is, creating automatically artificial sentence examples in order to increase the size of a training set. Because all example sentences for a given *intent* should express the same meaning, we have decided to use automatic back-and-forth translation to languages other than German in order to expand the dataset.

We managed to collect 20 example sentences per *intent*, they have been split into training and test sets in 4-to-1 ratio (16 sentences for training and 4 for testing), then augmented by generating 5 new sentences for each existing sentence in both sets. Translation was done using MarianMT model<sup>4</sup> [12] to convert each sentence from German into English, Spanish, Bulgarian, Arabic and Polish, then back to German.

The augmented dataset has been used to train the chatbot, in result NLU accuracy has been increased from 78% to 86%.

## 4 Evaluation

### 4.1 Implementation

In order to evaluate our system concept, we have implemented it using one of the available commercial chatbot frameworks – RASA. RASA<sup>5</sup> [4] is an open source framework that provides a set of tools to develop, test and deploy chatbots. Rasa consists of two main components; the Rasa NLU and Rasa Core. The Rasa NLU takes advantage of DIET architecture (Dual Intent and Entity Transformer) [5] to deliver a state of the art NLU engine. The Rasa Core combines different policies to manage dialogue flow. Adopting the RASA framework enabled us to accelerate the development process. Moreover, the RASA flexibility allowed us to extend and customise the pipeline in order to include our own architecture and custom features.

In order to train the RASA model, a developer needs to provide, among others *stories file* and *NLU file*. The NLU file contains training data for a NLU classifier (responsible for mapping input to *intents*) in the form of sets of sample texts grouped into *intents*, while the stories file contains storylines as sequences of intents and actions. Thus, we have provided our generated stories and augmented user's utterance examples. The RASA chatbot always follows one of the

---

<sup>4</sup> [https://huggingface.co/docs/transformers/model\\_doc/marian](https://huggingface.co/docs/transformers/model_doc/marian).

<sup>5</sup> <https://rasa.com/>.

provided storylines, and any deviation (whether caused by an unrecognisable user's utterance, or by recognised intent not fitting to any of the storylines) triggers the fallback mechanism.

## 4.2 NLU

The NLU engine maps textual data provided by the user to an intent. Our goal was to test the performance of the NLU engine while varying the number of training sentences. For a scenario that has 16 possible intents, 20 training sentences were collected for each intent. 5 sentences per intent sentences were used as test data. The rest of the training data were used in 3 experiments where the NLU engine was trained on 5, 10 and 15 sentences per intent respectively. Additionally two experiments were carried by cross validating multiple training and test data splits using the 20 sentences per intent. One experiment was performed on the original model and the other using BERT pre-trained language model.

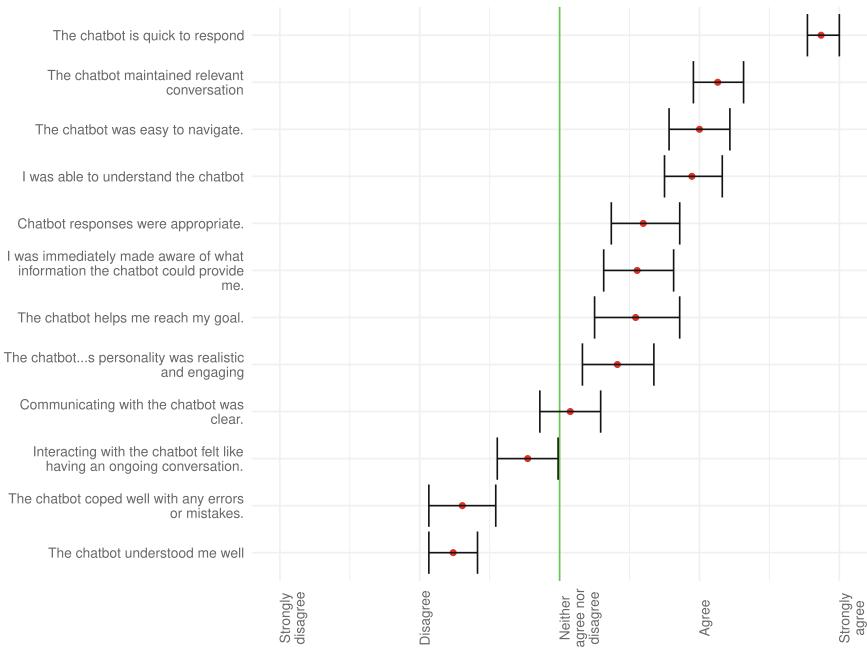
**Table 4.** Intent classification performance

Experiment	Precision [%]	Recall [%]	F1-score [%]
5-training sentences	83	79	78
10-training sentences	87	88	88
15-training sentences	86	85	85
cross validation	91	90	91
BERT + cross validation	82	82	81

The results, presented in Table 4, show that beyond 10 sentences per intent the accuracy can only slightly improve. We were able to achieve good results with 20 sentences per intent. As for the pre-trained model, it decreases the accuracy. This is due to the fact that our test sentences are collected from the data split of our data-collection and a pre-trained model would ambiguise the classification of our intents. A pre-trained model would be better suited for open-context chatbots that cover a larger number of intents.

## 4.3 Usability Evaluation

Based on CUQ (Chatbot Usability Questionnaire) [9] and AUS (Assessing User Satisfaction with Information Chatbots) [2], we designed a questionnaire that evaluates the chatbot functionality, quality of technology acceptance and the learning gain. Users were asked to interact with the chatbot for 3 different scenarios and then answer the 5-Likert scale online-questionnaire. 99 users tested the chatbot and filled the questionnaire, of which 71 were students of computer science (with specialty in AI) and 28 were healthcare students. A single sample bootstrapping t-test was carried out to check if the answers are positive (mean  $> 3$  := neither agree or disagree).

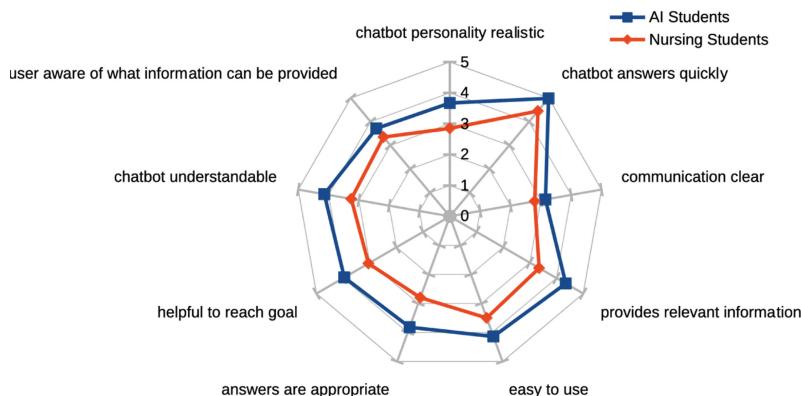


**Fig. 3.** Chatbot questionnaire results

The results (Fig. 3) were positive in terms of the student acceptance of the chatbot technology, they also pointed out to some improvement room in terms of functionality. The chatbot was proven to be beneficial and easy to use learning resource for students. Nevertheless, the chatbot did not understand all the inputs provided, which led to some dialogs not being as smooth as expected.

Besides the questionnaire, an interesting tendency was observed among many students. Despite them being asked to interact with the chatbot for 3 different scenarios, many repeated a scenario multiple times even after completing it successfully. The students were curious to experience different dialog paths, what suggests that our chatbot can be seen as a serious game that combines education and entertainment as an alternative for classical learning methods. This behaviour was especially common among computer science students, who often set on to explore system's capabilities and limitations.

Both groups of students have displayed a correlation in their answers, especially strong for questions about understanding the chatbot. However, the students of nursing subjects tended to be more critical in their gradings than Computer Science students (see Fig. 4).



**Fig. 4.** Comparison between answers provided by students of both groups

## 5 Future Work

In the future, in order to simplify the creation of new scenarios, the story generator will be modified to accept PDDL files as inputs. The complete system can then be deployed on university servers and be used for tutoring, as well as further evaluation.

It is also possible to create a graphic UI tool for scenario planning, where developer would move blocks instead of writing PDDL code.

Future research will also be directed into Natural Language Generation and Affect Control Theory in order to create an affective chatbot with answers that better fit the situation. This development may eliminate the need for manual writing of chatbot answers and hints for transitions between states.

## References

1. Aljameel, S.S., O'Shea, J.D., Crockett, K.A., Latham, A., Kaleem, M.: Development of an Arabic conversational intelligent tutoring system for education of children with ASD. In: 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 24–29, June 2017. <https://doi.org/10.1109/CIVEMSA.2017.7995296>
2. Balaji, D.: Assessing user satisfaction with information chatbots: a preliminary investigation, September 2019. <http://essay.utwente.nl/79785/>
3. Bauchat, J.R., Seropian, M., Jeffries, P.R.: Communication and empathy in the patient-centered care model-why simulation-based training is not optional. Clin. Simul. Nurs. **12**(8), 356–359 (2016). <https://doi.org/10.1016/j.ecns.2016.04.003>, <https://www.sciencedirect.com/science/article/pii/S1876139916300196>
4. Bocklisch, T., Faulkner, J., Pawłowski, N., Nichol, A.: Rasa: open source language understanding and dialogue management (2017). <https://doi.org/10.48550/ARXIV.1712.05181>, <https://arxiv.org/abs/1712.05181>

5. Bunk, T., Varshneya, D., Vlasov, V., Nichol, A.: Diet: lightweight language understanding for dialogue systems (2020). <https://doi.org/10.48550/ARXIV.2004.09936>, <https://arxiv.org/abs/2004.09936>
6. Djaouti, D., Alvarez, J., Jessel, J.P.: Classifying serious games: the g/p/s model. In: Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches, pp. 118–136. IGI global (2011)
7. Gebhard, P., et al.: Serious games for training social skills in job interviews. IEEE Trans. Games **11**(4), 340–351 (2019). <https://doi.org/10.1109/TG.2018.2808525>
8. Graesser, A.C., et al.: Autotutor: a tutor with dialogue in natural language. Behav. Res. Methods Instrum. Comput. **36**, 180–192 (2004)
9. Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., Mctear, M.: Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? In: Proceedings of the 31st European Conference on Cognitive Ergonomics, pp. 207–214. ECCE 2019, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3335082.3335094>
10. Jiao, A.: An intelligent chatbot system based on entity extraction using rasa nlu and neural network. JPHCS **1487**(1), 012014 (2020)
11. Johnson, W.L., Vilhjálmsson, H.H., Marsella, S.: Serious games for language learning: how much game, how much AI? In: AIED, vol. 125, pp. 306–313 (2005)
12. Junczys-Dowmunt, M., et al.: Marian: fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121. Association for Computational Linguistics, Melbourne, Australia, July 2018. <http://www.aclweb.org/anthology/P18-4020>
13. Luo, B., Lau, R.Y., Li, C., Si, Y.W.: A critical review of state-of-the-art chatbot designs and applications. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **12**(1), e1434 (2022)
14. Pham, X.L., Pham, T., Nguyen, Q.M., Nguyen, T.H., Cao, T.T.H.: Chatbot as an intelligent personal assistant for mobile language learning. In: Proceedings of the 2018 2Nd International Conference on Education and E-Learning, pp. 16–21. ICEEL 2018, ACM, New York, NY, USA (2018). <https://doi.org/10.1145/3291078.3291115>, <http://doi.acm.org/10.1145/3291078.3291115>
15. Schwarz, E.: Self-organized goal-oriented tutoring in adaptive hypermedia environments. In: Goettl, B.P., Halff, H.M., Redfield, C.L., Shute, V.J. (eds.) ITS 1998. LNCS, vol. 1452, pp. 294–303. Springer, Heidelberg (1998). [https://doi.org/10.1007/3-540-68716-5\\_35](https://doi.org/10.1007/3-540-68716-5_35)
16. Tamata, A.T., Mohammadnezhad, M.: A systematic review study on the factors affecting shortage of nursing workforce in the hospitals. Nurs. Open **10**(3), 1247–1257 (2023). <https://doi.org/10.1002/nop2.1434>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/nop2.1434>
17. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966). <https://doi.org/10.1145/365153.365168>
18. World health statistics 2022: monitoring health for the SDGs, sustainable development goals. World Health Organization (2022). <https://www.who.int/publications/i/item/9789240051157>



# Integrating an Ontological Reference Model of Piloting Procedures in ACT-R Cognitive Architecture to Simulate Piloting Tasks

Guy Carlos Tamkodjou Tchio<sup>1(✉)</sup>, Marc-Antoine Courtemanche<sup>1</sup>, Ange Adrienne Nyamen Tato<sup>1</sup>, Roger Nkambou<sup>1</sup>, and Valéry Psyché<sup>2</sup>

<sup>1</sup> Université du Québec à Montréal, Montreal, Canada

tamkodjou.tchio.guy.carlos@courrier.uqam.ca

<sup>2</sup> Université TÉLUQ, Quebec, Canada

**Abstract.** To accurately replicate the procedures and actions of piloting an aircraft, it is important to create an intelligent system capable of analyzing and executing tasks using established protocols in the field. In this study, we introduce a cognitive agent based on the ACT-R cognitive architecture that incorporates an ontological reference model into its declarative memory. The purpose of this is to simulate the performance of critical piloting tasks, such as take-off, in a manner similar to that of a human pilot. The agent accomplishes this by utilizing production rules stored in its procedural memory to deduce knowledge captured and formalized by the ontological reference model stored in its declarative memory. Our findings suggest that this approach is a key step towards developing a cognitive agent that can be tested in a real flight simulator, providing insights into how human pilots function in terms of their cognitive and affective behavior.

**Keywords:** ontology · OWL · domain ontology · task ontology · reference model · cognitive architecture · ACT-R · Pyactr · cognitive agent

## 1 Introduction

The resolution of critical situations in an aircraft piloting scenario always depends on the pilots, who may need to assume manual control of the aircraft if necessary [8]. However, piloting an aircraft is a complex task, and pilots can benefit from the support of a reliable artificial system [6]. Despite this, pilots may experience lapses in attention, leading to reduced performance in the cockpit. The causes of inattention can be diverse and may depend on factors such as perceptual, attentional, and emotional issues [15]. Given the decrease in attention that may occur in the cockpit, it is important for pilots to receive cognitive assistance [12]. A cognitive model that can track an individual user's operational

context and cognitive dynamics can serve as the basis for cognitive assistance during operations [16]. Our proposed cognitive model aims to demonstrate the functioning of a human pilot by taking into account the pilot's cognitive, affective, and behavioral dimensions. It should also demonstrate the cognitive cycle of a piloting task while considering the pilot's cognitive, affective, and behavioral states. Currently, our cognitive model simulates the execution of a piloting task based on reference procedures described through an ontology.

To formally represent knowledge related to the pilot's task, task environment, and cognitive processes, ontologies can be used. We have adopted the ontology defined by Courtemanche et al. [10], which serves as a reference model for piloting tasks and models complex procedural knowledge related to aircraft piloting procedures. The ontology allows for the automatic manipulation of knowledge by detailing the necessary information related to the execution environment for each task. It consists of a domain ontology and a task ontology. The domain ontology is specific to the execution environment of the complex task of piloting an aircraft, while the task ontology is the central element of the reference model, linking the different execution parameters. The resolution rules are closely linked to the execution environment, and the decomposition of the task execution with the domain theory is formalized by semantic links between the two ontologies, which supports the execution.

Based on Newell's criteria [2] and Sun's desiderata [5], which define various capacities, properties, and evaluation criteria of cognitive architectures, ACT-R has shown the best characteristics in modeling complex cognitive tasks. Therefore, we have decided to instantiate it in aviation to simulate a human pilot's mind. ACT-R [19] is a comprehensive, science-based cognitive architecture that has produced models representing the processes involved in driving a car, controlling a single-engine aircraft, allocating visual attention in a cockpit, using and acquiring skills for the flight management system, and modeling cognitive assistance in a cockpit [8]. This cognitive architecture integrates the major cognitive functions of memory, learning, perception, and actions, as well as their limits. ACT-R has a production system composed of declarative memory and procedural memory. Declarative memory contains facts such as "Ottawa is the capital of Canada," "Cameroon is a country in Africa," or "8+2=10," while procedural memory contains production rules that represent knowledge about how we do things. Examples of such knowledge include how to play chess, drive a car, or perform an arithmetic operation.

This paper aims to build a cognitive agent capable of performing piloting tasks using an ontological reference model containing expert knowledge. To represent the knowledge from the information related to the task, the task environment, and the cognitive processes of the pilot formally, we use the ontology defined by Courtemanche et al. [10], which models complex procedural knowledge related to aircraft piloting procedures. The ontological reference model is stored in the declarative memory of our cognitive agent, and the written production rules that allow for inferences on this knowledge are stored in its procedural memory. Our cognitive agent simulates the execution of piloting tasks by following the current procedures, using the production rules contained in its procedural

memory to infer the knowledge captured and formalized by the ontological reference model present in its declarative memory. The cognitive agent can track an individual user's operational context and cognitive dynamics, taking into account their cognitive, affective, and behavioral dimensions. Our ultimate goal is to test the cognitive agent in a real flight simulator to demonstrate and explain the functioning of a human pilot in a piloting context.

Our current work is part of a larger project aimed at developing a cognitive agent capable of performing complex piloting tasks in a flight simulator. The agent will be based on information related to the state of the human pilot (including cognitive and emotional factors), the task being performed by the pilot, and the aircraft environment.

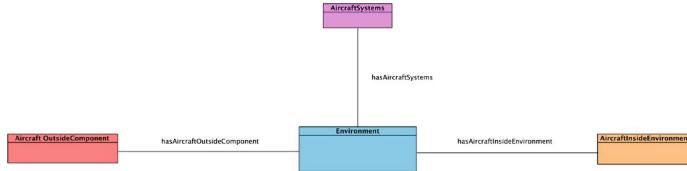
## 2 The Ontological Reference Model

The reference model formalizes knowledge of piloting procedures found in the technical documentation for executing a flight, both under normal and abnormal circumstances. It aims to provide a standardized framework for comparing pilot execution. The model uses Web Ontology Language (OWL) [22] to capture the complex logical relationships between concepts. The reference model consists of two interrelated ontologies: the domain ontology, which contains knowledge about the execution environment, including inside parameters, outside components, and aircraft systems, and the general task taxonomy specific to aviation, which decomposes tasks and anchors them semantically to the domain ontology. The tasks are decomposed and semantically anchored to the domain ontology making the execution framework highly related to the environment. More details about the reference model are available in Courtemanche et al. [10]. This reference model contains the data required for feeding the declarative memory of our cognitive agent. The reference of execution in OWL is manipulated to identify the right task to accomplish. In addition to providing a reference for selecting the right task to accomplish, the reference model is used for structuring the execution environment. The domain ontology is a taxonomy list of the environment's parameters and contains the dynamic data associated with the simulation. The cognitive agent uses this ontology to gather important data about the simulation and push new data to the environment. By using this standardized taxonomy about the environment, the reference of execution is semantically related to the domain of execution and is also providing an efficient way for making the data of the simulation environment more accessible.

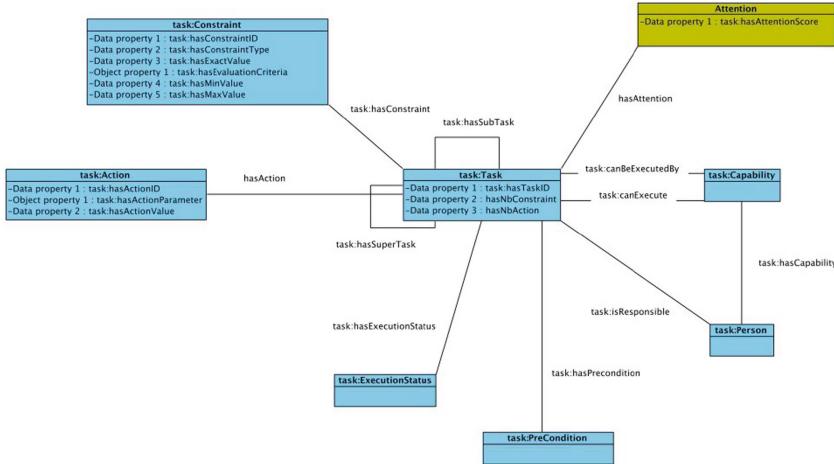
The class diagram depicted in Fig. 2 represents the domain ontology, while Fig. 1 displays the task ontology with an attentional layer [11]. For the purpose of our work, we only consider the classes represented by the blue squares in the task ontology. However, the cognitive agent may utilize the attention class in the future to enhance its functionality.

## 3 ACT-R Cognitive Agent Modeling for Piloting Tasks

The ACT family of architectures, developed by Anderson and his colleagues, is the oldest (since 1973) and best known, with the highest number of scientific



**Fig. 1.** Task ontology with an attentional layer.

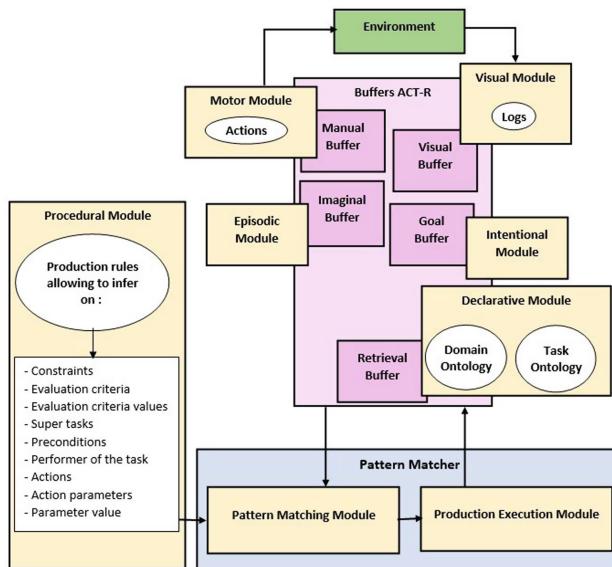


**Fig. 2.** Domain ontology excerpt.

articles devoted to it [18]. The aim of ACT-R (Adaptive Control of Thought - Rational) is to model the human mind [19], exploring the four dimensions of artificial intelligence: thinking like a human, thinking rationally, acting like a human, and acting rationally [7]. ACT-R is a production system consisting of a declarative memory that contains facts and a procedural memory that contains production rules. The declarative module recognizes what is presented to the model and calculates the activation of rules, while the procedural module calculates the utility of each activated rule and triggers the most appropriate one. Cognition emerges from the interaction between procedural and declarative structures.

Based on Anderson's cognitive architecture [19], we propose an architecture for our cognitive agent as shown in Fig. 3. The proposed cognitive agent is situated in the cockpit or an aircraft flight simulator. To interact with its environment, the agent has three main components: the sensory module, the pattern matcher, and the motor module. The sensory module enables the agent to perceive relevant information from its environment, such as the status of the aircraft, the position of the controls, and the actions of the human pilot. The pattern matcher allows the agent to recognize and process the perceived informa-

tion using its declarative and procedural memories. Finally, the motor module enables the agent to act on the environment by sending commands to the simulator or the aircraft systems. Currently, the cockpit is simulated using the ACT-R runtime environment, which provides a platform for testing and evaluating the cognitive agent's performance in realistic scenarios.



**Fig. 3.** Cognitive agent operating architecture.

The cognitive agent has several buffers working together, which are essential components of cognition [3]. Memory will be an indispensable part of our agent, allowing it to perform various functions, including but not limited to: Storing intermediate results of calculations, Learning, and Adapting to a changing environment.

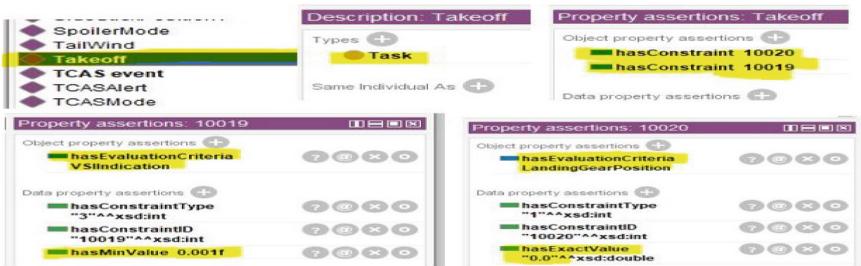
The memory of our agent is divided into short-term and long-term memory. Short-term memory is composed of sensory and working memory. The sensory memory stores and preprocesses recent perceptions, as well as actions to be performed. The Visual Buffer is related to sensory memory and contains logs captured from the environment, biases introduced relating to deviations, incapacities, and pilot behavior. The Manual buffer contains the tasks, constraints, and actions to be performed on the environment after the processing performed by the pattern matcher. The working memory contains items related to the task being executed. The Imaginal Buffer maintains an internal picture of the information associated with the current cognitive process, providing contextual information relevant to the current task. The Goal Buffer contains the goal to be achieved.

Long-term memory, on the other hand, is divided into procedural memory and declarative memory. Procedural memory stores knowledge about actions to take under certain conditions. In our agent, procedural knowledge is represented by a set of “if..then” rules. Declarative memory stores factual knowledge, including the ontological reference model. The Buffer Retrieval contains non-declarative, implicit knowledge.

The pattern matcher is a central module in the cognition of our cognitive agent, and it includes two modules: the production rule selection module and the production rule execution module. The production rule selection module searches the procedural memory for the production rule corresponding to the buffers’ current state. Only one production rule can be executed at any given time. Once executed, this production rule can modify the buffers and change the system’s state. Thus, cognition is a series of production rules triggered by our cognitive agent. The production rules are selected to be executed using the utility learning mechanism, which selects the rule with the highest utility. Additionally, the pattern matcher represents and processes information in a hybrid way that is both symbolic and connectionist [20].

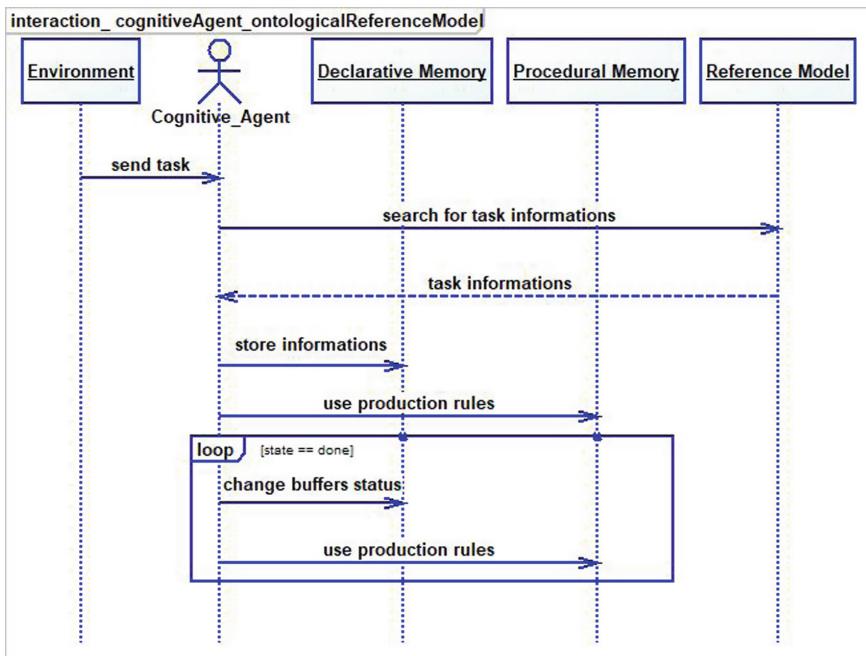
## 4 Management and Execution of Tasks by the Cognitive Agent

From the class diagram (Fig. 2), it appears that executing a task by the cognitive agent involves determining its constraints, its super task, its precondition, its execution manager, and its actions. Figure 4 below shows the information related to takeoff in the Protege ontology editor. This information is composed of task and domain ontology values.



**Fig. 4.** Information related to the Takeoff task.

In UML (Unified Modeling Language) formalism, a sequence diagram is an interaction diagram that expresses dynamically how different objects collaborate. In Fig. 5, we use this diagram to present the functioning of our cognitive agent in executing a task and its interactions with the reference model.



**Fig. 5.** Sequence diagram of the execution of a task by the cognitive agent.

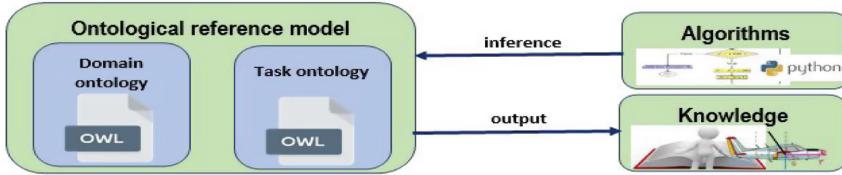
## 5 Methodology, Results and Model Validation

### 5.1 Methodology

The ACT-R theory has an official implementation in LISP and has subsequently been implemented in several programming languages, including Java (jACT-R and Java ACT-R), Swift (PRIM), Python2 (ccm), and Python3 (pyactr) [4]. The latest implementation, pyactr, is very close to the official Lisp implementation and was developed to facilitate the transfer of ACT-R modeling skills between Python and Lisp, and due to the popularity of the Python programming language. Pyactr is a Python library used to create and run cognitive models based on ACT-R, while Owlready2 is a Python module for ontology-oriented programming that allows loading OWL 2.0 ontologies and manipulating them transparently in Python.

We implemented our cognitive agent using the pyactr library and loaded the ontological reference model into the declarative memory of our agent using owlready2 to infer and perform piloting tasks. To perform a task, our cognitive agent retrieves it and searches the ontology for related information, which it places in its declarative memory. It also places the objective to be achieved in its goal buffer and uses its imaginal buffer to maintain an internal image of the information associated with the task being performed. The agent places the production rules in its procedural memory and uses the pattern matcher, a

central module in the cognition of our cognitive agent, to choose the production rule corresponding to the current state of the buffers to execute. Once executed, the system's state changes and the mechanism starts again until the goal buffer is in the desired state. Figure 6 presents the knowledge extraction in the ontological reference model by the ACT-R cognitive agent in a simple way.



**Fig. 6.** Inference made by ACT-R cognitive agent.

The Algorithm 5.1 gives details on how the cognitive agent executes a task. The functions and parameters used in the algorithm have the following meaning:

- **infos** is a variable that contains for the current task: the super task, the precondition, the performer, constraints and actions.
- **receive()** is a function that takes a task as a parameter and saves it in the declarative memory of our cognitive agent
- **pattern\_matcher(X)** searches in the procedural memory of the cognitive agent the rules allowing to execute the parameter X of the current task.
- **search\_information\_in\_referenceModel(X)** places in the **Infos** variable the information about: super task, precondition, performer, constraints, and actions.
- **print\_execution\_time(X)** displays the execution times of X.



**Fig. 7.** Task cognitive cycle.

---

**Algorithm 1.** Algorithm of task execution

---

**Input:** Reference model**Parameter:** task**Output:** execution time, precondition, performer, super task, actions, parameters, constraints, evaluation criteria

```

1: receive(task).
2: infos = search_information_in_referenceModel(task).
3: pattern_matcher(infos.superTask).
4: print_execution_time(infos.superTask).
5: pattern_matcher(infos.preconditions).
6: print_execution_time(infos.precondition).
7: pattern_matcher(infos.performer).
8: print_execution_time(infos.performer).
9: if len(infos.constraints) > 0) then
10:   pattern_matcher(infos.constraints).
11:   print_execution_time(infos.constraints).
12: end if
13: if len(infos.actions) > 0) then
14:   pattern_matcher(infos.actions).
15:   print_execution_time(infos.actions).
16: end if
17: return buffers

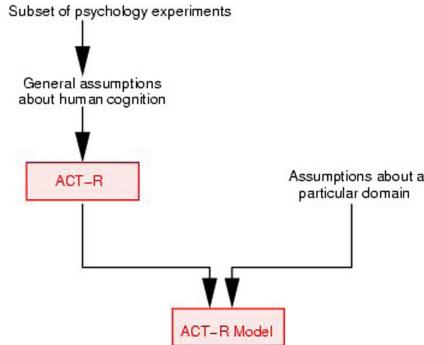
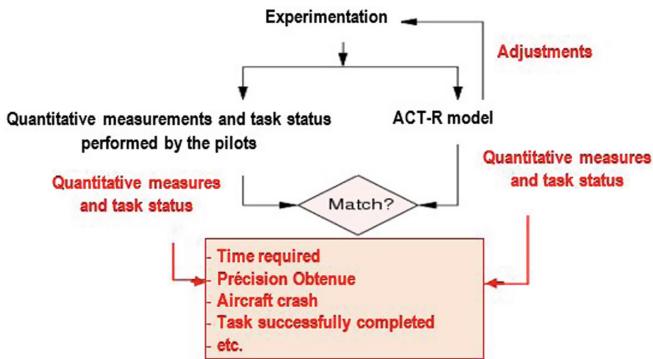
```

---

## 5.2 Results and Model Validation

The experiments carried out show that the model implemented successfully executed a piloting task by demonstrating the cognitive cycle of solving the task, as shown in the example of Fig. 7, which represents a part of the complex task of takeoff. The obtained results demonstrate that the cognitive agent verifies the constraints and performs the actions in a manner similar to a human expert over time. According to [10], the terminology elements represented by the reference model are the same as those of the X-Plane simulation environment and the A320 aircraft. Therefore, to verify the results provided by our cognitive model in a real aircraft flight environment, we plan to deploy the developed models with the X-Plane flight simulator. However, with some adjustments, it may also be possible to use our cognitive agent in a real flight simulator. As ACT-R is a scientifically based theory for modeling human cognitive performance [9], the deployment process will follow the scheme shown in Fig. 9.

As described in [20], the validation will consist of collecting the quantitative measurements produced by the ACT-R cognitive agent in the X-plane simulation environment and will compare them to the quantitative measurements made by the pilots. Aviation experts will then be invited to validate the results produced by the cognitive agent in X-plane. Thus validated, the cognitive agent will be used as a pilot simulating a piloting task in a particular context.

**Fig. 8.** Validation of an ACT-R model [20].**Fig. 9.** Deployment and tests.

## 6 Discussion and Open Problems

The results demonstrate that the cognitive agent can execute and present the cognitive cycle of a complex piloting task, providing insight into the cognitive processes involved in human piloting. Integrating an ontological reference model into the ACT-R cognitive architecture allows the cognitive agent to simulate piloting tasks according to expert procedures, similar to a human pilot. The appropriate rule to execute a piloting task at a given time is selected using utility learning to select the rule with the highest utility. However, as the piloting task is simulated, the ACT-R pattern matcher determines the time to check a constraint or execute an action.

According to Varela [1], cognition is embodied, situated, and social. Embodied cognition is determined by the structure of the organism, with the body being in the brain and not just the brain in the body. Situated cognition means that the environment's state determines cognition. Social cognition means that cognition is a collective cognitive system that emerges from the interaction of individual cognitive systems. One open problem is situating the cognitive agent

in an environment like Xplane. Another challenge is using the attentional dimension of piloting tasks presented in the task ontology (Fig. 1) to execute a piloting task.

## 7 Conclusion

In this paper, we introduce a cognitive agent based on the ACT-R cognitive architecture that utilizes an ontological reference model to perform tasks in the aviation domain. The declarative memory of the agent captures and formalizes knowledge from the domain ontology, which represents the internal and external environments and aircraft navigation systems, and the task ontology, which captures and formalizes piloting procedures. Using a class diagram, we modeled the complex structure of a task and the relationships between its components. When executing a task, the cognitive agent searches the knowledge formalized by the reference model to perform the task. We presented its behavior and interactions with other components using a sequence diagram. Our simulations demonstrated the cognitive execution cycle of a complex piloting task, and the results showed that the agent executed the task similarly to a human pilot.

As the reference model continues to incorporate the cognitive dimension associated with each task, a second perspective is to develop a version of the cognitive agent that integrates parameters such as the pilot's attention or emotions. Additionally, since cognition is embodied, social, and situated, a third perspective is to immerse the cognitive agent in a simulator such as Xplane or other flight simulators with suitable adaptations. Finally, a fourth perspective is to use the cognitive agent to manage deviations in a cockpit.

**Acknowledgement.** We acknowledge the support of CRIAQ, the Natural Sciences and Engineering Research Council of Canada (NSERC), CAE, Bombardier, and BMU.

## References

1. Varela, F.J., Thompson, E., Rosch, E.: *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, Cambridge (1991)
2. Newell, A.: *Unified theories of cognition*, Coll. "William James lectures 1987", Cambridge, Mass., Harvard University Press, pp. 17–18 (1990)
3. Atkinson, R.C., Shiffrin, R.M.: Human memory: a proposed system and its control processes. In: *Psychology of Learning and Motivation*, vol. 2, pp. 89–195. Academic Press (1968). [https://doi.org/10.1016/S0079-7421\(08\)60422-3](https://doi.org/10.1016/S0079-7421(08)60422-3)
4. Brasoveanu, Adrian, Dotlačil, Jakub: Computational Cognitive Modeling and Linguistic Theory. LCM, vol. 6. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-31846-8>
5. Sun, R.: Desiderata for cognitive architectures. *Philos. Psychol.* **17**(3), 341–373 (2004)
6. Insaurralde, C.C., Blasch, E.: Uncertainty in avionics analytics ontology for decision-making support. *J. Adv. Inf. Fusion* **13**(2), 255–274 (2019)

7. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Pearson, Boston (2021)
8. Oliver, W., Marc, H., Nele, R.: ACT-R model for cognitive assistance in handling flight deck alerts. In: International Conference on Cognitive Modelling, Montreal, Canada (2019)
9. Smart, P.R., Scutt, T., Sycara, K., Shadbolt, N.R.: Integrating ACT-R cognitive models with the Unity game engine. GI Global, Hershey, Pennsylvania, USA (2016)
10. Courtemanche, M.A., Tato, A., Nkambou, R.: Ontological reference model for piloting procedures. In: Crossley, S., Popescu, E. (eds.) Intelligent Tutoring Systems (ITS 2022). LNCS, vol. 13284, pp. 95–104. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_9](https://doi.org/10.1007/978-3-031-09680-8_9)
11. Ghaderi, M., Courtemanche, M.A., Ben Abdessalem, H., Nkambou, R., Frasson, C.: Attentional tasks model: a focus group approach. In: Krouská, A., Troussas, C., Caro, J. (eds.) Novel and Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022). LNNS, vol. 556, pp. 297–307. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-17601-2\\_29](https://doi.org/10.1007/978-3-031-17601-2_29)
12. Estes, S., Burns, K., Helleberg, J., Long, K., Stein, J., Pollack, M.: Digital copilot: cognitive assistance for pilots. In: Proceedings of the AAAI Fall Symposium on Cognitive Assistance in Government and Public Sector Applications (2016)
13. Oltramari, A., Lebiere, C.: Mechanisms meet content: integrating cognitive architectures and ontologies. In: AAAI Fall Symposium: Advances in Cognitive Systems (2001)
14. Abrahão, E., Hirakawa, A.R.: Task ontology modeling for technical knowledge representation in agriculture field operations domain. In: Proceedings of the 2017 Second International Conference on Information Systems Engineering (ICISE), pp. 12–16 (2017)
15. Dehais, F., Roy, R.N., Scannella, S.: Inattentional deafness to auditory alarms: inter-individual differences, electrophysiological signature and single trial classification. Behav. Brain Res. **360**, 51–59 (2019). <https://doi.org/10.1016/j.bbr.2018.11.045>
16. Zhang, Z., Russwinkel, N., Prezنسki, S.: Modeling individual strategies in dynamic decision-making with ACT-R: a task toward decision-making assistance in HCI. Procedia Comput. Sci. **145**, 668–674 (2018). <https://doi.org/10.1016/j.procs.2018.11.064>
17. Taatgen, N.A.: Cognitive modelling: a new look at individual differences. Dutch J. Psychol. 167–176 (1999)
18. Kotseruba, I., Tsotsos, J.K.: 40 years of cognitive architectures: core cognitive abilities and practical applications. Artif. Intell. Rev. **53**(1), 17–94 (2018). <https://doi.org/10.1007/s10462-018-9646-y>
19. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychol. Rev. **111**(4), 1036–1060 (2004)
20. Raluca, Budiu: About ACT-R (2013). <http://act-r.psy.cmu.edu/about/>
21. Google Research: Google colab notebook (2023). <https://colab.research.google.com>
22. W3C: OWL 2 web ontology language: structural specification and functional-style syntax (second edition) (2012). <https://www.w3.org/TR/owl2-overview>



# Generating Pedagogical Questions to Help Students Learn

Oleg Sychev<sup>(✉)</sup> and Marat Gumerov

Volgograd State Technical University, Lenin Ave, 28, Volgograd 400005, Russia  
[o\\_sychev@vstu.ru](mailto:o_sychev@vstu.ru)

**Abstract.** One of the important features of intelligent tutoring systems is providing feedback so that the learners who have problems can understand and fix their errors. Two popular approaches to implementing it are demonstrating worked examples and providing explanatory feedback about the errors the learner made. Pedagogical questions - small, specific questions aimed at the thought process required to make the correct conclusion are often overlooked. Generating pedagogical questions to help the learner is beneficial because answering questions stimulates information retrieval and thinking. It also gives the tutoring system detailed information about the learner's misconceptions. However, creating systems of pedagogical questions manually requires a lot of effort, even in relatively simple domains. In this paper, we propose four kinds of pedagogical questions that can be generated from a formal description of the taught skill, provide examples of pedagogical questions, and discuss the advantages and disadvantages of their use.

**Keywords:** Pedagogical questions · Subject-domain modeling · Intelligent tutoring systems

## 1 Introduction

One of the important tasks of any tutor - human or software - is supporting students who make errors when solving problems and are stuck because they do not understand why their solution (or a step in solution) is wrong and how to do it correctly. A number of pedagogical interventions can be used to help students. The tutor can point out the relevant part of lecture material (or textbook), demonstrate how to solve this problem correctly, explain why the student's answer is wrong, or ask a series of small topic-related questions to let the student remember the necessary facts and realize the causes of their error.

Most of these methods are implemented in intelligent tutoring systems and other educational helper applications. Grevise et al. describe a system that provides links to the descriptions of relevant concepts when students make syntax errors in their programs [6]. Demonstrating correct actions is known as hinting [24] or showing worked examples; it is used to build example-tracing tutors [1].

---

The reported study was funded by VSTU, project number 60/438-22.

Generating smaller questions is a less-used approach. Razzaq and Hefferman [19] describe a technique of using scaffolding questions and prove its efficiency, but they do not go farther than breaking the problematic problem-solving step into smaller steps and do not provide a technique for identifying of generating scaffolding questions. Pedagogical questions can do more by assessing the line of reasoning required to perform a step correctly.

Small questions about the knowledge necessary to solve the task at hand stimulate thinking and information retrieval and can lead to realizations and better absorbing information than hints, simply telling the student what is true. Pedagogical questions can turn interactions with the learning system into a dialogue with the tutoring agent which brings it closer to human communications. However, the question-generating strategy must be optimized to avoid repeatedly asking dull, trivial questions until the student reveals that their knowledge is poor enough for these questions to be useful.

Software engineering education is a good domain to implement pedagogical questions because it includes learning a significant number of well-defined thought processes that must be learned precisely: from understanding an expression to designing databases and applications. In the previous work [7], we developed a method of building sets of reasoning errors that are possible when solving a learning problem based on the formal description of the line of thought required to solve the problem. In this paper, we propose a subsystem of automatically generating pedagogical question sequences that determine which errors are factually made based on the same data.

## 2 Related Works

The three main venues of research about generating pedagogical questions in automatic tutoring systems concern automatic tutors, scaffolding techniques, and question generation.

### 2.1 Automated Tutors in Software Engineering Education

A big number of automatic tutors have been developed for software engineering education, from constructing expressions [10] to writing simple programs [11]. Many of these tutors rely on animation and visualization of program execution to demonstrate how it works.

Jeliot 3 [15] is a program execution visualization tool for teaching Java programming language. It executes program code step-by-step, similarly to a debugger, and produces animations, visualizing the inner processes of the code's execution. Jeliot presents basic functionality which many other code visualization systems expand; Jeliot itself can be used for self-learning or classroom demonstrations.

Putting the learner in the position of a program's executor is another approach to enhance teaching programming. Both UUhistle [21] and TRAKLA2

[14] systems feature exercises, built upon the program’s visualization by making students interact with data structure’s visual representations, manipulating them to match the way a computer will execute the given program. This can be enhanced by pop-up multiple choice questions, as it is done in VILLE [18] – a program visualization system where an exercise author can create pop-up questions to appear at certain states of program visualization. These approaches are good to engage learners and integrate knowledge about program execution, but the only form of feedback they give is visualizing the current program state and its changes. If the learner does not know what to do next, they need a way to understand it step by step.

Another approach to automated tutoring is represented by the Problets series of automatic tutor developed by A. Kumar [5]. They concentrate on different aspects of programming, allowing students to learn different topics one by one instead of showing the entire process of program execution. These tutors split the answer to a learning problem into smaller sub-answers (steps) that the learner can enter [9,13]. Immediate feedback is provided by color-coding answers to steps based on their correctness [12,13]: correct, incorrect or incomplete [3,12], sometimes accompanied by text messages, telling the learner what was wrong. An on-demand feedback option is also often present [9,12], providing detailed description of the correct way to solve a problem, serving as a worked example.

## 2.2 Scaffolding in Tutoring Systems

Scaffolding in automatic tutoring systems is providing additional information or questions on how to make a step in solving a problem by breaking it into smaller steps [19]. It can be done by hints or questions. There is a number of studies concerning the efficacy of using the scaffolding [2,8,19,24], but their results are mixed. It can depend on the subject domain, learners, and the method of providing scaffolding. Belland et al. [2] in their meta-analysis state that scaffolding in general provides significant benefits when used in a problem-solving context which is important for teaching software engineering where problem-solving is often the chief learning activity.

Risha et al. [20] describe an automatic generation of a step-by-step problem-solving guide in the form of a trace table to a code snippet. A student is asked the result of the code snippet’s execution, and if their answer is incorrect, they can follow the code trace step-by-step. This lets the student quickly skip learning problems they can perform correctly and concentrate on those that cause errors. However, the level of scaffolding is low: it shows program execution, not the line of thought that is necessary to predict program execution. Many other tutors require solving learning problems at this level as a part of the regular learning process and do not consider this scaffolding.

One of the most important concerns is the domain independence of techniques for generating scaffolding hints. The work done by Risha et al. generalized to a single specific class of problems (Java Code Tracing which can be generalized to tracing program code in imperative programming languages); it is unclear if it can be used outside this domain. Using human-authored scaffolding hints or

tasks for each learning problem makes creating problem banks time-consuming and the scaffolding unsystematic. O'Rourke et al. [17] propose a generalized technique of scaffolding by formalizing the thought process that can be used to solve a given problem in the form of an algorithm that can be written similarly to a program code in a special domain-independent language called Thought Process Language (TPL). TPL code is executed by an interpreter which broadcasts events based on the executed statements (assignments, conditionals, loops and returns) that are displayed using domain-specific interface hooks. They considered three scaffolding modalities: tutorial progressions, just-in-time feedback, and fading worked-out examples. All these are based on displaying hints about solving the learning problem correctly.

Most of these scaffolding techniques focus on solving problems step-by-step and guiding the user through this process, showing what to do correctly. While often a useful practice, this can lead users to spend a lot of time either “debugging” their understanding of the problem (if the final answer is asked first), or entering all the smaller answers to the problem-solving steps (if the whole problem is broken down into steps). Reading hints or watching visualizations is a passive form of learning while scaffolding questions stimulate the active participation of the learner. Also, scaffolding questions can be used not just to show the correct way to solve the problem but also to help the learner understand what is wrong with their answer.

### 2.3 Question Generation

Writing a big amount of scaffolding questions manually is undesirable because of high labour costs, the complex structure of dependencies between scaffolding questions (see Fig. 11 in [22]), and the difficulty of keeping generated questions systematic. These problems can be solved by automatic question generation.

Myller proposed a method of automatic generation of prediction questions for Jeliot 3 [16], which is similar to O'Rourke's approach to explaining TPL code. Myller's approach relies on assigning static (as opposed to O'Rourke's dynamically constructed) questions to statements of Java code visualized using Jeliot. As the java code is executed, the system can check if a question for the current instruction is defined and, if found, display it to the user. The system creates a list of answers for the question consisting of a correct answer, which is the outcome of the line's execution, and several randomly chosen distractors. These questions concern solving the problem (predicting results of code execution) but not the thought process required to solve the problem.

Another question-generation approach is described in a paper by Zavala and Mendoza [23]. In this work, a question is an independent exercise (not a step within a larger learning problem) and is defined by question text (stem), a list of answer options if the question is multi-choice and a generator script. Both the stem and the answer options use embedded variables that act like placeholders for problem-specific values. The generator script is used to generate the values for embedded variables as well as the correct answer if the answer is textual. The use of a generator script allows for more variety when creating an exercise. While

this method allows the creation of several similar questions from a template, the question templates, including the generator scripts, defining kinds and structure of questions for each new learning situation must be defined by the author.

There is a lot of work on question generation from natural-language texts (e.g., [4]), but this approach is most suitable for verifying if the learner read the text. The number and precision of the generated questions are not enough to ensure that the learner mastered the technique, described in the text. Given the current state of question generation, the only viable way of generating scaffolding questions is using a description of the thought process required to solve the problem in some formal language (e.g., TPL [17]).

### 3 Generating Pedagogical Questions

We propose developing an automatic system that can maintain discussion with a learner who has problems while performing exercises in an automatic tutor by generating scaffolding questions depending on the learning situation and the learner's answer to previous questions.

#### 3.1 Thought Process Formalization

We based our approach to formalization of the problem-solving thought process on the model of “lines of thought” described by Kammenov et al. [7]. We define **Thought Process** as a decision tree that represents the line of thought required to make a step in solving a problem (**Hinting Tree**) or determine if the step was made correctly (**Verifying Tree**). In any current problem-solving situation, there is a specific part in the decision tree of the Thought Process that determines the outcome (the data for the next step for a Hinting Tree, or the correctness of the step and kinds of errors in a Verifying Tree). Let's call this path through the Thought Process' structure the *correct reasoning path*.

The reasoning steps that constitute the line of thought are therefore **nodes** of the **Thought Process** tree. The tree nodes correspond to reasoning steps proposed in [7], (i.e., **Questions**, **Actions**, **Branching** and **Iterations**); they can also describe **Outcomes** - conclusions made in particular branches.

When working with the **Branching** nodes, **Thought Branches** are the lower-level lines of thought that represent some sub-problems that have their own outcomes, impacting higher-level reasoning. In a Verifying Tree, the outcomes can be “correct” and “incorrect” and they can be aggregated using logical operators “and” and “or”. We call the **Branching** nodes **Logical Aggregations** if they aggregate different **Thought Branches** in that way, or **Independent Branching** if they link several lines of thought that can be checked in any order because only one of them can result in a conclusion in any given learning situation.

At the authoring stage both the elements of the subject-domain and the building blocks of the Thought Process (Branches, Nodes and Outcomes) are assigned natural-language strings that will be used as templates to automatically generate text of scaffolding questions and answers to them.

### 3.2 Kinds of Pedagogical Questions

We identified several kinds of pedagogical questions that can be generated automatically based on the Thought Process model. The questions are asked in a particular **learning situation**, including problem formulation, the current problem-solving step, variable values, and correct reasoning path according to the tree. The learning situation can describe a learner's error (Verifying Tree is used) or the situation when the learner does not understand what to do next (Hinting Tree is used). These questions are multiple choice questions with a single or multiple correct answers, so that each answer variant can lead to the next question (depending on the answer) or explanation (if the learner's problem was identified), shown to the learner.

- *Variable value* questions: “What is the value of <given variable>?”. A set of possible answers is based on what kind of objects from the learning situation the variable in question can hold. This kind of question allows to check if the learner performs actions correctly.
- *Final node* questions: “Which <reasoning step> determines the outcome in the current situation?”. These questions allow the system to quickly check what made the learner make a wrong move during problem-solving and to adapt the questioning strategy. These are single-choice questions whose sets of possible answers are formed from the nodes of the current branch of the decision tree.
- *Node* questions: asks the learner about the results in a particular node given the current learning situation. They are mostly used for question nodes because action nodes are covered by the variable value questions. The set of answers is defined by possible answers to question in the given node (labeled at its outgoing edges). These questions can identify reasoning errors when the learner asked the right question but gave a wrong answer to it because they did not know relevant facts of the subject domain. For independent branching nodes these questions ask for the results of the whole node, which can be followed by the questions about particular branches. For logical aggregation, node questions contains two parts: it asks for the aggregation results and for results for each branch. Iteration nodes produce three kinds of node questions: the question about aggregation results, the set of iterated objects, and individual results for each object.
- *Structural* questions: concern the edges of the decision tree “Given the <answer to a node question> what should you do next?”. These questions concern situations when the learner answered the question correctly but does not understand the consequences of this answer, i.e. the algorithm of reasoning. The set of answers is created from the decision-tree nodes that use variables known at this point.

### 3.3 Generating Question and Answer Formulations

As described above, all generated questions are linked to certain elements of the Thought-Process Tree—variables, nodes, and edges. So we can attach all the

1) Why do you think that the + operator at position 2 can be evaluated next?

- The left operand of the + operator at position 2 does not require further evaluation
- The right operand of the + operator at position 2 does not require further evaluation
- The internal operand of the + operator at position 2 does not require further evaluation

This is incorrect, since the right operand of the + operator at position 2 does require further evaluation.

Let's see why the right operand of the + operator at position 2 requires further evaluation.

2) Which operator can prevent the evaluation of the + operator at position 2 from the right?

- There is no such operator
- The + operator at position 8
- The \* operator at position 6
- The / operator at position 4

Correct.

3) Why do you think that the right operand of the + operator at position 2 does not require further evaluation?

- Due to the associativity of the + operator and the / operator
- Because the + operator at position 2 does not require a right operand
- Because of how parentheses affect the expression
- Because there are only evaluated operators to the right of the + operator at position 2
- Because of the precedences of the + operator and the / operator

Let's figure it out.

4) In the absence of parentheses' effect on the expression, how is the order of evaluation of adjacent operators determined?

- Left operator is always evaluated first
- The right operator is always evaluated first
- We need to compare the precedence of the + operator and the / operator
- We need to check what is the associativity of the + operator and the / operator

This is incorrect. In the absence of the effect of parentheses, the order of evaluation of adjacent operators is determined by their precedence – so the precedence should be compared next.

5) Compare the precedences of the + operator and the / operator

- The + operator has a higher precedence
- They have equal precedence
- The / operator has a higher precedence

Correct.

6) How is the order of evaluation of operators with different precedence determined?

- We need to check what is the associativity of the operator + and operator /
- Operator with higher precedence is always evaluated first
- Operator with lower precedence is always evaluated first

Correct.

So we've discussed that the / operator at position 4 should be evaluated before the + operator at position 2. So in this case, the + operator at position 2 cannot be evaluated next because the right operand of the + operator at position 2 requires further evaluation.

**Fig. 1.** An example of the question sequence for the “order of expression evaluation” problem

information required for generating natural-language parts of question formulations to the corresponding elements of the Thought-Process Tree. We use a template-based approach for constructing the question texts and sets of possible answers so that the questions can be adapted to any learning situation while remaining coherent and understandable.

Let us consider the Thought-Process Tree for determining the order of expression evaluation described in [7] to provide an example of generating pedagogical

questions. Figure 1 shows a sequence of questions that can be generated using that tree.

In this sequence, the first question is a Node question for an Aggregation node. The general template for this type of question is “Why do you think that `$description?`”, where `description` is a string parameter generated from the template stored in the node. For the node in this example, the `description` template is “`$X ${result} ? ‘can’ : ‘cannot’`” be evaluated next” where `X` is a string parameter describing the particular operator that the student selected (in this example, “the + operator at position 2”) and `result` is a boolean parameter which determines if the `description` is generated for the *true* or *false* result of the aggregation (in this example, *true*, as it is chosen to be equal to the student-selected answer to the node). Answer texts, as well as the explanations, are generated similarly from the templates attached to the Thought Branches that that node aggregates (e.g., “the left operand of `$X` does `${result} ? ‘not’ : ‘’` require further evaluation” and so on).

The correctness of answers to generated questions is determined using a software reasoner that verifies and grades students’ answers based on the formulae already present in the Thought-Process Tree. The templates’ parameters are either determined by the software reasoner or specifically chosen by the system to create the necessary texts.

With this scaffolding system of templates (i.e., templates on the question-type level, node level, and variable level), we can make the process flexible without asking the author of the Thought-Process Tree to provide too much information—for example, the tree for determining the order of expression evaluation, consisting of 24 nodes combined in 3 Thought Branches, requires 237 text templates, most of which are relatively simple sentences or phrases. This work should be done once per developed tutor so the workload can be considered small compared to the number of generated questions.

### 3.4 Generating Sequences of Questions

One of the most important and complex problems when using scaffolding pedagogical questions is choosing the questions to ask. When a human teacher uses this technique, they are limited by classroom time (pedagogical questions to help a student solve a particular problem are, by their nature, individual), and experienced teachers mostly rely on their knowledge about most common errors during these exercises. In an automatic tutor, the developers must be aware about the possibility of asking too many trivial questions which can decrease learners engagement. While using reinforcement learning to build an optimal strategy of asking pedagogical questions for a particular domain is a promising method, to perform it we need manual “cold start” strategies to gather data for machine learning. The starting point to generate a set of pedagogical question when a learner made error that they cannot understand is the correct reasoning path through the Verifying tree which leads to a particular conclusion about the cause of the error – the conclusion that the learner missed for some reason.

---

**Algorithm 1.** Strategy of asking pedagogical questions

---

- 1: Determine Deciding Variables: a set of variables which receive values in the correct reasoning path through the Thought Process and whose values are used at least once later in the correct reasoning path.
  - 2: **if** Deciding Variables isn't empty **then**
  - 3:     Perform a Faulty Variable Search on a list of Deciding Variables.
  - 4: **end if**
  - 5: Determine Final Nodes: final nodes are the nodes that have at least one outcome linking them to a conclusion node for the current Thought Branch. The correct final node is the final node that lies on the correct reasoning path — i.e., the reasoning step determining the final decision in this situation.
  - 6: **if** Final Nodes contain at least two nodes **then**
  - 7:     Ask a Final Node question with Final Nodes as the set of answers to determine the learner's chosen final node.
  - 8: **else**
  - 9:     Assume that the learner chose the correct final node.
  - 10: **end if**
  - 11: Determine the lowest common ancestor (LCA) of the learner's chosen final node and the correct final node.
  - 12: **if** LCA is not the learner's chosen final node **then**
  - 13:     Ask a Structural question about the outcome leading to the LCA.
  - 14: **end if**
  - 15: Start with the current node being LCA
  - 16: **while** No learner's error is found **do** ▷ Do a step-by-step questioning to fix the learner's understanding of the Thought process structure
  - 17:     Ask a Node question about the current node.
  - 18:     Ask a Structural question about the current node's correct outcome.
  - 19:     Move to the node following the current node's correct outcome.
  - 20: **end while**
- 

We can implement different strategies of generating pedagogical questions using the same thought-process model; their efficiency must be compared experimentally. We propose one possible algorithm (see Algorithm 1) of using pedagogical questions which tries to find balance between precision (asking the learner about everything necessary to solve the task to find the first error) and the number of questions asked (first nodes of the decision tree are often trivial). To understand this strategy of questioning, it is necessary to keep in mind the fact that many of the nodes in the decision tree use variables whose values were determined before, and these values are mentioned in the question text, so it is unfeasible to ask node and structural questions concerning these nodes until we are sure that the learner calculated the required variable correctly. So questions verifying variable values should, typically, come first. Also, questions about required nodes allow quick surveying the learner's line of thought and so can be used to detect reasoning errors quicker when used at the beginning.

In this strategy, Faulty Variable Search and checking the understanding of the structure of the Thought Process can be performed linearly (from the start of the correct reasoning path; this method asks a lot of questions but allows

precise finding of the first reasoning error in every situation) or using dichotomy (significantly fewer questions, but can give the wrong results if some of the learner's errors were corrected by their later errors).

This proposed strategy produces complex trees of possible pedagogical questions, even for relatively simple Thought Processes. For the tree for determining the order of expression evaluation that we used as an example, we generated 74 unique pedagogical questions; the number of unique possible question sequences depends on the given learning situation (in this case, the expression and student's answer) and can be on the order of thousands (for example, for the relatively simple expression  $(X - A)[B + C * D]$  when considering if the + operator can be evaluated, we generated 5 576 different question sequences depending on the student's answers) which shows a high degree of the method's adaptability and proves that manual development of systems of pedagogical questions is not viable. The question sequences tend to be relatively short, with 14 questions in a sequence at most and the average sequence length from 5 to 8 depending on the learning situation, so it is short enough not to overload the student.

The sequence of questions in Fig. 1 was generated using this strategy for the situation when the student had to determine the order of evaluation for the expression  $X + A/B * C + D/K$  and chose the operator + after X variable first. In this sequence, the first question is a Node question for the Logical Aggregation node, the second question is a Variable Value question, the third question is Final Node question, the fourth and sixth questions are Structural questions, and the fifth question is a Node question for the Question node.

## 4 Evaluation

To evaluate the proposed method, we generated pedagogical questions for determining the order of expression evaluation. We asked 5 experts who taught CS0 and CS1 courses for students majoring in Computer Science in a university. They were able to simulate student interaction with the system in different learning situations and were asked to fill a short survey on 5-point Likert scale (1 – strongly disagree; 5 – strongly agree). Table 1 shows the survey results.

As the table shows, teachers rated pedagogical questions favorably (most average values were equal or above 4 “agree” level) with the lowest value for simplicity of question formulations. In free-text interviews, they also raised this problem most often: automatically generated question formulations sometimes became too complex while human teachers, taking into account the nuances of a particular situation, could often say it simpler. Teachers also supposed changing some of the complex terms (e.g., “operator associativity” or “operators with the strict order of operands”) to relevant objects (particular operators) in the learning situations or their lists, but it contradicts the goal of teaching the relevant concepts. This problem was caused more by the description of the subject domain than the presented algorithm; it can be at least partially solved by working more on the natural-language formulation of the target domain.

Some of the experts also expressed concerns regarding some of distracting answers because they thought them potentially misleading, even though the

**Table 1.** Expert survey on the system of pedagogical questions.

Question	Avg	Std.dev.
The question formulations were grammatically correct	4.6	0.55
Questions and possible answers can be easily understood by a student	3.4	1.14
Question formulations are compact	4.4	0.89
The sequences of pedagogical questions are effective for determining the error causes	4	1.22
Answering sequences of pedagogical questions can help students learn the topic	4.2	1.3
Question sequences are optimal: they allow quick finding of the error cause	4	0.71
Question sequences are reasonable: a user can easily grasp why these questions are asked in the specified order	4	1.22

system explains why students are wrong if they choose a wrong option. This complaint mostly concerned the situations when the wrongness of the answer was too obvious for the teacher. This requires studying what is better: propose a potentially misleading answer and explain why it is wrong or never propose it.

## 5 Conclusion

Pedagogical questions let tutoring systems determine the exact cause of the error the student made and correct the relevant misconception. Unlike explanations and examples, questions make the student an active participant in fixing their errors; answering questions stimulates thinking and information retrieval. They also increase the students' trust because the system does not hand them information but asks for it. Practically, it is enough to detect and correct the first reasoning error the student made while performing a problem-solving step: the subsequent errors will be detected and fixed while solving further problems.

The disadvantage of pedagogical questions is the amount of time needed to answer them. Their automatic generation takes the load off the teacher, but not from the students. It is not practical to ask pedagogical questions for each error the student makes. Determining the conditions to ask a set of pedagogical questions effectively is a good direction for further research. Pedagogical questions are useful to fix persistent errors; tutoring system can keep track of the pedagogical questions the student answers correctly and avoid repeating them.

Our further research will concern integrating pedagogical question modules we developed with the CompPrehension intelligent tutoring system [22] and experimenting with students to study the efficiency of automatically generated pedagogical questions and improving the method of their generation.

We will also consider the applicability of this technique to other domains than teaching programming and other formal courses like mathematics. For example, we are currently developing Thought-Process Trees for teaching English natural language, covering the tasks of determine the order of adjectives and generally determining the order of words in a sentence.

## References

1. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: A new paradigm for intelligent tutoring systems: example-tracing tutors. *Int. J. Artif. Intell. Educ.* **19**(2), 105–154 (2009). <http://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai19-2-02>
2. Belland, B.R., Walker, A.E., Kim, N.J., Lefler, M.: Synthesizing results from empirical research on computer-based scaffolding in stem education: a meta-analysis. *Rev. Educ. Res.* **87**(2), 309–344 (2017). <https://doi.org/10.3102/0034654316670999>
3. Fernandes, E., Kumar, A.N.: A tutor on scope for the programming languages course. In: Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 2004), pp. 90–93. Association for Computing Machinery, New York, NY, USA (2004). <https://doi.org/10.1145/971300.971332>
4. Foucher, S., Pascual, D., Richter, O., Wattenhofer, R.: Word2course: creating interactive courses from as little as a keyword. In: Proceedings of the 14th International Conference on Computer Supported Education (CSEDU), vol. 1, pp. 105–115. INSTICC, SciTePress (2022). <https://doi.org/10.5220/0011064700003182>
5. Getseva, V., Kumar, A.N.: An empirical analysis of code-tracing concepts. In: Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education (ITiCSE 2022), vol. 1, pp. 262–268. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3502718.3524794>
6. Gréuisse, C., Rothkugel, S.: An SKOS-based vocabulary on the swift programming language. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 244–258. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-62466-8\\_16](https://doi.org/10.1007/978-3-030-62466-8_16)
7. Kamennov, Y., Sychev, O., Orlova, Y.: Covering possible reasoning errors for intelligent tutoring systems: order of expression evaluation case. In: Crossley, S., Popescu, E. (eds.) Intelligent Tutoring Systems (ITS 2022). LNCS, vol. 13284, pp. 65–74. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_6](https://doi.org/10.1007/978-3-031-09680-8_6)
8. Kim, N.J., Belland, B.R., Walker, A.E.: Effectiveness of computer-based scaffolding in the context of problem-based learning for stem education: Bayesian meta-analysis. *Educ. Psychol. Rev.* **30**(2), 397–429 (2017). <https://doi.org/10.1007/s10648-017-9419-1>
9. Kumar, A., Dancik, G.: A tutor for counter-controlled loop concepts and its evaluation. In: 33rd Annual Frontiers in Education (FIE 2003), vol. 1, pp. T3C–7 (2003). <https://doi.org/10.1109/FIE.2003.1263331>
10. Kumar, A.N.: The effectiveness of visualization for learning expression evaluation: a reproducibility study. In: Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2016), pp. 192–197. Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2899415.2899427>

11. Kumar, A.N.: An epistemic model-based tutor for imperative programming. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 213–218. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78270-2\\_38](https://doi.org/10.1007/978-3-030-78270-2_38)
12. Kumar, A.: A tutor for using dynamic memory in C++. In: 32nd Annual Frontiers in Education, vol. 1, pp. T4G–T4G (2002). <https://doi.org/10.1109/FIE.2002.1158013>
13. Laengrich, M., Schulze, J., Kumar, A.N.: Expression tasks for novice programmers: turning the attention to objectivity, reliability and validity. In: 2015 IEEE Frontiers in Education Conference (FIE), pp. 1–8 (2015). <https://doi.org/10.1109/FIE.2015.7344070>
14. Malmi, L., Karavirta, V., Korhonen, A., Nikander, J., Seppälä, O., Silvasti, P.: Visual algorithm simulation exercise system with automatic assessment: TRAKLA2. *Inform. Educ.* **3**(2), 267–288 (2004). <https://doi.org/10.15388/infedu.2004.19>
15. Moreno, A., Myller, N., Sutinen, E., Ben-Ari, M.: Visualizing programs with Jeliot 3. In: Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2004), pp. 373–376. Association for Computing Machinery, New York, NY, USA (2004). <https://doi.org/10.1145/989863.989928>
16. Myller, N.: Automatic generation of prediction questions during program visualization. *Electron. Notes Theor. Comput. Sci.* **178**, 43–49 (2007). <https://doi.org/10.1016/j.entcs.2007.01.034>. Proceedings of the Fourth Program Visualization Workshop (PVW 2006)
17. O'Rourke, E., Andersen, E., Gulwani, S., Popović, Z.: A framework for automatically generating interactive instructional scaffolding. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), pp. 1545–1554. Association for Computing Machinery, New York, NY, USA (2015). <https://doi.org/10.1145/2702123.2702580>
18. Rajala, T., Laakso, M.J., Kaila, E., Salakoski, T.: VILLE: a language-independent program visualization tool. In: Proceedings of the Seventh Baltic Sea Conference on Computing Education Research (Koli Calling 2007), vol. 88, pp. 151–159. Australian Computer Society Inc., AUS (2007)
19. Razzaq, L., Heffernan, N.T.: Scaffolding vs. hints in the assistance system. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 635–644. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_63](https://doi.org/10.1007/11774303_63)
20. Risha, Z., Barria-Pineda, J., Akhuseyinoglu, K., Brusilovsky, P.: Stepwise help and scaffolding for java code tracing problems with an interactive trace table. In: Proceedings of the 21st Koli Calling International Conference on Computing Education Research (Koli Calling 2021). Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3488042.3490508>
21. Sorva, J., Sirkia, T.: UUhistle: a software tool for visual program simulation. In: Proceedings of the 10th Koli Calling International Conference on Computing Education Research (Koli Calling 2010), pp. 49–54. Association for Computing Machinery, New York, NY, USA (2010). <https://doi.org/10.1145/1930464.1930471>
22. Sychev, O., Penskoy, N., Anikin, A., Denisov, M., Prokudin, A.: Improving comprehension: intelligent tutoring system explaining the domain rules when students break them. *Educ. Sci.* **11**(11), 719 (2021). <https://doi.org/10.3390/educsci11110719>

23. Zavala, L., Mendoza, B.: On the use of semantic-based AIG to automatically generate programming exercises. In: Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE 2018), pp. 14–19. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3159450.3159608>
24. Zhou, Y., Andres-Bray, J.M., Hutt, S., Ostrow, K., Baker, R.S.: A comparison of hints vs. scaffolding in a MOOC with adult learners. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 427–432. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78270-2\\_76](https://doi.org/10.1007/978-3-030-78270-2_76)



# Attention Assessment of Aircraft Pilots Using Eye Tracking

Maryam Ghaderi, Amin Bonyad Khalaj, Hamdi Ben Abdessalem<sup>(✉)</sup>,  
and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal,  
Montréal H3C 3J7, Canada

{maryam.ghaderi, amin.bonyad.khalaj,  
hamdi.ben.abdessalem}@umontreal.ca, frasson@iro.umontreal.ca

**Abstract.** Today, airplane transportation plays a significant role in the widespread movement of people and in the commercial activity of nations. Modern air travel can make remarkable progress by enhancing cultural and social services, growing economic activities, enhancing regional employment levels, and expanding tourism. The aviation industry has developed advanced planes and provided pilots with numerous control systems and instruments. In order to operate a modern high-speed aircraft, pilots must exchange instructions and information with air traffic controllers (ATC). In spite of many efforts and investigations in the field of aviation, tragic air accidents still occur worldwide. During the flight, pilots are subject to a variety of emotional, mental, and physical conditions, which affect their performance and attention. In this study, several pilots with diverse skills and experiences took part in different flights with different scenarios. Based on eye-tracking features, and as well as the amount of visual attention they paid to each window and tool within the simulation of an A320 aircraft, the amount of their visual attention is calculated.

**Keywords:** Visual Attention · Eye Tracking · Attention Assessment · Fixation Duration

## 1 Introduction

This study aims to use eye tracking to understand a person’s attention state in aviation by analyzing their eye movements and visual processes [1]. An attention process is a complex mental function that involves filtering out irrelevant information and focusing on a specific aspect of the environment. The task requires mental control and resistance to distractions. With training and practice, individuals can develop this skill, improving their performance and achieving their goals [2, 3].

High-level of attention is a crucial aspect of aviation [4], as it can significantly impact the safety and efficiency of flight operations, and it is necessary for communication, decision-making, and managing fatigue. To ensure that all operations are carried out safely and effectively, pilots, air traffic controllers, maintenance workers, and other

aviation professionals must maintain a high level of situational awareness and attention to detail. Pilots must pay close attention in the cockpit to monitor and change a variety of flight factors, such as airspeed, altitude, heading, and navigation. In order to maintain situational awareness and react to potential threats and hazards, and respond in emergency situations, they must also continuously monitor flight instruments and scan the surrounding environment.

Visual information is an essential aspect of the process of acquiring and transmitting information to the brain. The visual system is responsible for processing information from the environment through the eyes and sending it to the brain for further analysis and interpretation [5]. This allows the brain to prioritize the most important visual information and process it more efficiently, leading to better perception and understanding of the environment. In order to attend to visual information, the brain must first detect and recognize relevant visual stimuli. The photoreceptors—specialized cells that capture light and transmit signals to the brain via the optic nerve—start this process in the retina. After entering the brain, visual information is processed in a number of steps, requiring different levels of processing and interpretation [6].

Eye tracking technology involves detecting the movements and gaze of a person's eyes to determine what they are looking at or paying attention to. In the context of flight, eye-tracking technology can be used in several ways, such as piloting, flight simulation, and cabin management. Eye tracking can be used to control the movement of aircraft or to monitor the visual attention of pilots to detect signs of stress or fatigue. Furthermore, eye tracking can be utilized to control pilot attention and comfort levels to improve cabin design and services. In addition, eye tracking can be used in flight simulators to track a pilot's gaze and determine where they are directing their attention. For example, acquired information from eye tracking can be used to modify and control various elements in the simulation, such as bringing up information relevant to the object the pilot is looking at or changing the view in the cockpit. Finally, the aim of using eye tracking in flight simulators is to increase realism and give pilots a more immersive training environment.

To measure a pilot's attention during a flight, eye tracking devices must first follow the pilot's gaze and locate where they are looking. These data can then be used to calculate how much time the pilot spent looking at various components in the cockpit, such as instruments, controls, or the aircraft's exterior. In addition, the eye tracking devices may also measure other parameters such as pupil dilation and position, fixation, saccade, and blink rate which may provide additional insight into the level of attention and engagement of the pilot. Eye tracking can provide useful insights into a pilot's attention for training and performance evaluation. Generally, eye tracking technology could improve comfort, safety, and effectiveness in a variety of flight operations applications.

A brief description of the related works is presented in Sect. 2. Section 3 focuses on the experimental process, including the structure and steps of the experiments. The relationship between attention and eye tracking is explained in Sect. 4. In Sect. 5, we present an analysis of the experiment, and in Sect. 6, we discuss the results of the experiment. The conclusion and future directions of research are presented in Sect. 7.

## 2 Related Work

Eye tracking is an experimental method of detecting and directing eye movement, gaze points, fixation position, and fixation duration. It is a typical method for observing the allocation of visual attention. Thorough primary information on eye tracking research, including history and details, is provided in [7]. Eye tracking was first studied by Charles Bell, who explained the brain's control over eye movements and their effect on visual orientation. Fitts and coworkers conducted the first eye-tracking studies in a flying simulator in 1949 and 1950 [8, 9].

Previous studies have explored human behavior using eye tracking, including stress [10], problem-solving [11], and visual attention [12], as well as in virtual reality. Eye tracking has been used in VR to study gender classification and different types of visual attention, such as on peer-learners, instructors, and screens with various manipulations. Eye tracking analyzed children's learning in classrooms using object-of-interests and different manipulations, such as hand-raising percentages, visualization styles, and sitting positions in immersive VR [13]. Furthermore, Eye movements have been used to predict psychological behaviors such as personality traits, cognitive load, and IQ test performance using machine learning methods [14].

Paper [4] discusses how pilots can manage attention and workload during high or low workload periods. They recommend strategies such as recognizing unexpected events and prioritizing tasks. Another paper [15] introduces the attentional task model, which categorizes all flight tasks according to their attention scores and establishes a baseline level of workload and attention requirements.

By identifying regions of high visual saliency using eye movements, researchers can determine where users focus their attention within the scene, and use this information to develop more effective visual displays. Eye movement indicators, such as fixation count, average fixation duration, visit count, and first fixation duration, were used in one study [6] to create more engaging and user-friendly 3D scene maps, ultimately improving user experience and map quality [6].

## 3 Experiments

An experiment was conducted aiming to gather pilot CW<sup>1</sup> and PD<sup>2</sup> during a takeoff procedure in an Airbus A320 in real time. It consisted of six different scenarios with varying weather, time, and conditions as well as the possibility of a failure. The first to third scenarios relate to standard takeoff sessions, while scenarios four to six relate to failure sessions. Different scenarios were given to participants to perform the takeoff procedure. A pilot monitor and a pilot are required to operate the Airbus A320. In this experiment, the participants served as pilots and the experimenter as pilot monitors. The experiment was conducted at the University of Montreal and with participants from CAE and Bombardier who all signed written consent forms following an ethics certificate.

---

<sup>1</sup> Cognitive Workload.

<sup>2</sup> Pupil Dilation.

### 3.1 Participants

Participants from CAE and Bombardier participated in the experiment at the University of Montreal. The study included 13 male subjects (all between 24 and 49 years of age) with an average age of 36 years. The 13 participants included 7 pilots with piloting experience, piloting license and A320 piloting experience. The other 6 participants were engineers at Bombardier and CAE who were familiar with most aircraft procedures but without holding a piloting license.

### 3.2 Materials

**Eye-Tracking Module:** Gaze point is the most affordable, research-grade eye tracker on the market, used in the experiment to measure pupil dilation.

Visual attention can be investigated using the eye-tracking technique. Our attention tends to be drawn to objects that are relevant to the task at hand or objects that have attracted our attention [16]. Seeing only what we're paying attention to is the result of the cognitive system allocating sufficient resources for visual processing. A significant part of visual attention is eye movement, which consists of fixations (stationary phase) and saccades (rapid, ballistic phase) [16]. As well as measuring eye movements and gaze point location, eye tracking is an effective tool for understanding human attention [16, 17].

### 3.3 Procedure

The experiment environment was comprised of a participant (for each experiment) as the pilot and the experimenter as the pilot monitor. Ethics committee and partners' approval were obtained for the experiment's procedure. Participants received a detailed description of the A320 takeoff procedure a week before their experiment to familiarize themselves with its handling characteristics. There are several different visual displays mentioned in the document (for example, navigation display, primary flight display, engine/warning display, system display, and engine display). It also included a rejected takeoff, an engine failure after V1 procedure and a standard takeoff procedure. In order to generate a more cognitive workload for every scenario, participants were not aware of the scenarios beforehand.

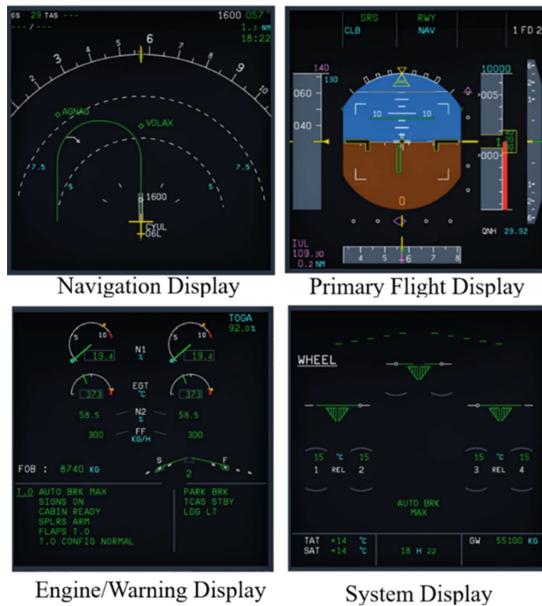
## 4 Attention and Eye Tracking

Piloting a plane requires a visual attention for all the different tasks of the flight. In order to obtain visual information, the eyes are the primary means, and eye movements reflect cognitive tasks and visual thinking processes. As a result, eye-tracking techniques can be used to gather eye movement data and analyze the distribution pattern of visual attention among observers. Based on the studies, area of pixels (pixel as the smallest unit), can be used to divide the target screen into areas of interest (AOI), as shown in Fig. 1. The whole display window that participants see in the A320 simulator during the experiment is shown in Fig. 1. Based on the experiment we did, there are five important areas

and dashboards named “Navigation Display (ND)”, “Primary Flight Display (PFD)”, “Engine/Warning Display (E/WD)”, “System Display (SD)”, and the “Runway”, which were used in the experiment Fig. 2.



**Fig. 1.** Areas of interest for takeoff



**Fig. 2.** Flight instruments for takeoff

Visual factors and visual saliency can be studied using eye-tracking technology by researchers and practitioners. When referring to visual saliency, we refer to the extent to which an object or feature in a visual scene attracts attention or catches the eye.

In eye tracking, a person's gaze is tracked as he or she looks at a visual display, such as a computer screen or a poster. Eye tracking records where and how long people look in a scene, giving insight into what they find salient. As a result of this information, it is possible to understand how visual factors in a scene are related to object saliency. Thus, an eye-tracking system can link visual factors and visual saliency. The main eye-tracking indexes which can be used to determine the relationship between the eye tracking index and visual attention [6], and their specific meaning are presented based on the gaze point documentation:

- **Fixation POG:** By using the internal fixation filter, POG (Point of Gaze) data determines the user's point-of-gaze, expressed as FPOGX and FPOGY. In other words, Fixation Point of Gaze (POG) refers to the specific location where a person's eyes are focused on when they are observing a visual display.
- **FPOGS:** POG start times which are expressed in seconds since system initialization. Using the FPOGS in eye tracking can provide insight into the initial focus of attention, which can be used to identify the most salient or noticeable features.
- **FPOGD:** The fixation POG duration in seconds.
- **FPOGID:** Number associated with the fixation POG.

As mentioned earlier, pixels (pixel as the smallest unit), can be used to divide the target screen into areas of interest (AOI). We divided the screen into 5 areas of interest based on area of pixels.

point in a coordinate is named by its ordered pair of the form of  $(x, y)$ . The first number corresponds to the x-coordinates and the second to the y-coordinate. For screen coordinates, the origin is the upper-left corner of the screen. Usually, the full position of a window is represented as a rectangle with two points defining its upper-left and lower-right corners.

Each window coordinate, which consists of four numbers separated by commas or spaces, defines the boundaries of each area of interest. Two points on the screen are represented by these four numbers. "X" and "Y" coordinates of the first point are indicated by the first two numbers, while "X" and "Y" coordinates of the second point are indicated by the final two numbers. The screen's upper-left corner is marked by "0,0," while its lower-right corner is marked by "1,1." The screen's centre is marked by [0.5,0.5]. By using these two points, we can define the position and size of a window. Specifically, the first point defines the upper-left corner of the window, while the second point defines the lower-right corner. It is possible to define a window's position and size using these two points. The boundaries of each area of interest are as follows:

- Visual Display 1: [0, 0.66] and [0.18, 1]
- Visual Display 2: [0.19, 0.66] and [0.38, 1]
- Visual Display 3: [0.66, 0.66] and [0.84, 1]
- Visual Display 4: [0.85, 0.66] and [1, 1]
- Visual Display 5: Runway
- Visual Display 6: All other coordinates.

Thus, the **Fixation POG** has been used to find the visual display that pilots was looking at each portion of the second. The **FPOGD** has been used a sign indicating a change in the point of view. According to [18], the duration of the fixation POG can be used as the only effective factor, which shows the distribution of visual attention during each second of the flight.

$$S = \frac{FPOGD}{\sum FPOGD} \quad (1)$$

This formula considers the portion of fixation time. The proportion of fixation time on a specified AOI, can be used to measure the distribution of visual attention during the flight phase (or any phase segment). The idea of this formula is to divide the duration of the fixation POG during the occurrence of each event by the total time of that event. By doing this, visual attention will be independent of the total duration of the event. In this way, we can see what portion of time the pilot (participant) has looked at each display window. If only one display window was observed for the entire duration of the event, the S will be equal to one (or 100% in scale of 0 to 100), which means that the pilot paid full attention to that display window during the event. Hence, the gaze duration (sum of fixation durations on a specific AOI) was used as the basis for evaluating visual attention distribution. However, we focus only on the distribution of attention during each second in this study. The goal is to find out which display window the participant is looking at each second of the flight.

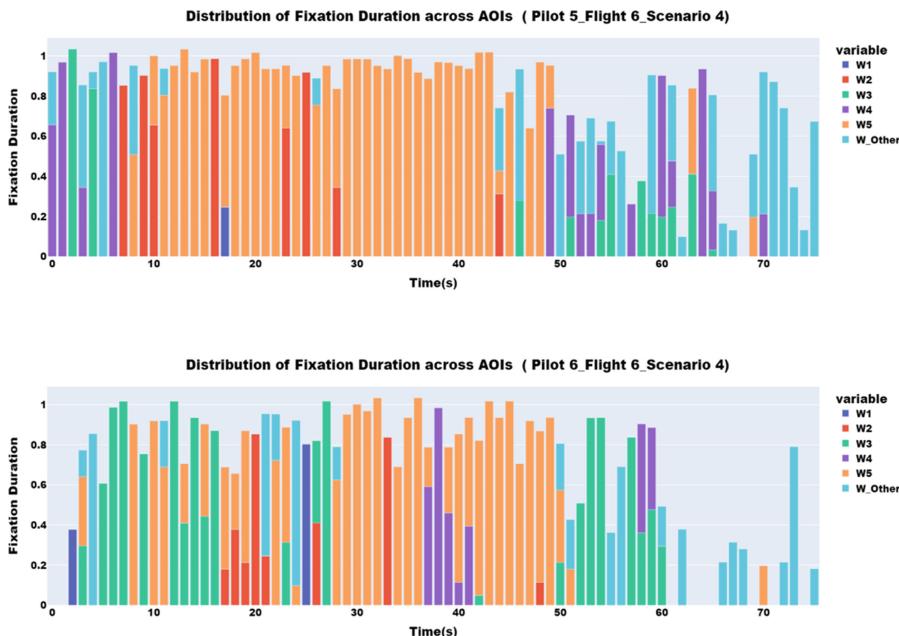
## 5 Analysis

A participant may gaze at one or more visual displays per second, as explained in the previous sections. Knowing what visual displays the participant was looking at, at every second of a flight is crucial to determining his correct and proper performance. Each period of the flight includes special tasks, so the participant is expected to pay attention to special visual displays at those times. Consequently, comparing the areas that the participant looked at every second with the areas that he was supposed to look at gives us valuable information about the participant's ability to control the flight and his level of performance.

Two participants' data output examples are shown here. As shown in Fig. 3, the first one exhibits the data output of an official A320 pilot with over 250 formal flight hours and eight years of experience during a rejected takeoff (sixth pilot). And the second one shows the output of a non-licensed pilot for the same scenario. This pilot has flown in the simulator with the A320 for testing purposes but is not a licensed pilot. He is an A320 engineer who knows the ins and outs of the aircraft (fifth pilot). In Figs. 3 and 4, we can see which screen participants 5 and 6 were watching throughout the flight.

Figure 3 gives us the information about participant 6 and 5, flight 6 and scenario 4. As you can see Participants' fractions of a second spent on each visual display are displayed in different colors. For example, a participant's attention was diverted to three areas one, two, and three in the 2<sup>th</sup> second, but in the 10<sup>th</sup> second, the participant focused entirely on the runway. It should be noted that, for different windows, the sum of fractions of seconds is the maximum one since our measurement standard is one second.

During a flight, the visual displays a participant looks at can provide valuable information about their ability to control the flight. Checking whether they looked at the appropriate displays can be determined by comparing where they looked with where they were supposed to look. Pilots with licenses (participant 6) demonstrated better focus during a rejected takeoff scenario than non-licensed engineers with no licenses (participant 5). Thus, monitoring a pilot's visual attention during critical scenarios can enhance pilot training and flight safety.



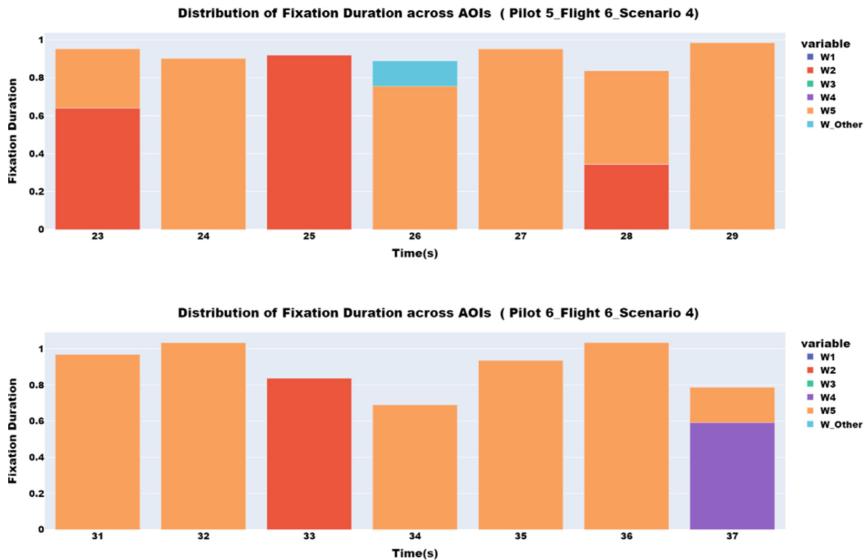
**Fig. 3.** Distribution of fixation duration across AOIs Related to Pilot 6 and 5

## 6 Discussion and Results

As shown in the figure below (Fig. 4), the duration of the fixation POG at each Area of Interest is shown in seconds, 3 s before and after the engine failure. The figure on the right represents the 6th participant, and the figure on the left represents the 5th participant. In the 3 s leading up to the engine failure, the 6th participant looked at the primary flight display and the most at the runway and had everything under control. The pilot's attention was only focused on the front window (Runway) once the engine failed, so he could keep the plane on its course by controlling the pedals.

However, the figure on the left shows that before the engine failure, the 5th participant mostly looked at the primary flight display and the front window (Runway), failing to pay attention enough to the other runway, which is very critical at the time of failure. Additionally, he was not only focused on the runway at the time of the engine failure

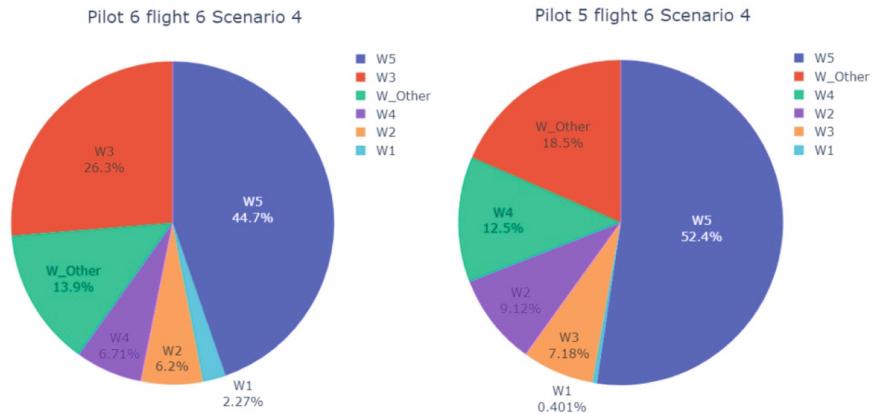
but was also looking around. He randomly looked at visual display 2 (primary flight display), the runway and around, indicating a lack of control of the aircraft and the fact that he does not know where to look at.



**Fig. 4.** Distribution of fixation duration across AOIs 3 s before and after engine failure (Related to Pilot 6 and 5)

Based on the pie chart (Fig. 5), we can see how pilots' attention is divided between different visual displays (AOIs) throughout the whole flight. Visual displays are represented by slices of the circle. Based on the size of each slice, we can determine how much attention the pilot paid to each visual display during the flight. The figure on the top represents the 6th participant, and the figure on the down represents the 5th participant. As you can see, both participants looked at the runway for almost half of the flight (this is not unexpected since the pilot's primary goal is to keep the plane on course, so spending so much time in the front window (runway) makes sense). However, it is the amount of time participants spend looking at different dashboards that distinguishes the experienced pilot participant from the non-pilot participant. As it is shown, participant 6, who was an expert pilot, spent more than a quarter of the flight looking at dashboard 3 (Engine/Warning Display), which is quite reasonable given the engine failure and shows he has control of the flight. He looked at the dashboard 2 (Primary Flight Display) and 4 (System Display) for about 6% of the total flight time (6.2% and 6.71% respectively).

However, despite the importance of dashboard 3 (Engine/Warning Display), only 7% of the 5th pilot's attention was focused on this dashboard. A greater amount of attention was paid to dashboard 2 (Primary Flight Display) and dashboard 4 (System Display). It shows he had no control over the flight, as he looked at different dashboards at random and was not aware of where to look at all times.



**Fig. 5.** Pilot attention distribution across areas of interest (AOIs) during flight.

## 7 Conclusion

In conclusion, we measured the fixation duration during takeoff to investigate the attention distribution of expert and novice pilots. Attention can be measured by fixation duration, which measures the time spent looking at a particular location. There is generally a correlation between longer fixation durations and higher levels of attention, as the longer someone looks at a particular location, the more information they are likely to absorb. In another word, the fixation duration can thus be used to measure the amount of attention paid to different objects or locations. Locations or objects with longer fixation durations are likely to have received more attention. There was a clear difference in fixation duration between the two groups, with the expert pilot maintaining constant attention on the relevant display window, while the novice pilot was less focused.

In this study, fixation duration was demonstrated to be a useful measure of attention allocation, and it was revealed that expert pilots have a more developed and efficient system for allocating attention. By utilizing this information, expert pilots can be differentiated from novice pilots, and training programs can be designed to improve aviation safety.

Overall, our study brings important novel insights into the relationship between attention and performance in complex tasks and stresses the importance of attention during critical phases of flight. The results of this study point to new avenues for future research in this area, aimed at increasing the safety of air travel.

**Acknowledgments.** We acknowledge the support from NSERC-Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

## References

1. Glaholt, M.G.: Eye Tracking in the Cockpit: A Review of the Relationships Between Eye Movements and the Aviators Cognitive State (2014)

2. Bigelow, S.J.: Senior Technology Editor in the Data Center and Virtualization Media Group at TechTarget Inc. <https://www.skybrary.aero/>
3. Wickens, C.D., McCarley, J.S.: Attention-Situation Awareness (A-SA) Model of Pilot Error (2007)
4. Ghaderi, M., Ben Abdessalem, H., Frasson, C.: An analysis of mental workload involved in piloting tasks. In: Krouská, A., Troussas, C., Caro, J. (eds.) Novel & Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022). Lecture Notes in Networks and Systems, vol. 556. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-17601-2\\_21](https://doi.org/10.1007/978-3-031-17601-2_21)
5. Bull, D.R.: Chapter 2—The Human Visual System. Elsevier Ltd., Amsterdam, The Netherlands (2014)
6. Yang, B., Li, H.: A visual attention model based on eye tracking in 3D scene maps. ISPRS Int. J. Geo-Inf. **10**(10), 664 (2021). <https://doi.org/10.3390/ijgi10100664>
7. Wade, N., Tatler, B.W.: The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research. Oxford University Press, USA (2005). <https://doi.org/10.1093/acprof:oso/9780198566175.001.0001>
8. Charles, B.: On the motions of the eye, in illustration of the uses of the muscles and nerves of the orbit. Philos. Trans. R. Soc. Lond. **113**, 166–186. <https://doi.org/10.1098/rstl.1823.0017>
9. Fitts, P.M., Jones, R.E., Milton, J.L.: Eye movements of aircraft pilots during instrument-landing approaches. Aeronaut. Eng. Rev. **9**, 24–29 (1950)
10. Hirt, C., Eckard, M., Kunz, A.: Stress generation and non-intrusive measurement in virtual environments using eye tracking. J. Ambient. Intell. Humaniz. Comput. **11**(12), 5977–5989 (2020). <https://doi.org/10.1007/s12652-020-01845-y>
11. Eivazi, S., Bednarik, R.: Predicting problem-solving behavior and performance levels from visual attention data. In: Proceedings of 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI, pp. 9–16. ACM, New York (2011)
12. Bozkir, E., Geisler, D., Kasneci, E.: Assessment of driver attention during a safety critical situation in VR to generate VR-based training. In: ACM Symposium on Applied Perception 2019. ACM, New York (2019). <https://doi.org/10.1145/3343036.3343138>
13. Bozkir, E., et al.: Exploiting object-of-interest information to understand attention in VR classrooms. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), Lisboa, Portugal, pp. 597–605 (2021). <https://doi.org/10.1109/VR50410.2021.00085>
14. Kasneci, E., Kasneci, G., Trautwein, U., et al.: Do your eye movements reveal your performance on an IQ test? A study linking eye movements and socio-demographic information to fluid intelligence. PLoS ONE **17**(3), 1–23 (2022). <https://doi.org/10.1371/journal.pone.0264316>
15. Ghaderi, M., Courtemanche, M.A., Ben Abdessalem, H., Nkambou, R., Frasson, C.: Attentional tasks model: a focus group approach. In: Krouská, A., Troussas, C., Caro, J. (eds.) Novel & Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022). NiDS 2022. Lecture Notes in Networks and Systems, vol. 556. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-17601-2\\_29](https://doi.org/10.1007/978-3-031-17601-2_29)
16. Abdelrahman, Y., et al.: Classifying attention types with thermal imaging and eye tracking. In: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 3, Article 69, 27 p. (2019). <https://doi.org/10.1145/3351227>
17. Mancas, M., Ferrera, V.P.P., Riche, N., Taylor, J.G.G. (eds.): Springer, New York (2016)
18. Jin, H., et al.: Study on how expert and novice pilots can distribute their visual attention to improve flight performance. IEEE Access **9**, 44757–44769 (2021). <https://doi.org/10.1109/ACCESS.2021.3066880>



# KEMMRL: Knowledge Extraction Model for Morphologically Rich Languages

Daniel Vasić<sup>1,3</sup>(✉) , Branko Žitko<sup>2,3</sup> , Ani Grubišić<sup>2,3</sup> , and Angelina Gašpar<sup>2,3</sup>

<sup>1</sup> Faculty of Science and Education, University of Mostar, Matice hrvatske b.b.,  
Mostar 88000, Bosnia and Herzegovina

*daniel.vasic@fpmoz.sum.ba*

<sup>2</sup> Faculty of Science, University of Split, Rudera Boškovića 33 Split,  
21000 Split, Croatia

*{bzitko,ani}@pmfst.hr*

<sup>3</sup> Catholic Faculty of Theology, University of Split,  
Zrinsko-Frankopanska 19, 21000 Split, Croatia

*agaspar@kbf-st.hr*

**Abstract.** There is a growing interest in automatic text processing and knowledge extraction from text repositories which often requires building new language resources and technologies. We present the KEMMRL model designed for the under-resourced but morphologically rich Croatian language. The proposed model uses natural language processing techniques, state-of-the-art deep learning algorithms and a rule-based approach to generate knowledge representations. The output of the newly developed HRtagger and HRparser methods in combination with the KEMMRL model is knowledge represented in the form of an ordered recursive hypergraph. Since the performance of KEMMRL is highly dependent on the applied deep learning methods, we evaluated them using hr500k reference corpus in the training and testing phase and manually designed out-of-domain Semantic Hypergraph Corpus (SemCro). The results of standard evaluation metrics showed that the HRtagger and HRparser achieved significantly better results than other state-of-the-art methods. These methods also showed the best results in measuring the structural similarity of hypergraphs, the highest average similarity to the manually annotated semantic hypergraphs and the number of semantic hyperedges correctly annotated by the model. The semantic hypergraph proved to be an ideal structure to capture and represent knowledge from more complex sentences without information loss. Researchers and developers of similar morphologically rich languages can customize and extend KEMMRL to their requirements. This article highlights the potential benefits of implementing the KEMMRL model into an Intelligent Tutoring System (ITS), and future research may focus on developing and testing such implementations.

**Keywords:** Knowledge extraction · natural language processing · deep learning techniques · morphologically rich languages · knowledge representation · semantic hypergraph

## 1 Introduction

People exchange or gain knowledge using languages, images, symbols, and other communication tools. Efficient communication or knowledge exchange depends on one's ability to convey information clearly and accurately so that its receiver can easily understand or decode it. Similarly, machines interpret and process natural language only if it is machine-readable. The automatic conversion from unstructured data into structured one is known as knowledge extraction, which can be arduous due to polysemy, free word order, complex morphology, etc. Artificial Intelligence researchers have developed various natural language processing (NLP) techniques to allow machines to read, understand and derive meaning from a human language.

Natural language processing (NLP) is a subfield of computer science, linguistics, and artificial intelligence that studies the interaction between computers and natural languages. NLP tasks include natural language understanding (NLU, transforming human language into a machine-readable format) and natural language generation (NLG, converting information from computer representations into human-readable text). NLP techniques can be applied to several modules of an Intelligent Tutoring System, such as knowledge representation, user modelling, feedback and assessment, and dialogue management. This can enable the system to provide a more personalized and engaging learning experience by better understanding learners' needs and interactions.

The KEMMRL converts natural language text into a semantic hypergraph using standard deep learning techniques in the natural language processing phase and specially defined rules in the knowledge extraction phase. We created knowledge extraction rules that convert syntactic annotations into specific labels to be embedded in the semantic hypergraph [18]. Compared to other methods for semantic parsing in [8], KEMMRL allows different natural language processing (NLP) and natural language generation (NLG) tasks. It provides mechanisms for natural language inference, co-reference resolution, topic modelling, etc. The advantage of using a semantic hypergraph to represent knowledge is that it allows the reconstruction of a sentence in its original form and the creation of new concepts from existing ones. However, its main disadvantage is its inability to reduce vagueness in automatic knowledge inference.

The motivation behind this research is to apply semantic hypergraph in an Intelligent Tutoring System (ITS) [22] more precisely in its communication module designed for question-answer pair generation. The hypergraph can help the system identify relevant concepts and associations that may be overlooked by traditional keyword-based approaches and other similar approaches, resulting in more accurate and comprehensive answers to learner queries.

Our main contributions are as follows:

- We designed a model for knowledge extraction and semantic hypergraph-based knowledge representation.
- The KEMMRL flexibility and extensibility allow researchers of similar languages to adapt the model to their requirements.

- Manually designed Semantic Hypergraph Corpus (SemCro) is freely available under the CC-BY SA license at the CLARIN.SI repository of the Slovene National Consortium of the European Research Infrastructure (CLARIN).

The following section provides an overview of the related work.

## 2 Related Work

Traditional NLP methods require handcrafted features and their manual design is a lengthy process. Deep learning, aiming at new applications(e.g. summarization, intelligent tutoring, etc.) can acquire new effective feature representation from training data. State-of-the-art NLP methods for morphologically rich languages are based on deep learning and show good results for Croatian language [16].

### 2.1 Methods for POS Tagging

The emergence of recurrent neural networks has enabled the development of methods for POS tagging without manually defining the features but using a gate mechanism (Long-Short Term Memory LSTM) and two-way connections (Bidirectional LSTM) between neural networks layer. Even though it can be a problem for under-resourced languages that lack annotated corpora, transfer learning is the solution to overcome it. We propose a similar model that uses MaxOut on the output of the transformer and single-layer neural network with Softmax activation for the output predictions, The advancement in modelling attention [24] in neural networks allowed the model to focus only on relevant parts of the input. These unsupervised learning models are state-of-the-art for multiple NLP tasks. In [6], the authors propose a model for POS-tagging based on the open-source T'n'T implementation called HunPos [13]. An improvement of this model presented in [15] refers to the optimization of the algorithm and the training corpus. Instead of the HunPos implementation, an algorithm based on Conditional Random Fields (CRF) was used. A specific set of features used as input to the CRF algorithm contributed to the improvement of POS tagging for the Croatian language [15].

### 2.2 Methods for Dependency Parsing

Most modern parsers are statistical and data-driven, relying on manually annotated corpora. Typically, data-driven parsers do not require formal grammar but use statistics instead. Data-driven dependency parsing solutions for under-resourced languages include the approaches that use annotation projection [3] and/or transfer learning [16, 23]. The first Croatian SETimes.HR corpus [4] was developed for data-driven parsing and other NLP tasks. In [2], the authors propose a graph-based dependency parsing approach with predefined rules for parsing non-projective trees. This approach achieved the unlabeled attachment score

(UAS) of 73.4% and labelled attachment score (LAS) of 69.7% for the SETimes. They evaluated it using an hr500k corpus. In [5], the authors obtained better results, the UAS of 85.3% and the LAS of 79.4% employing graph-based dependency parsing and assigning the universal dependency labels for dependency trees.

### 2.3 Methods for Knowledge Extraction and Semantic Hypergraph-Based Knowledge Representation

Recent research on big data [11] shows that they come in various forms (text, XML, emails, images) or mostly (80%) in textual form. The manual extraction of these textual knowledge repositories requires effort and time, so there is a growing interest in automatic knowledge extraction. Knowledge extraction methods commonly identify triplets, concepts, and dependency relations in the text.

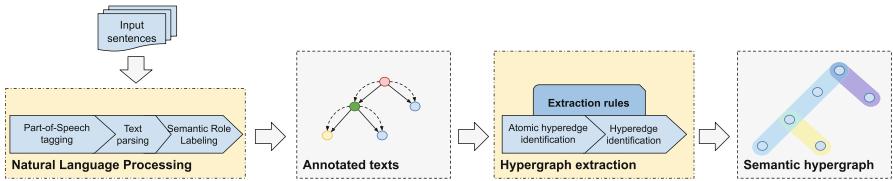
Knowledge representation formalisms are semantic networks, conceptual graphs, existential graphs, predicate logic, etc. Methods presented in [7] for concept extraction from the text and determining their interrelationships rely on rules and semantic knowledge bases. Knowledge extraction can refer to a task of learning, grouping, or classification. The automatic extraction of semantic networks from text implies the extraction of triplets and the merging of these triplets into a unique graphic representation of knowledge. The main disadvantage of these methods is their inability to present all the knowledge contained in the text because they are focused on the extraction of triplets. Although triplets represent most knowledge in the text the representation of more complex sentences in this form would result in information loss because of an incomplete and incorrect knowledge representation. The knowledge representation technique that can meet these prerequisites is a semantic hypergraph [19]. The semantic hypergraph can entirely capture the richness of knowledge contained in natural language text.

## 3 The KEMMRL Architecture and Core Functions

This section presents a model for knowledge extraction from a text in the Croatian language. The proposed model uses natural language processing techniques, state-of-the-art deep learning algorithms and a rule-based approach to generate knowledge representations. The knowledge extraction model is rule-based and the output of HRtagger and HRparser is knowledge represented in the form of a semantic hypergraph.

### 3.1 Knowledge Extraction

Knowledge extraction from the text in a morphologically rich language such as Croatian is a demanding task. In highly inflected language, one word can have multiple forms that convey the same meaning. The proposed model for knowledge extraction and hypergraphic representation is shown in Fig. 1.



**Fig. 1.** Model for knowledge extraction and semantic hypergraph-based knowledge representation

The proposed model for knowledge extraction and semantic hypergraph-based knowledge representation consists of three components.

- The text processing component uses Croatian text as input and produces syntactic and semantic annotations as output.
- The hypergraph extraction component consists of a hypergraph extraction model that generates a semantic hypergraph as the output. It uses a rule-based approach to extract the hypergraph in two phases.
- The inference component is an inference mechanism that uses retrieval and inference rules. It relies on a structured source such as a dictionary to describe the properties of the predicates over which it makes inferences.

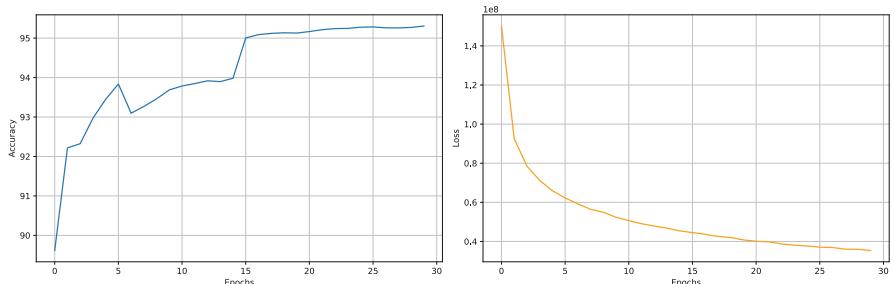
The proposed model for hyperedge extraction and knowledge representation relays on syntactic annotation methods defined by the MULTTEXT-East [12] Morphosyntactic Specifications and the Universal Dependencies [17]. The KEMMRL model supports languages that comply with these two specifications. The KEMMRL is a rule-based model independent of the tagging and parsing methods but dependent on the tagging and parsing specifications. Thus, it is easily extendable to many morphologically rich languages, including Romanian, Macedonian, Torlak and other languages in line with the MULTTEXT-East specifications. The higher accuracy of the tagging methods, the better quality of knowledge extraction and representation.

We used POS tags and dependency tree annotations to create an algorithm for hypergraph building in two phases. In the first phase, the vertices of hyperedges were formed and semantically labelled. In the second phase, the model built hyperedges using semantic labels and dependency tree labels. We used a manually annotated hr500k corpus [14] as a reference training and evaluation corpus. The deep learning model for POS tagging and dependency parser uses transformer-based architecture. In the feature extraction phase it uses CSEBert model [23]. The model proposed in this article is modular and can be easily upgraded by using different transformer-based architectures in the text processing stages.

### 3.2 HRtagger

As a sub-component of the text processing component, HRtagger is a deep learning-based method that uses transfer learning to achieve state-of-the-art results. It uses Bidirectional Encoder Representation from Transformers (BERT) [9] in the training phase and the CSEBert model to initialize the neural network with the pre-trained weights for feature extraction. For parameter optimization in the neural network, the HRtagger uses the Adam optimization algorithm instead of the SGD [21] optimizer with an adaptive learning rate method [10] and L2 regularization. The default value of 0.9 is for beta1 and 0.999 for beta2. In initial training, the neural network uses the first 10 warm-up steps. The reason for this neural network training is that the pre-trained data of the model differ from the training data. If the model immediately adjusts the weight values to a fixed learning rate, it would likely overfit the learning data and give no optimal results.

The neural network is trained with an adjustable number of steps with a maximum of 2000 steps. After 25 iterations, there is no improvement in parameter optimization and neural network training stops. Figure 2 shows the training process, the accuracy of HRtagger, and the value of the loss function of the evaluation data set. The model gradually adjusts the learning rate, increasing it over the warm-up period to avoid overfitting. At this stage, the learned weight values of the CSEBert model slowly fit sequence classification.



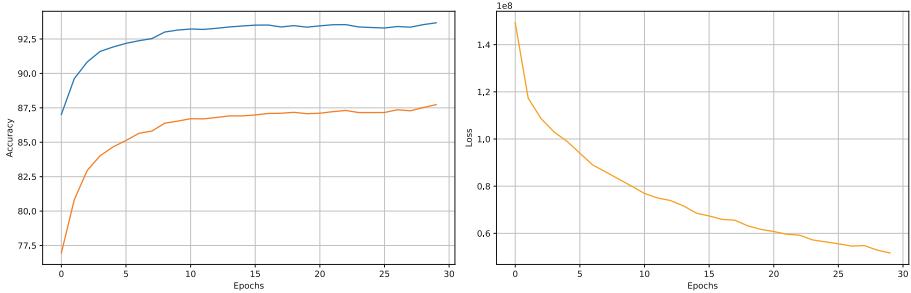
**Fig. 2.** Accuracy and loss function values of HRtagger during training

### 3.3 HRParser

The HRparser is another sub-component of the text processing component and a deep learning-based method. A neural network parser uses a similar configuration for feature extraction as the HRtagger. It employs a dynamic oracle for arc-eager parsing and performs real-time training. The neural network architecture is based on the BERT model and initialized with CSEBert weights.

The HRparser is trained on the hr500k corpus and gradually adjusts the learning rate and warm-up training with adaptable iterations to avoid overfitting.

Adam algorithm with an adaptive learning rate and L2 regularization are used instead of the SGD optimization algorithm. The default value of 0.9 is for Beta1 and 0.999 for Beta2. During the training, the model makes 10 warm-up iterations, gradually lowers the learning rate and does not update the weight values of the CSEBert model. After 26 iterations, the model shows no decrease in loss function or increase in model accuracy, so training stops. Figure 3 shows the values of a loss function and model accuracy during the training phase.



**Fig. 3.** Accuracy and loss function values of HRparser during training

### 3.4 Semantic Hypergraph-Based Knowledge Representation

As mentioned before, the output of the knowledge extraction is knowledge represented in the form of a semantic hypergraph. It is a graph whose vertices and hyperedges have specific labels representing the roles of hyperedges in the sentence. They also determine the semantics of the event expressed in it.

The semantic hypergraph used in this research is an ordered recursive hypergraph. Its ordering property is as essential as the word order in the sentence. The recursive property allows hyperedges to connect with other hyperedges similar to sentence building blocks. These properties make an ordered recursive hypergraph an ideal structure for knowledge representation.

Let  $V = \{v_1, \dots, v_n\}$ , the function  $\prod : V \mapsto U_{v \subseteq V} \pi(v)$  for every vertex  $v \in V$  associates all permutations of set  $\pi(v)$ .

$$\pi(v) = \left\{ \begin{array}{l} \{v\} \mid v \text{ is not a set} \\ \bigcup_{i=1}^{n!} \pi_i(v) \mid v \text{ is a set, } \pi_i(v) \text{ } i^{\text{th}} \text{ per. of } v \end{array} \right. \quad (1)$$

The permutation of the set as shown in Eq. 1 determines all possible iterations of the set. If  $V = \{v_1, \dots, v_n\}$  is a set, then the first permutation  $\pi_1(V)$  is defined with the sequence  $(v_1 \dots v_n)$ , and the second permutation  $\pi_2(V)$  is equal to  $(v_1 \dots v_n v_{n-1})$ , while for the last permutation  $\pi_{n!}(V) = (v_n \dots v_1)$ .

The order of hyperedges determines their recursivity. A hyperedge of order  $k$  contains at least one hyperedge of order  $k - 1$ , at least one hyperedge of order

$k - 2$  and so on up to a hyperedge of order 0. The recursivity of a hyperedge is defined inductively starting from the set of vertices  $P_o = V = v_1, \dots, v_n$ . hyperedge of order 0 is now defined in Eq ??.

$$\Pi(P_0) = \Pi(v_1) \cup \dots \cup \Pi(v_n) = v_1, \dots, v_n \quad (2)$$

Let  $P_1 = 2^{P_0} - \emptyset = \{\{v_1\}, \dots, \{v_1, \dots, v_n\}\}$  a set of all possible hyperedges of order 0 of set V. In this way, an ordered and recursive hypergraph is defined as an ordered pair  $(V, E^{hyper})$  where  $V = \{v_1, \dots, v_n\}$  is a set of vertices, and  $E^{hyper}$  subset  $\bigcup_{i=0}^n \Pi(P_i) = \mathcal{H}_V$  set of all possible hyperedges of highest order n. With such a defined, arranged and recursive hypergraph, the structural foundations for the semantic hypergraph  $H$  have been defined.

## 4 Evaluation Results

The complexity of the knowledge extraction task makes the evaluation task complex too. The text processing component of the KEMMRL relies on state-of-the-art deep learning methods for POS tagging and dependency parsing. Hence, the accuracy of the KEMMRL depends on the accuracy of the applied ML methods, which require evaluation. We evaluated these methods in the testing phase on the hr500k corpus. To further check the accuracy of the HRtagger and HRparser methods, we used an out-of-domain corpus different from the training one [20], the SemCro corpus. We applied the algorithm for measuring the structural similarity of hypergraphs described in the previous section and the metric for measuring correctly labelled semantic roles. The highest average structural similarity to the reference hypergraph was considered the best. The evaluation metric computed the percentage of correctly labelled semantic roles out of all ones. We also evaluated the HRtagger and HRparser methods by comparing them with existing Classla and ReLDI [1] methods and evaluated on the SemCro corpus.

### 4.1 Evaluation Results for HRtagger and HRparser

KEMMRL extracted knowledge by applying rules over syntactic information, which meant the higher the POS tagging precision, the higher the knowledge extraction accuracy. HRtagger method was compared with state-of-the-art methods for the Croatian language and tested on the SemCro corpus. Typically, methods created by training on thematic texts are biased to the text type and genre (science, literature, media, etc.). Therefore, methods developed on the SemCro corpus would likely obtain lower results if used on a different text type or genre. However, the SemCro corpus encompasses various grammatical structures that are not domain-specific but belong to a general language. Table 1 shows the evaluation results obtained for the HRtagger method using the measures of precision, recall, and F-score.

**Table 1.** Evaluation results for HRtagger

Method	hr500k corpus			SemCro corpus		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Classla	94%	94%	94%	95%	94%	95%
HRtagger	95%	95%	95%	94%	93%	94%
ReLDI	92%	92%	92%	92%	92%	92%

Although the HRtagger achieved better results on the hr500k corpus, the evaluation results for the SemCro were slightly worse because it overfitted the training data and evaluated on out-of-domain text the HRtagger was not so efficient. So, we used different hyperparameters and applied regularization methods to reduce overfitting. Comparison to the state-of-the-art POS tagging methods in the knowledge extraction phase revealed that the HRtagger achieved the best results and created knowledge that was more similar to the gold standard. Presumably, because the KEMMRL does not consider morphological features but only POS tags (if the word is a noun or verb etc.). We evaluated the accuracy of HRparser using standard evaluation metrics Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) to compute the percentage of words assigned to the correct head with the proper relation. Also, a label accuracy score (LS) counted the percentage of correctly labelled tokens at the level of all sentences. The HRparser was also evaluated on hr500k and SemCro corpus.

**Table 2.** Evaluation results for HRparser

Method	hr500k corpus		SemCro corpus		
	UAS	LAS	UAS	LAS	LS
Classla	86%	91%	86%	75%	86%
HRparser	89%	94%	95%	86%	88%
ReLDI	86%	90%	77%	53%	66%

As shown in Table 2, the HRparser method achieved the best results, followed by the Classla and the ReLDI methods. Those results were not surprising because the HRparser achieved similar scores when testing on the hr500k corpus. The HRparser architecture used a contextual vector representation and the transition-based method, which resulted in better scores. Using the BERT model, the HRparser achieved significantly better results. HRparser achieved a 9% increase in UAS score for a given sentence and an 11% increase in LAS score on SemCro. This significant improvement opened the possibility of developing a robust HRparser method for knowledge extraction based on the syntactic annotation of text.

## 4.2 The Evaluation of Semantic Hypergraph Representations

We evaluated semantic hypergraph representations using measures for computing the structural similarity of hyperedges and the number of correctly labelled hypergraph vertices. We used methods for extracting hyperedges  $\beta$  and labelling vertices with semantic roles  $\alpha$ . The algorithm shown in 1 was used to evaluate the quality of the  $\beta$  method. The proposed similarity measure captured hyperedges whose average similarity was maximal. Evaluating the structure of a hypergraph consisted in evaluating the method for extracting hyperedges  $\beta$ . Table 3 shows similarity metrics scores obtained for HRtagger and HRparser on the SemCro, compared to other methods.

**Table 3.** The results of similarity metrics

Method	Average similarity
Classla	64.64%
HRtagger + HRparser	67.55%
ReLDI	62.54%

The HRtagger and HRparser showed the best results in measuring the structural similarity of hypergraphs. They used an algorithm for comparing the similarity of the adjacency matrices of two hypergraphs. As expected the method with the best parsing accuracy (due to a dependency tree-based structure) obtained the best results in measuring the structural similarity of hypergraphs. We also evaluated the percentage of semantic roles and labels of hyperedge vertices that the HRtagger and HRparser successfully identified. The  $\alpha$  evaluation method checked the quality of knowledge extracted using HRtagger and HRparser methods. Table 4 shows the results of the  $\alpha$  evaluation method, using Precision, Recall and F1 metrics.

**Table 4.** Results indicating the percentage of correctly identified semantic roles and labels in the SemCro corpus.

Method	Precision	Recall	F1 score
Classla	45.79%	47.79%	45.73%
HRtagger + HRparser	47.70%	47.57%	47.03%
ReLDI	44.88%	44.87%	44.35%

All methods performed approximately the same. We calculated the percentage of semantic roles and labels of hyperedge vertices that the HRtagger and HRparser successfully identified in each sentence. We compared automatically

identified semantic roles to manually annotated ones in the SemCro corpus, representing the gold standard. The HRtagger + HRparser methods showed the best results when applying the alpha method. Although the HRtagger showed slightly lower results, the high precision of HRparser compensated for it.

## 5 Conclusion and Future Work

This paper presents the model for knowledge extraction for morphologically rich languages (KEMMRL) as the Croatian language and semantic hypergraph representation. An extensive evaluation demonstrates the efficiency and accuracy of the newly developed HRtagger and HRparser methods compared to the state-of-art methods for text processing in the Croatian language. Most systems that use text in natural language can represent knowledge using semantic hypergraphs. Automatic similarity metrics can capture hypergraphs' similarity. A semantic hypergraph is the most suitable form for knowledge representation and evaluation based on a similarity algorithm. The application of the semantic hypergraph is not limited to knowledge-based systems but also other natural language processing tasks such as text summarization. Lengthy texts can be converted into semantic hypergraphs and then compressed by reducing the number of edges and vertices in the semantic hypergraph. The KEMMRL uses a statistical model for hyperedge identification and building without handmade rules. Future research will include more precise knowledge extraction rules and a statistical model for semantic hypergraph building. The objective is to apply this model as knowledge representation in an intelligent tutoring system. The implementation of a semantic hypergraph into modules of an ITS can enable the system to represent complex knowledge structures and relationships between concepts more effectively. This can lead to more personalized and adaptive tutoring experiences for learners, as the system can better understand their knowledge gaps and provide targeted feedback and resources.

**Acknowledgements.** The paper is part of the work supported by the Office of Naval Research Grant No.N00014-20-1-2066

## References

1. ReLDIanno – text annotation service for processing slovenian, croatian and serbian – CLARIN Slovenia. <https://www.clarin.si/info/k-centre/web-services-documentation/>. Accessed 21 Jan 2022
2. Agić, Ž., Merkler, D., Berović, D.: Parsing croatian and serbian by using croatian dependency treebanks. In: Proceedings of the 4th Workshop on Statistical Parsing of Morphologically-Rich Languages, pp. 22–33. Association for Computational Linguistics, Seattle, Washington, USA (2013). <https://aclanthology.org/W13-4903>
3. Agić, Ž., Tiedemann, J., Merkler, D., Krek, S., Dobrovoljc, K., Može, S.: Cross-lingual dependency parsing of related languages with rich morphosyntactic Tagsets. In: Proceedings of the EMNLP’2014 Workshop on Language Technology for Closely

- Related Languages and Language Variants, pp. 13–24. Association for Computational Linguistics, Doha, Qatar (2014). <https://doi.org/10.3115/v1/W14-4203>. <https://aclanthology.org/W14-4203>
4. Željko Agić, Ljubešić, N.: The SETIMES.HR linguistically annotated corpus of croatian, pp. 1724–1727 (2014). <http://nlp.ffzg.hr/resources/corpora/>
  5. Željko Agić, Ljubešić, N.: Universal dependencies for croatian (that work for serbian, too), pp. 1–8 (2015). <http://universaldependencies.github>, <http://bsnlp-2015.cs.helsinki.fi/bsnlp2015-book.pdf>
  6. Željko Agić, Ljubešić, N., Merkler, D.: Lemmatization and morphosyntactic tagging of croatian and serbian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pp. 48–57 (2013). <http://www.nljubesic.net/resources/corpora/setimes/>, <http://nlp.ffzg.hr/resources/models/tagging/>
  7. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web, pp. 2670–2676 (2007)
  8. Batanović, V., Cvetanović, M., Nikolic, B.: A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. PLOS ONE 15, e0242050 (2020). <https://doi.org/10.1371/journal.pone.0242050>
  9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). <http://arxiv.org/abs/1810.04805>
  10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159 (2011). <https://doi.org/10.1109/CDC.2012.6426698>, <http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html>
  11. Eberendu, A.C.: Unstructured data: an overview of the data of big data. Int. J. Comput. Trends Technol. 38, 46–50 (2016). [https://doi.org/10.14445/22312803\\_IJCTT-V38P109](https://doi.org/10.14445/22312803_IJCTT-V38P109)
  12. Erjavec, T.: Multext-east: morphosyntactic resources for central and eastern European languages. Lang. Res. Eval. 46, 131–142 (2012). <https://doi.org/10.1007/s10579-011-9174-8>, <https://link.springer.com/article/10.1007/s10579-011-9174-8>
  13. Halácsy, P., Kornai, A., Oravecz, C.: Hunpos-an open source trigram tagger, pp. 209–212 (2007). <https://doi.org/10.5555/1557769>, <http://mokk.bme.hu/resources/hunpos/>
  14. Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., Erjavec, T.: Training corpus hr500k 1.0 (2018). <http://hdl.handle.net/11356/1183>, slovenian language resource repository CLARIN.SI
  15. Ljubesic, N., et al., (eds.) Proceedings of the 10th International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016. European Language Resources Association (ELRA) (2016). <http://www.lrec-conf.org/proceedings/lrec2016/summaries/340.html>
  16. Ljubešić, N., Lauc, D.: BERTić- The transformer language model for bosnian, croatian, montenegrin and serbian, pp. 37–42 (2021). <https://www.clarin.si/info/k-centre/>, <http://arxiv.org/abs/2104.09243>
  17. de Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D.: Universal dependencies. Comput. Linguist. 47, 255–308 (2021). <https://doi.org/10.1162/COLI-a-00402>, <http://universaldependencies.org/>
  18. Menezes, T., Roth, C.: Semantic hypergraphs. CoRR abs/1908.10784 (2019). <http://arxiv.org/abs/1908.10784>
  19. Menezes, T., Roth, C.: Semantic hypergraphs. <https://arxiv.org/abs/1908.10784> (2019). <https://doi.org/10.48550/ARXIV.1908.10784>

20. Paroubek, P., Chaudiron, S., Hirschman, L., Chaudiron, S., Hirschman, L.: Principles of evaluation in natural language processing. *Revue TAL* **48**, 7–31 (2007). <http://www.technolangue.net>
21. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951). <https://doi.org/10.1214/AOMS/1177729586>
22. Stankov, S., Rosić, M., Žitko, B., Grubišić, A.: Tex-sys model for building intelligent tutoring systems. *Comput. Educ.* **5**, 1017–1036 (2008)
23. Ulčar, M., Robnik-Šikonja, M.: Finest Bert and Crosloengual Bert, pp. 104–111. Springer International Publishing (2020)
24. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017). <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>



# Synthesizing Didactic Explanatory Texts in Intelligent Tutoring Systems Based on the Information in Cognitive Maps

Viktor Uglev<sup>1</sup> and Oleg Sychev<sup>2</sup>

<sup>1</sup> Siberian Federal University, Zheleznogorsk, Russia  
vauglev@sfu-kras.ru

<sup>2</sup> Volgograd State Technical University, Volgograd, Russia

**Abstract.** The paper deals with the problem of synthesizing texts that explain decisions of an Intelligent Tutoring System for the learner. Solving this problem is important for increasing human trust in the ITS decisions by finding effective arguments and synthesizing explanatory texts. We describe an approach to preparing and concentrating raw data that is similar to the process used for mapping. The texts explaining the system's decisions are generated using the method of parametric maps, which are visualized as Cognitive Maps of Knowledge Diagnosis: the transition from particular maps to the combined map and the development of a system of arguments based on these data. We show the structure of an explanatory text and the accompanying visualization of a cognitive map. We demonstrate the explanation synthesis on the example of a graduate student in the course “Simulation modeling”. The explanation consists of the text and supporting visualization. Preliminary analysis has shown significant student interest in receiving explanations, containing verbal and non-verbal (cognitive maps) components, from the Intelligent Tutoring System about its decisions.

**Keywords:** Intelligent Tutoring Systems · cognitive visualization · decision making · explainable AI · explanatory text · Cognitive Maps of Knowledge Diagnosis

## 1 Introduction and Related Works

Using Intelligent Tutoring Systems (ITS) in online learning involves making difficult decisions corresponding to the learning situation. A human learner is more complex than any current tutoring system, including intelligent systems, and not inclined to trust unknown sources; this lowers learners' trust in recommendations and decisions made by ITS [1, 9]. Human-computer communication happens through the ITS interface and can take different forms: text, synthesized speech, images, and changes in the availability of navigation widgets. When ITS decisions are not clear to the learner and the system simply states them, it can irritate the learner and lower the learning speed [2, 12]. This problem can be solved by organizing a dialogue between ITS and the learner to let the learner

learn more about the reasons for the ITS decisions while the system can update the learner model using the information acquired through the explanatory dialogue.

Dialogue in ITS can be divided into two chief categories: dialogue related to the studied material (content-related dialogue) and dialogue related to learning process (didactic dialogue). Some ITS use static content-related dialogue entered by authors or teachers (e.g., in the form of scaffolding questions [11]). It can be dynamically generated based on ontologies of subject domains (e.g., [10, 17, 24]). But developing specific subject-domain models is very time-consuming and their usage is limited to the courses teaching the relevant subject domain.

Didactic dialogue is more generalizable because decisions making and explanatory text synthesis in ITS can be unified if the course specifics are described in general terms by a learning-process ontology (e.g., [5]) or using other methods of learning-unit organization (e.g., [22, 25]). So we can develop a generalized method of generating explanations which allow answering questions like “Why and how this decision was made?” or “Why do I need this?”. Let’s call this kind of dialogue a didactic dialogue. It can be useful in forming personalised learning collections and personalised learning trajectories when showing hints and recommendations about topics to study next, during an assessment of the learning situation, and while discussing the situation [19]. The dialogue mode can be initiated by the learner or by the ITS. The topic of didactic dialogue can be either a part of a course (e.g., a term, assignment, learning unit, topic, course or curriculum) or other learning entities (e.g., competencies, standards, methods of grading and decision-making, learner’s actions). This makes it important to use both verbal and non-verbal (e.g., cognitive maps) methods of supporting the dialogue. The didactic dialogue starts from a well-reasoned explanatory text which starts the dialogue. In this paper, we consider the problem of synthesizing explanatory texts that start didactic dialogues with an ITS.

Table 1 shows a comparative analysis of modern ITS with developed explanatory text capabilities. They mostly lack complex solutions using flexible (for different entities) support of the didactic dialogue process with explanations and supporting cognitive visualization. No ITS has “Yes” in all the columns. So further research in this field is relevant.

In this paper, we consider an approach to explaining didactic decisions in ITS based on the information concentrated for cognitive maps (including their visualization). We briefly sketch the technology of concentrating and mapping data, describe the stages of synthesizing explanatory text and the templates used, consider a case study of didactic explanation and formulate recommendations for developing and using the explainable AI approach for ITS.

## 2 Method of Synthesizing Explanatory Text

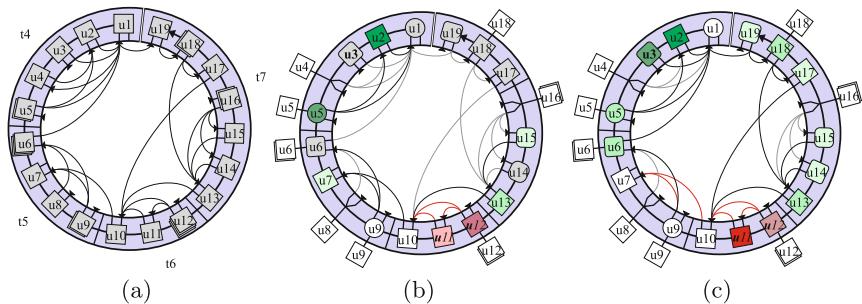
### 2.1 Data Concentration and Mapping

Consider the model of an e-learning course, consisting of course elements, connections between them, and supporting entities. ITS logs (as a digital learning

**Table 1.** Comparison of didactic-explanation characteristics of different ITS

System	Interdisciplinary connections	Changing the subject and level of abstraction	Explanation of decision making	Graphic supplement
GIFT [8]	Yes	No	No	No
Mastery Grid [3]	No	No	No	Yes
ITS with knowledge graphs [18]	No	No	No	Yes
AT-TECHNOLOGY [15]	Yes	No	No	No
DeepTutor [14]	No	Yes	Yes	No

footprint) time and frequency data concerning working with course elements, answers to tests and assignments, and survey data. These data require generalization (concentration) both for determining the control action and synthesizing its explanation. For this, a chain of concentration is used: statistical concentrations → metric concentration → semantic concentration → logical concentration. At the stage of metric concentrations, mapping techniques are used: the basic model of the e-learning course in the Cognitive Map of Knowledge Diagnosis (CMKD) notation (see Fig. 1a) is overlaid with the digital footprint data regarding the chosen analysis aspect, time dynamics and decision-making level (see the visualized example of a cognitive map for competency aspect in Fig. 1b; it is explained in detail in [20]). These cognitive maps will be discussed in Sect. 3 during a case study.

**Fig. 1.** Basic CMKD for the course “Simulation modeling” (a) and its individualized variants for the competence (b) and course (c) aspects.

## 2.2 Explanatory Form

Explanatory form is shown as the ITS reaction to an event (e.g., passing a control point in the course) or at the learner’s initiative (e.g., by pressing the button

to ask for explanation while watching learning material). Further in this paper, we will consider using the explanatory form to explain the system's decision as a reaction to an event which is the most important case. Didactic explanation combines verbal (text) and non-verbal (cognitive map in the visualized form) parts.

The basic explanatory text generated by ITS must include the following:

- the sentence of qualitative evaluation of learner's actions or the learning situation  $\alpha$ ;
- sentences, briefly describing the current situation, emphasizing 1–3 relevant parameters  $\beta_j$ ;
- sentences  $\gamma_k$ , formulating recommendations for further work with the course elements (mentioning the expected result regarding the learner's goals) for furthering pedagogical intervention and increasing the learner's motivation;
- the sentence that invites the learner for further explanatory dialogue  $\delta$ .

The non-verbal part of the explanation is shown as a CMKD (basic, extended, or simplified), focusing attention on the subject matter. It is oriented at convincing the learner in the potential usefulness of following the system recommendations. The invitation to continue dialogue is supplemented with the necessary controls (request constructor).

According to our method, to generate the verbal part of the explanatory form, ITS should perform the following actions, based on the cross-cutting approach to decision-making on the course level [22]:

1. statistic concentration of data regarding the chosen (by ITS) pedagogical intervention for each learning unit  $u_i$  with overlaying of these data on the individualized course map for each relevant aspect (the relevant aspects are selected according to the current learning problems as shown in [20]);
2. uniting all the problematic elements  $u_i'$  in the combined parametric map, characterizing the situation as a whole, which can be extended by inserting additional entities (competences, learner's goals, etc., which can be semantically linked to the relevant units  $u_i'$ );
3. evaluation of the significance of the learning problems using expert-systems methods and sorting values in  $u_i'$  according to their significance;
4. simplifying the combined map by removing the less-significant elements, emphasizing the elements  $u_i''$ , which are most significant for achieving the result of the pedagogical intervention;
5. selecting 1–2 key problems from  $u_i''$  and synthesizing the sentence  $\alpha$  according to them;
6. iterating over  $u_i''$  and synthesizing sentences  $\beta_j$ , reflecting the evaluation of mastering learning units or their groups, indicating the found causes of problems in learning and the linked entities;
7. iterating over  $u_i''$  and synthesizing (for each problematic unit or their group) 1–3 sentences  $\delta_k$ , mentioning the learner's action that will have the best effect according to the ITS estimation: these sentences include didactic evaluation

- of learning units (their groups) stating found causes of problems and the relevant entities, enhanced with personally-oriented arguments according to the learner's model;
8. showing the sentences  $\alpha$ ,  $\beta_j$  and  $\delta_k$ ;
  9. selecting the most relevant analysis aspect according to the goal of pedagogical intervention and showing the relevant CMKD (typically, in the simplified form);
  10. showing the invitation to continue dialogue  $\delta$  and request constructor.

The method shown above is based on the ITS ability to concentrate data into CMKDs and give the learner an option to interact with the map.

The learner, after reviewing the content of the explanatory form, can initiate dialogue (ask to synthesize a new form by their request), jump to one of the mentioned learning units using its hyperlink, return to one of the previous forms (later during the dialogue), start the process of clarifying their preferences (correcting personal goals), or close the explanatory form.

### 2.3 Text Synthesis

In the developed ITS AESU, we employed the template-based text synthesis technique. Didactic texts use dictionaries of user-created learning entities which are entered when a course is published in the tutoring system. Teachers cannot affect the general structure of the explanation and its sentences; they can only enter a textual description of entities in their courses. The text is synthesized according to the specifics of a particular learning situation and the learner's goals.

Let us consider the generation of each block of explanatory text in the situation when the system formulates recommendations on further study of the course after the mid-term exam. The first block of explanatory text ( $\alpha$ ) consists of three parts: statement of the facts (grade) and generalized problem; quality of the learner's work when studying the course, and a brief listing of the learned objects which are recommended for further study.

**Grade and Generalized Problem.** This sentence is generated according to the following template (translated to English):

Your learning results have  $P_{1.1}$  dynamics (the current grade is  $P_{1.2}\%$   $If_1\{$  and it  $P_{1.3}$  by  $P_{1.4}\%$  in comparison with the previous results})  $If_2\{$ , but  $P_{1.5}$  problems arise because of inadequate absorption of some elements of the previously studied material}.

The parameters  $P$  and conditions  $If$  are determined as follows:

- $P_{1.1}$  is a qualitative estimate, calculated using fuzzy-logic methods (see [23]) from quantitative grades for the current and previous exams and can be one of the following: {highly positive, positive, slightly positive, slightly negative, negative, highly negative};

- $P_{1.2}$  and  $P_{1.4}$  are calculated from the grades in the digital learning footprint;
- $P_{1.3}$  is an evaluation of the difference between the latest and previous grade, which takes a value from the following set {increased, same, decreased};
- $P_{1.5}$  is a fuzzy estimate of the level of problems in the current learning situation, calculated by an expert system, and can be one of the following {no, some, significant, very significant};
- the block  $If_1$  is shown only if  $P_{1.1}$  is “slightly positive”;
- the block  $If_2$  is shown only if  $P_{1.1}$  is not “highly positive”.

The usage of quantitative fuzzy estimates allows making the explanatory text closer to natural-language dialogue and stimulates the learner to ask for further clarification of the summarized statements made by the ITS.

**Quality of the Learner’s Work.** This part uses the following template:

Your intensity of learning is estimated as  $P_{2.1}$ , while your intention to achieve your stated goals is estimated as  $P_{2.2}$ .

Both qualitative parameters of this sentence are estimated by the ITS based on the data in the digital learning footprint, including the data concentrated for the combined CMKD. Intensity of learning is based on the evaluating the frequency of using learning materials in the ITS and time spent learning; the method of estimating the intent to achieve student’s stated goals is described in [19]. The values of  $P_{2.1}$  and  $P_{2.2}$  can be one of the following: {low, average, high}.

**Recommendation for Further Study.** If there are problems in the learning situation (i.e., the value of  $P_{1.5}$  is different from “no problems”), the explanatory message is enhanced with the sentence, recommending the learning units that, according to the system estimates, can help with solving the learning problems most. The following sentence template is used:

Please,  $If_3\{\text{first of all}\}$  pay more attention to  $If_1\{P_{3.1}, P_{3.2}\}$ .

The parameters  $P$  and the output condition  $If$  are determined as follows:

- $P_{3.2}$  is the list of the most significant learning units from  $u_i$ , which were detected during analysis of the combined CMKD (1–3 units), which are joined by commas and the conjunction “and” depending on their number;
- $P_{3.1}$  changes depending on the number of recommended units (i.e., singular “learning unit” or plural “learning units” depending on the number of elements in  $P_{3.2}$ );
- the block  $If_3$  is shown if  $u_i$  contains more than 3 values for  $P_{3.2}$ .

Russian is a synthetic language [6] so it has more diverse morphology than partly analytical languages like English. The template method of text synthesis allowed us to avoid problems with word form changes. The user-depending words and collocations are entered in the system dictionaries in the required case; an effort was made to formulate all the template phrases so that each user-depending parameter is used in one case only.

### 3 Case Study

For a case study of explanatory text generation, we used the learning situation of a graduate student majoring in “Informatics and Computing” (3rd semester, Siberian Federal University) in the ITS AESU<sup>1</sup>). The learner studies the course “Simulation modeling” (code *d3.2*; the basic CMKD for this course is shown in Fig. 1a), which consists of 19 learning units, combined into four topics (from  $t_4$  to  $t_7$  according to the cross-cutting numbering of semester’s topics). The ITS has the results of mid-semester exams for all the semester courses (grades and competence profile according to [20]), survey data (the student indicated their preferences regarding learning units, competences *PK-7*, *OK-3* and *UK-1*, and skill development  $\psi_6$ ). We consider the case of the explanation that is shown after the event “showing the results of mid-term exams for the course *d3.2*” by the ITS planner.

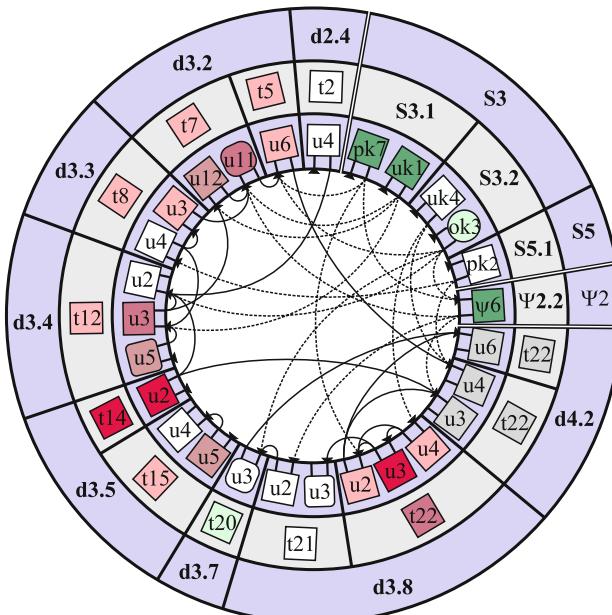
The ITS decided to recommend the student to work more with some of the more significant (in the context of the current learning situation) learning units of the course *d3.2*, which was supported by generating textual arguments and cognitive visualization (the expected reaction of the student is working on the recommended learning units). The following actions were performed automatically:

1. the data regarding knowledge and competence profile of the student were gathered from their digital learning footprint; they were overlaid on the individualized parametric multi-aspect map (the map is shown in the CMKD notation in the course aspect in Fig. 1c and shown in the competence aspect in Fig. 1b for the competency *UK-1*); for each generated map the problematic units were selected according to the method described in [20] (these were the learning units  $u_1, u_6, u_7, u_9, u_{10}, u_{11}, u_{12}, u_{18}$  and  $u_{19}$ );
2. the combined parametric map was generated (which united all “problematic” elements (with zero or negative certainty factor values; visualized as white and red figures) both from the target course (*d3.2*, 41 learning units were selected) and the linked courses (directly or indirectly) (the set  $u_i'$ , uniting the learning units from the 6 courses of the current semester, one course from the previous semester, and one course from the next semester) and also the personal preferences and key learning entities that were closely linked to the problematic elements (the competences *PK-2*, *PK-7*, *UK-1*, *UK-4*, *OK-5* and the personal goal  $\psi_6$ );
3. for each of the elements from  $u_i'$  the complex estimate of the degree of their “level of being problematic” was made (a production-based expert system with composite reasoning rules [4]) and sorted according to it (e.g., for  $u_6$  the certainty coefficient of being problematic was 0.62, while for  $u_7$  it was 0.17);
4. the new set  $u_i''$  was created, which included the most significant elements from  $u_i'$  ( $u_6, u_{11}, u_{12}$ ) for the course *d3.2* and other entities (27 elements

---

<sup>1</sup> <https://aesfu.ru/>.

total); a simplified combined parametric map was synthesized (it is visualized in Fig. 2);



**Fig. 2.** A combined parametric map in the CMKD notation for the current learning situation.

5. the expert system analyzed all the elements from the combined parametric map belonging to the course  $d3.2$  (in  $u_i$ " the most significant problems were found in the competence aspect for the competence  $UK-1$  which was a priority for the student, which was caused by the low mastery of the learning unit  $u_7$  "Monte-Carlo method" (the learning material from the course core which is semantically linked to the competence  $UK-1$  which is important for the student and determines the success in studying the following learning unit  $u_{10}$ ) and  $u_{10}$  "Simulation of analog processes using pseudo-random numbers" (this learning unit is semantically linked to  $\psi_6$  and causes problems in studying the dependent units  $u_{11}$  and  $u_{12}$ ); this resulted in the synthesis of the following  $\alpha$  block of text (hyper references are underscored on the upper block in Fig. 3);

Your learning results have positive dynamics (your grade is 76%; it increased by 3% compared to the previous grade), but significant problems arise because of inadequate absorption of some elements of the previously studied material. Your intensity of learning is estimated as average, while your intention to achieve your stated goals is estimated as low. Please, first of all, pay more attention to "Monte-Carlo

method" ( $u_7$ ) and "Simulation of analog processes using pseudo-random numbers" ( $u_{10}$ ).

#### Description of the learning situation

Your learning results have **positive dynamics** (your grade is **76%**; it **increased by 3%** compared to the previous grade), but significant problems arise because of inadequate absorption of some elements of the previously studied material. Your intensity of learning is estimated as **average**, while your intention to achieve your stated goals is estimated as **low**. Please, first of all, pay more attention to "[Monte-Carlo method](#)" ( $u_7$ ) and "[Simulation of analog processes using pseudo-random numbers](#)" ( $u_{10}$ ).

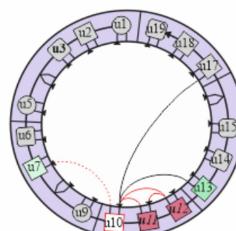
#### Recommendations for working with the course - 4

##### Recommendation №1 about theme №2 (important)

The found problems in the learning unit "[Random variables and their characteristics](#)" ( $u_6$ ) belong to the assignments, verifying the development of the competence "[Can estimate modes of operation of IT-subsystems of space systems and take them into account during development](#)" (PK-7) (the competence development level is estimated as **not developed**) and "[Can analyse professional information, highlight the most important structure, draw up documents, and present with well-grounded conclusions and recommendations](#)" (OK-3) (the competence development level is estimated as **likely not developed**), while you indicated that you consider these competencies important for your professional development; the low level of their development keeps your from reaching the goal "[Study the methods of working in specialized software for modeling and engineering design](#)" ( $\psi_6$ ).

##### Recommendation №2 about theme №2 (critical)

The found problems in the learning unit "[Random variables and their characteristics](#)" ( $u_6$ ) belong to the assignments, verifying the development of the competence "[Can estimate modes of operation of IT-subsystems of space systems and take them into account during development](#)" (PK-7) (the competence development level is estimated as **not developed**) and "[Can analyse professional information, highlight the most important structure, draw up documents, and present with well-grounded conclusions and recommendations](#)" (OK-3) which are important for you, and also let you better understand the problematic unit "[Formalizing transaction streams in distributed systems](#)" ( $u_5$ ).



Map legend and explanation for its interpretation

#### Detailed explanation for the selected recommendation (№1)

In the topic 2 "[Distributions of pseudo-random variables and their simulation](#)" ( $d_5$ ) you should work more on the learning unit "[Random variables and their characteristics](#)" ( $u_6$ ), which will allow to better develop the competences "[Can estimate modes of operation of IT-subsystems of space systems and take them into account during development](#)" (PK-7) and "[Can analyse professional information, highlight the most important structure, draw up documents, and present with well-grounded conclusions and recommendations](#)" (OK-3) which are important for you, and also let you better understand the problematic unit "[Formalizing transaction streams in distributed systems](#)" ( $u_5$ ).

#### Ask question to the tutoring system



[RETAKE THE INTERMEDIATE TEST](#)

[GO BACK TO THE COURSE](#)

**Fig. 3.** Explanatory form for the student described in the case study enhanced by the visualization of simplified CMKD; course "Simulation modeling" (translated from Russian).

6. for each element of  $u_i''$ , belonging to  $d3.2$ , the synthesis of the sentences  $\beta_j$ , was conducted to describe the found problem as in the following example for the learning unit  $u_6$  (the middle block in Fig. 3):

The found problems in the learning unit "[Random variables and their characteristics](#)" ( $u_6$ ) belong to the assignments, verifying the development of the competence "[Can estimate modes of operation of IT-subsystems of space systems and take them into account during development](#)" (PK-7) (the competence development level is estimated as **not developed**) and "[Can analyse professional information, highlight the most important structure, draw up documents, and present with well-grounded conclusions and recommendations](#)" (OK-3) (the competence development level is estimated as **likely not developed**), while you indicated that you consider these competencies important for your professional development; the low level of their development keeps your from reaching the goal "[Study the methods of working in specialized software for modeling and engineering design](#)" ( $\psi_6$ ).

7. the sentences  $\delta_k$  were synthesised for each element of  $u_i''$  to describe the recommended actions and their expected outcomes as in the following example for the unit  $u_6$  (the lower block of the explanatory text in Fig. 3):

In the topic 2 “Distributions of pseudo-random variables and their simulation” ( $t_5$ ) you should work more on the learning unit “Random variables and their characteristics” ( $u_6$ ), which will allow to better develop the competences ‘Can estimate modes of operation of IT-subsystems of space systems and take them into account during development’ (PK-7) and “Can analyse professional information, highlight the most important, structure, draw up documents, and present with well-grounded conclusions and recommendations”(OK-3) which are important for you, and also let you better understand the problematic unit “Formalizing transaction streams in queueing models” ( $u_{11}$ ). This recommendation is given to you for the second time!

8. the text  $\alpha$ ,  $\beta_j$  and  $\delta_k$  is shown to the student;
9. the most important analysis aspect for the recommendations in  $\alpha$  was identified as the course aspect (for  $u_{11}$  and  $u_{12}$ ), which allowed to generate the simplified CMKD (see Fig. 3), emphasizing the necessary cause-effect links by red colour, to enhance the generated text;
10. the sentence  $\delta$  was shown to invite the student to discuss the system to get more information about the ITS decision: see the lower part of the window in Fig. 3 with the request constructor (see Fig. 4) that is collapsed for now to let the student concentrate on the explanation first.

Ask question to the tutoring system		
Kind of the question Why	Subject Learning material fragment	Value u_7 - Monte-Carlo method
Process of interest Assessed / defined	Parameter Subject assessment	Value Not successful enough
<input checked="" type="checkbox"/> Visualize values for entities that aren't mentioned in the selected recommendation <input type="checkbox"/> Visualize external links to other types of entities on the map <input type="checkbox"/> Go beyond the current scope of analysis for the explanation		
<input type="button" value="PERFORM REQUEST"/> <input type="button" value="HELP"/>		

**Fig. 4.** Request constructor form in ITS AESU with filled fields (translated from Russian).

After reading the explanatory text shown above, the graduate student followed the system recommendation and jumped to the learning unit  $u_{10}$ . In general, the student whose situation was shown in this example, received 114 didactic explanations while studying the course “Simulation modeling” in 62%

of occurrences he used the links in the text or visualized map to jump to the recommended learning material, while in 38% of occurrences, the student started a dialogue with the ITS to deepen their understanding of the reasons of ITS decisions.

## 4 Results and Discussion

Textual and graphic explanations of the ITS decisions, synthesized according to the method described above were used to support experimental online learning for graduate students majoring in “Informatics and computing” in Siberian Federal University enrolled in the 2020 year (see [22] for details of method). For each course, we formed individual learning trajectories, taking into account learners’ goals. So, while the curriculum was the same for all the learners, its content differed by the lists of learned topics and levels of control of knowing each topic. Students received recommendations from ITS; they could ask for explanations of these recommendations in text form, which was supported by cognitive visualization in the CMKD notation. We recorded all learners’ dialogues with the ITS, the frequency of asking for explanations of ITS decisions, and the following jumps to the pages with learning content. At the end of each course, the students completed a survey about their trust in ITS decisions.

The high percentage of asking for explanations (up to 68% cases of asking for didactic explanations) is consistent with the findings of Bodily and Verbert [2], “if students know why they receive a particular recommendation, it could increase their trust in the system along with the likelihood of them following feedback provided by the system”. Our results are also consistent with most of the findings of Ilves et al. [7], but we did not find a lowering in the quality of learning for the different groups of learners (see [22]).

We consider these results preliminary; they require a more thorough study (e.g., to determine the comparative effectiveness of different parts of the explanation and improve them). However, our current findings let us recommend the following:

- for making decisions and explanation generation, online course model (structural component) and target activities model (functional component) require not just developing hierarchies of entities, but also links between them and graded tasks (tests and assignments);
- learner surveys must contain information necessary to find personal preferences on different education levels (from a particular course to the whole curriculum);
- the textual explanation must be compact to avoid overloading the learner with the volume of information and too many details [16];
- templates for repeated sentences (e.g., for the parts  $\beta_j$  and  $\delta_k$ ) should have several variants (e.g., using synonyms) to increase text diversity [13].

The disadvantage of the proposed approach is the very labour-intensive process of creating the knowledge base of the course (curriculum) that is necessary

for making automatic decisions in the learning situation and explaining these decisions. The advantage is applying a unified approach of analysing learning situations to a wide range of courses. In our experience, an experienced teacher can specify a new course and gather personal preferences using ITS services without spending too much time: a course can be parameterized in 1–1.5 h, while formalization of the existing bank of assignments and tests requires 1–2 working days. The data for CMKD is concentrated automatically and can be used to generate textual explanations according to our method, which makes explanatory forms enhancement of existing decision-supporting techniques which does not require additional input from the teacher.

## 5 Conclusion

The quality of communication between a learner and an ITS depends heavily on the level of trust in the system's decisions. Combining methods of text synthesis and cognitive visualization to explain AI decisions increases the effectiveness of online learning. We found a steady interest to the didactic explanations provided by the ITS for its decisions both supported by cognitive visualizations (up to 68% of cases) and without graphic support (in about 30% of cases).

Our further work will be aimed at complex support of ITS decision with natural-language dialogue with elements on cognitive visualizations on the curriculum level. Adding the ability to make further requests for explanations will make the system more interactive. We also plan to perform in-depth evaluation of the cross-cutting approach described in [22] and graphic notations CMKD and UGVA ([21]) to find its influence on the students' learning gains. It is also important to upgrade the sentence templates based on the experimental data, increase their variety, and adding support of other natural languages than Russian.

## References

1. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020). <https://doi.org/10.1016/j.inffus.2019.12.012>
2. 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. LAK 2017, Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3027385.3027403>
3. Brusilovsky, P., Rus, V.: Social Navigation for Self-Improving Intelligent Educational Systems, pp. 131–145. Army Research Laboratory (2019). <https://www.pitt.edu/peterb/papers/SocNav4SIS.pdf>
4. Fullér, R., et al.: Fuzzy Reasoning and Fuzzy Optimization. Turku Centre for Computer Science Turku, Finland (1998)
5. Gavrilova, T.A., Kokoulina, L.: Using ontology engineering to design an artificial intelligence course. In: Uskov, V.L., Howlett, R.J., Jain, L.C. (eds.) Smart Education and e-Learning 2019. SIST, vol. 144, pp. 201–207. Springer, Singapore (2019). [https://doi.org/10.1007/978-981-13-8260-4\\_19](https://doi.org/10.1007/978-981-13-8260-4_19)

6. Haspelmath, M., Michaelis, S.M.: Analytic and synthetic: typological change in varieties of European languages. In: Language Variation-European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), pp. 3–22. John Benjamins, Amsterdam (2017)
7. Ilves, K., Leinonen, J., Hellas, A.: Supporting self-regulated learning with visualizations in online learning environments. In: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, pp. 257–262. Association for Computing Machinery, New York, USA (2018). <https://doi.org/10.1145/3159450.3159509>
8. Kelsey, E., Ray, F., Brown, D., Robson, R.: Design of a domain-independent, interactive, dialogue-based tutor for use within the GIFT framework. In: Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (giftsym3), pp. 161–168 (2015)
9. Mashbitz, E., Andrievskays, V., Komissarova, E.: Dialog in a Tutoring System. Higher School, Kiev (1989). (in Russian)
10. O'Rourke, E., Andersen, E., Gulwani, S., Popović, Z.: A framework for automatically generating interactive instructional scaffolding. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1545–1554. CHI 2015, Association for Computing Machinery, New York, NY, USA (2015). <https://doi.org/10.1145/2702123.2702580>
11. Razzaq, L., Heffernan, N.T.: Scaffolding vs. hints in the assistance system. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 635–644. Springer, Heidelberg (2006). [https://doi.org/10.1007/11774303\\_63](https://doi.org/10.1007/11774303_63)
12. Risha, Z., Barria-Pineda, J., Akhuseyinoglu, K., Brusilovsky, P.: Stepwise help and scaffolding for java code tracing problems with an interactive trace table. In: Proceedings of the 21st Koli Calling International Conference on Computing Education Research. Koli Calling 2021, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3488042.3490508>
13. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method. Wiley, Hoboken (2016)
14. Rus, V., Stefanescu, D., Niraula, N., Graesser, A.C.: DeepTutor: towards macro- and micro-adaptive conversational intelligent tutoring at scale. In: Proceedings of the First ACM Conference on Learning @ Scale Conference. p. 209–210. L@S 2014, Association for Computing Machinery, New York, NY, USA (2014). <https://doi.org/10.1145/2556325.2567885>
15. Rybina, G., Slinkov, A.: The implementation of the ontological approach to control of the processes of designing integrated expert systems based on the problem-oriented methodology. In: Kovalev, S.M., Kuznetsov, S.O., Panov, A.I. (eds.) RCAI 2021. LNCS (LNAI), vol. 12948, pp. 354–364. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86855-0\\_25](https://doi.org/10.1007/978-3-030-86855-0_25)
16. Stefanuk, V.L.: Learning levels in intelligent tutoring systems. IEICE Trans. Inf. Syst. **78**(9), 1103–1107 (1995)
17. Sychev, O., Anikin, A., Penskoy, N., Denisov, M., Prokudin, A.: CompPrehension - model-based intelligent tutoring system on comprehension level. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 52–59. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-80421-3\\_6](https://doi.org/10.1007/978-3-030-80421-3_6)
18. Troussas, C., Krouská, A.: Path-based recommender system for learning activities using knowledge graphs. Information **14**(1), 9 (2023)
19. Uglev, V., Rogozhnikova, M.: Individualized natural language dialogue with the students in the intellectual education system. In: Interactive Systems and Tech-

- nologies: the Problem of Human-Computer Interaction, vol. 11, pp. 255–259 (2015). [http://conf-is.ulstu.ru/sites/default/files/IS\\_2015\\_Part\\_II.1.pdf](http://conf-is.ulstu.ru/sites/default/files/IS_2015_Part_II.1.pdf)
- 20. Uglev, V., Sychev, O.: Creating and visualising cognitive maps of knowledge diagnosis during the processing of learning digital footprint. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 93–98. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-80421-3\\_11](https://doi.org/10.1007/978-3-030-80421-3_11)
  - 21. Uglev, V., Sychev, O.: Evaluation, comparison and monitoring of multiparameter systems by unified graphic visualization of activity method on the example of learning process. Algorithms **15**(21), 468 (2022). <https://doi.org/10.3390/a15120468>
  - 22. Uglev, V., Sychev, O., Gavrilova, T.: Cross-cutting support of making and explaining decisions in Intelligent tutoring systems using cognitive maps of knowledge diagnosis. In: Crossley, S., Popescu, E. (eds.) Intelligent Tutoring Systems ITS 2022. Lecture Notes in Computer Science, vol. 13284, pp. 51–64. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_5](https://doi.org/10.1007/978-3-031-09680-8_5)
  - 23. Zadeh, L.A., Aliev, R.A.: Fuzzy Logic Theory and Applications: Part I and Part II. World Scientific Publishing, Singapore (2018)
  - 24. Zouri, M., Ferworn, A.: An ontology-based approach for curriculum mapping in higher education. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0141–0147. IEEE (2021). <https://doi.org/10.1109/CCWC51732.2021.9376163>
  - 25. Zykina, A., Kaneva, O., Munko, V.: The development of approaches for obtaining automated solution on the formation of the curriculum. Mod. Inf. Technol. IT-Educ. **14**(4), 931–937 (2018). <https://doi.org/10.25559/SITITO.14.201804.931-937>



# Using the ITS Components in Improving the Q-Learning Policy for Instructional Sequencing

Amel Yessad<sup>(✉)</sup>

Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75252 Paris Cedex 05, France  
[amel.yessad@lip6.fr](mailto:amel.yessad@lip6.fr)

**Abstract.** In this paper, we aim to optimize the sequencing of learning activities using the Q-learning, a reinforcement learning method. The Q-learning agent decides each time which activity to propose to the student. The sequencing policy we propose is guided by the aim to improve efficiently the student knowledge state. Thus, the Q-learning learns a mapping of the student knowledge states to the optimal activity to perform in that state.

In this paper, we tackle two main issues in implementing the Q-learning off-policy: the combinatorial explosion of the student knowledge states and the definition of the reward function allowing to improve efficiently the student knowledge state. We rely on the student model and the domain model to answer these two challenges.

We carried out a study to evaluate the approach we propose on simulated students. We show that our approach is more efficient since it achieves better learning gain with fewer activities than a random policy or an expert based policy.

**Keywords:** Instructional Sequencing · Sequencing policy · Reinforcement learning · Q-learning · student model · domain model · Intelligent tutoring system

## 1 Introduction

In intelligent tutoring systems literature, we can distinguish between two main categories to adapt instructional sequencing in terms of the granularity of the instructional activities and time scales: task-loop (or outer loop) and step-loop (or inner loop) [1, 12]. Task-loop adaptivity involves data-driven decisions the system makes to select instructional activities for the student. Step-loop adaptivity involves data-driven decisions the system makes in response to individual actions a student takes within an instructional activity. In this research, we focus on task-loop adaptivity which consists on the sequencing of learning activities (lectures, problems, quizzes, etc.) that are most suitable according to the student characteristics [1]. It has been widely studied but remains a hard challenge

in ITS since the adaptation process depends on several dimensions of the student (knowledge state, emotional state, motivation, etc.) often hidden and not directly observable by the systems.

Several research [2, 5, 6] have shown the interest of reinforcement learning (RL) for instructional policies, as RL models can learn complex and latent relationships between instructional tasks, students actions, and knowledge outcomes. Given the importance of sequential student-teacher interactions in education, there is a real interest in applying RL to improve the state-of-the-art technology for ITS. In particular, the problem of sequencing the learning activities according to the student characteristics fits well a RL problem [2].

Unlike supervised learning and unsupervised learning, a RL agent learns from direct interaction with its environment. RL is suitable for systems where the agent needs to learn a policy on what to do in different situations to maximize a long-term utility [9]. Concretely, it is for an agent to learn how to map states to actions.

Often, instructional sequencing approaches rely on a student model that is a partially observable Markov decision process (POMDP) [10]. POMDP models the fact that the student's knowledge state is a latent variable inferred from some observation function which maps states to probability distributions of observations (student answers). The Bayesian knowledge tracing model (BKT) [4] is a POMDP which has been used to implement cognitive mastery learning in intelligent tutoring systems. Surprisingly, BKT is not considered in the RL framework, because a reward function is not explicitly specified [5]. We propose in this research to connect the BKT to a RL agent by explicitly defining a reward function.

Using the RL raises two challenging tasks: the definition of the environment state as an input of the RL agent and the definition of the reward function guiding the learning of the optimal policy. Several model-Based Policies [8] and model-free policies [3] were proposed in the state of the art. In this research, we propose to use the Q-learning [14], a model-free algorithm to learn a sequencing policy. The Q-learning seems to be a good method since we haven't any explicit model of the optimal decision. We hypothesize that the optimal policy can be reached through a trial and error strategy.

In this research, we rely on the two main components of an ITS, the student model and the domain model to improve the training process of the Q-learning algorithm. On the one hand, the student model is used to infer the student knowledge state each time the student performs a learning activity. The Q-learning learns a mapping of the student knowledge states to a learning activity to recommend in that state. In this paper, we consider only the knowledge state of the student as input to the Q-learning agent. The objective is to learn a sequencing policy that improves efficiently the knowledge state of the students. It is well known that the student's knowledge state may be a complicated function of the history of activities and estimating the parameters of the student model may require enormous amount of data [5]. But we make the assumption that the

student model, when available, allows the convergence of the Q-learning to an efficient sequencing policy.

Moreover, we propose a reward function that quantifies the student learning gain according to the student knowledge states inferred by the BKT model [4]. More precisely, we define the student's learning gain as the number of new knowledge components acquired by the student since the last recommended learning activity. The reward function we propose aims to maximize the expectation of future learning gains. On the other hand, the domain model allows us to reduce the combinatorial explosion of the number of student knowledge states.

RL models need to be trained on student historical data to converge to a good sequencing policy [7]. However, in education, policy evaluation becomes challenging because of the experience is scarce and expensive and the data are sometimes unavailable. In this paper, we answer the issue of data availability by carrying out a first study on simulated students.

We first formalise the problem of instructional sequencing by defining the main elements of the Q-learning policy and the BKT student model. Then, we present our approach and focus on the contributions of the domain experts to reduce, on the one hand the combinatorial explosion of the student knowledge states and on the other hand to define the reward function. Finally, we present the study we carried out to evaluate the proposed approach and its results.

## 2 Reinforcement Learning

The reinforcement learning (RL) problem is usually modeled as a Markov decision process (MDP) giving rise to a sequence of observed, states, actions and rewards:  $s_0, a_0, r_1, s_1, a_1, r_2, s_2, \dots, s_T$ . The  $s_t$  is the state of the environment,  $a_t$  the action taken by the agent and  $r_t$  the reward received by the agent at time-step t. A RL agent learns to decide actions in different states through trial and error strategy. RL algorithms can be mainly divided into two categories model-based and model-free. In the context of deciding learning activities, the objective is to explore the consequences of decisions on the students and we haven't any, a priori model of the optimal decision. Thus, the model-free methods are more adapted for the context of instructional sequencing. In this work, we focus on the Q-learning, a type of model-free RL algorithms. It relies on a Q-table that associates observed states with actions and a Q-function that maximizes a "reward function" received by the Q-learning agent from the environment after each action.

The Q-table is the key component of the Q-learning algorithm and it is used to calculate the expected maximum future rewards for each learning activity at each student knowledge state. The Q-table guides the Q-learning agent to choose the "optimal" learning activity for each student to maximize her learning gain. Each value of the Q-table is first initialized randomly and then learned iteratively via the Bellman recurrence equation:

$$\text{New } Q(s, a) = Q(s, a) + \alpha [ R(s, a) + \gamma \max_{a'} Q'(s', a') - Q(s, a)]$$

New Q-Value      Immediate reward       $\max_{a'} Q'(s', a')$   
 learning rate      Discount rate       $-Q(s, a)$

where  $R(s, a)$  is the immediate reward received when selecting the action  $a$  in the state  $s$ ,  $\alpha$  is the learning rate ( $0 < \alpha \leq 1$ ) and  $\gamma$  is the discount rate ( $0 < \gamma \leq 1$ ), reflecting the importance of the immediate reward comparing to the future rewards.

### 3 BKT Model

The BKT model [4] is a student modeling method. It is a two state hidden Markov model (see Fig. 1) that is used to infer the student's mastery of a knowledge component (KC). The BKT assumes that student knowledge is represented as a set of binary variables - one per KC (the skill is either mastered by the student or not). Observations in BKT are also binary: a student gets a problem (or a problem step) either right or wrong [15].

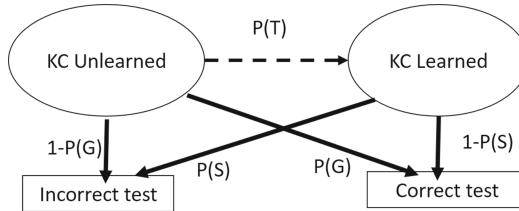
A student has an initial mastery parameter  $L_0$  and at every practice opportunity (or observation), a student who has not mastered the KC has some probability of attaining mastery (the parameter  $T$ ). If a student has mastered a KC, they will answer a question correctly unless they “slip” with some probability (the parameter  $S$ ), and if the student has not mastered the KC, they can only guess correctly with some probability (the parameter  $G$ ).

In BKT, each KC is modeled separately without considering any relationships between KCs. Other work on individualized BKT models [15] describes different approaches for defining and learning student-specific parameters. In this research, we use the Individualized BKT with student-specific  $p(L_0)$  and  $p(T)$  [4]. These specific parameters depend on the KC complexity and the student class (see 7.2).

The inference model of the BKT gives us a vector of real values (comprise between 0 and 1)—each value represents the probability of mastery for a KC. In the rest of the paper, this vector related to a student is named the knowledge state of the student.

## 4 Proposed Approach

In this work, we aim to learn a sequencing policy that assigns learning activities to students in order to maximize their learning gains. Our contributions are the following:



**Fig. 1.** The BKT model and its parameters T, S and G

1. we propose a domain-based method to reduce the risk of combinatorial explosion of the number of students' knowledge states,
2. we define a reward function that relies on the knowledge states inferred by the BKT model,
3. we evaluate the sequencing policy we propose by comparing it with two baseline approaches

## 5 Contributions of Human Expertise to the Q-Learning Function Approximation

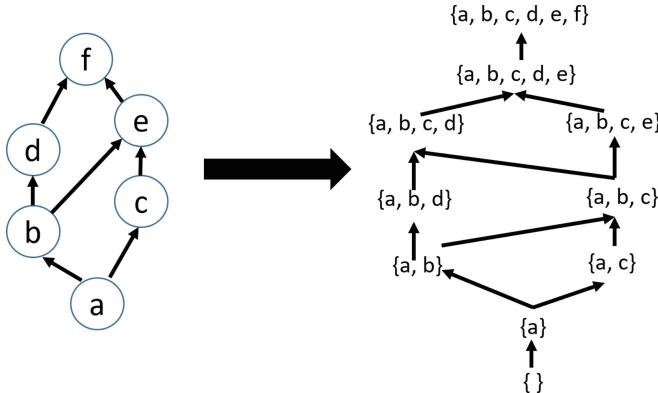
In this section, we present the contributions of domain experts to the Q-learning function approximation. In our case, we identify two major issues that can prevent the convergence of the Q-learning: the combinatorial explosion of the number of the Q-table states (usually called curse of dimensionality) and the definition of an irrelevant reward function.

On the one hand, when the state-action space is small enough, the iterative approximation of the Q-learning algorithm is proved to converge to the true Q-values [13]. However, the combinatorial explosion of the number of possible states in even a modest-size environment makes this table based implementation of Q-learning unfeasible. In our case, we can observe that there is a risk of combinatorial explosion of the number of the Q-table states when the number of the KCs increases. For example, if we have  $n$  KCs in the domain model and for each KC we consider a binary random variable representing its mastery (1 if the KC is mastered by the student, 0 otherwise), we obtain  $2^n$  knowledge states. Thus, the number of states increases considerably with the number of KC. To answer this issue, we propose an approach based on the domain model.

On the other hand, the definition of a relevant reward function is a challenged task because it guides the sequencing policy towards the objective [11], in our case, achieving efficiently the mastering the domain KCs by the students.

### 5.1 Competence Structure

To answer the combinatorial explosion of the number of possible knowledge states, we ask domain experts to identify prerequisite relationships between the



**Fig. 2.** Example of domain model (left graph) and competence structure (right graph).

KCs in order to reduce the number of possible knowledge states of the students. Two KCs  $\mathbf{X}$  and  $\mathbf{Y}$  are related by a prerequisite relationship  $\mathbf{X} \rightarrow \mathbf{Y}$  if and only if the mastery of  $\mathbf{Y}$  implies the mastery of  $\mathbf{X}$  and the non mastery of  $\mathbf{X}$  implies the non mastery of  $\mathbf{Y}$ . Thus, given prerequisite relationships between KCs, some knowledge states are improbable and thus considered as non admissible.

For instance, if we consider the domain model of the Fig. 2:

$\{a \rightarrow b, a \rightarrow c, b \rightarrow d, b \rightarrow e, c \rightarrow e, d \rightarrow f, e \rightarrow f\}$ . The competence structure is obtained by deriving all the admissible knowledge states of the domain model. We can observe that all combinations of KC are not admissible. For instance, the knowledge state  $\{a, d\}$  which means that both the KCs  $a$  and  $d$  are mastered whereas all others are not mastered, cannot be a knowledge state since the KC  $b$  has to be mastered before mastering the KC  $d$ .

Thus, only 10 knowledge states are admissible instead of  $2^6$  possible knowledge states (6 is the number of KCs).

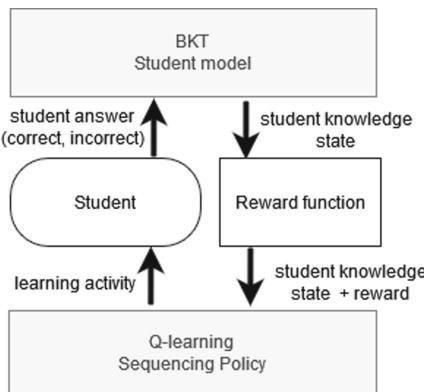
In summary, the more the KS are connected by prerequisite links, the less the number of knowledge states is.

As we consider each knowledge state as an entry of the Q-table, considering the prerequisite relationships between the KCs allows to reduce considerably the size of this table. Given  $N$  KCs, we reduce the number of states at least from  $2^N$  to  $3 * 2^{N-2}$  (we are reducing by a quarter the number of knowledge states) if we consider only one prerequisite relationship in the domain model. In the best case we have only  $N$  states if each KC is chained to one other KC with a prerequisite relationship.

## 6 Definition of the Reward

The Q-learning agent performs an iterative process until reaching the knowledge state where all the KC are mastered by the student. In each learning episode, the Q-Learning selects a learning activity to present to the student, based on the

Q-table and the  $\epsilon$ -greedy exploration/exploitation strategy. When the student answers the testing activity associated to the selected learning activity, the BKT model infers the new mastery of the KCs worked on the learning activity. Then, the new knowledge state is communicated to the Q-learning agent. Finally, the agent calculates the reward corresponding to the move from the previous knowledge state to the new knowledge state in order to update the Q-table entry according to the Bellman equation (see Fig. 3).



**Fig. 3.** Connection between the Q-learning and the BKT student model

The reward function is defined as the following:

$$R(s, a) = \sum_{i=1}^N (s'_i - s_i) \quad \text{if } s'_i > s_i$$

where  $s'$  is the new knowledge state of the student inferred by the BKT model after selecting the learning activity  $a$  and  $N$  the number of KCs. This reward function means that the reward depends on the number of new mastered KCs. More the observations allow an improving of the KCs mastery more the reward is significant. We make the choice that the cumulative rewards rely on the distance between knowledge states to quantify the learning gain of the student. We evaluate the relevance of this function in the study presented in the next section.

## 7 Evaluation

We evaluate our sequencing policy using simulated students. The evaluation consists on comparing the Q-learning (Ql) with two baseline approaches: random-based approach (Ra) and expert-based approach (Ex). The three approaches serve to design a sequencing policy and aim to achieve efficiently the mastering of all KCs. In this evaluation, we are interested in the number of activities recommended to the student before mastering all KCs.

## 7.1 Random-Based and Expert-Based Approaches

To evaluate our approach  $Ql$  we compare it to two baseline approaches for designing a sequencing policy: Random-based approach ( $Ra$ ) and Expert-based approach ( $Ex$ ).

$Ra$  is straightforward, we randomly select a learning activity by excluding the learning activities already proposed.

$Ex$  is based on the competence structure presented in Sect. 5.1. The principle of  $Ex$  is that as long as all the prerequisite KCs of a KC  $X$  have not been mastered, the learning activities about these KCs still recommended to the student. Once all the prerequisite KCs are mastered, the system starts to recommend learning activities to work on  $X$ .

We note that for these two baseline approaches, we also use the BKT model to infer the knowledge state of the simulated student. This allows us to evaluate only the impact of the reinforcement learning as a strategy of instructional sequencing.

## 7.2 Simulated Students

for the study, we generate two synthetic datasets in two different ways. On the one hand, we generate a dataset consisting of 20,000 simulated students that is used to learn the parameters of the BKT model for each KC. To do so, we first define three student classes with a probability distribution for each student to belong to one of these classes. Then, for each student class, we set the success of a student in an activity in function of four parameters inspired from the BKT parameters: *prior*, *learn*, *guess* and *slip*. The parameters *prior* and *learn* depend on the complexity of the *KC* and the student class. The parameters *guess* and *slip* depend on the type of the performed activity (quiz, open-ended question, etc.). In addition and based on the prerequisite relationships between *KCs*, we have added the assumption that a student cannot answer correctly a testing activity on  $X$  without having first mastered enough all its prerequisite KCs. In this simulation, we generate data for 20 KCs, 50 prerequisite relationships and 100 activities for each KC. All these constraints for the data generation, allow us to have a sufficiently clear characterization of the simulated data.

The evaluation of the fitted BKT model obtains an average RMSE of 0.45180 and an average AUC of 0.710705.

On the other hand, we generate in real time student's interactions that are used to learn the sequencing policy with the  $Ql$  approach. Each time an activity is performed by the simulated student, an observation (right or wrong) is sent to the Q-learning agent and based on the current knowledge state of the simulated student. This observation is then used by the BKT model to update the knowledge state of the simulated student. In this way, the Q-learning agent receives iteratively the environment state and decides an activity to the simulated student. We note that students from the three classes are generated with fixed probabilities.

### 7.3 Results

Table 1 is a summary of results for the proposed approach *Ql* (a Q-learning based policy) and the two used baseline approaches: *Ra* (a random policy) and *Ex* (an expert-based policy). For each policy, we compute after one hundred executions the mean and the SD (standard deviation) of the number of activities proposed to the students before acquiring all KCs,

**Table 1.** The mean and the SD (standard deviation) of the number of activities proposed to the students before acquiring all KCs, after one hundred simulations. The total number of simulated activities is 2000 (20 KCs and 100 activities per KC)

Sequencing policies	Mean	SD
Random policy ( <i>Ra</i> )	754	469.57
Expert-based policy ( <i>Ex</i> )	516.54	19.66
Q-learning based policy ( <i>Ql</i> )	<b>479.54</b>	49.28

The results show that the Q-learning policy is in average the most efficient even if it has a standard-deviation greater than the expert-based approach. This is explained by the random initialisations of the Q-table where some ones make longer the convergence time of the learning process.

As expected, the random policy gives the worst results, the standard-deviation between the executions shows a big variations between the executions.

The expert based approach is less efficient to the Q-learning policy but close. This result is interesting because it gives us insight into whether RL-based policies lead to better instructional sequences than relying on experts for sequencing. However, we could also explain it that the expert policy is less efficient because certain prerequisite KCs are not necessary to master some KCs and increase thus the mean number for the expert approach. But, we know that is quite hard to have a consensus on the domain model.

T-test on these results showed a significant difference between the Q-learning policy and both the Expert-based policy ( $p = 0.02$ ) and the Random policy ( $p = 0.03$ ).

Moreover, we observe clearly in the learned Q-table that the Q-learning reproduces most of the prerequisite relationships between the KC. This observation is interesting because it may provide a solution to the cold-start issue of the Q-learning when the system interacts with real students. Indeed, we can initialize the Q-table with a dataset of simulated students before using the Q-learning on real data.

## 8 Conclusion

In this research, we propose to use the Q-learning algorithm to decide the more suitable next learning activity to recommend to the students. The sequencing policy aims to improve efficiently the student knowledge state. Student knowledge states are inferred dynamically by using the BKT student model.

The proposed approach poses the problem of the combinatorial explosion of the student knowledge states. For that, a domain-based method is proposed to reduce the number of the admissible knowledge states and thus the size of the Q-table.

Based on the objective of improving efficiently the student knowledge states, a reward function is defined.

We carried out a study on simulated data to evaluate the efficiency of the proposed Q-learning based approach by comparing it with two baseline approaches: random-based approach and expert-based approach. The results show that our approach is the most efficient one but very close to the expert-based approach.

There are several research directions for future work, in particular, to provide evidence about the benefits of this approach by experimenting it on real data.

## References

1. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction based on adaptive learning technologies. In: *Handbook of Research on Learning and Instruction*, vol. 2, pp. 522–560 (2016)
2. Bassan, J., et al.: Reinforcement learning for the adaptive scheduling of educational activities. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2020)
3. Clement, B., Roy, D., Oudeyer, P.Y., Lopes, M.: Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174* (2013)
4. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**, 253–278 (1994)
5. Doroudi, S., Aleven, V., Brunskill, E.: Where's the reward? *Int. J. Artif. Intell. Educ.* **29**(4), 568–620 (2019)
6. Efremov, A., Ghosh, A., Singla, A.: Zero-shot learning of hint policy via reinforcement learning and program synthesis. In: *EDM* (2020)
7. Mandel, T., Liu, Y.E., Levine, S., Brunskill, E., Popovic, Z.: Offline policy evaluation across representations with applications to educational games. In: *AAMAS*, vol. 1077 (2014)
8. Sen, A., et al.: Machine beats human at sequencing visuals for perceptual-fluency practice. *International Educational Data Mining Society* (2018)
9. Singla, A., Rafferty, A.N., Radanovic, G., Heffernan, N.T.: Reinforcement learning for education: opportunities and challenges. *arXiv preprint arXiv:2107.08828* (2021)
10. Sondik, E.J.: The optimal control of partially observable Markov decision processes. PhD thesis, Stanford University (1971)
11. Tessler, C., Mankowitz, D.J., Mannor, S.: Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074* (2018)
12. VanLehn, K.: Regulative loops, step loops and task loops. *Int. J. Artif. Intell. Educ.* **26**, 107–112 (2016)
13. Watkins, C.J., Dayan, P.: Q-learning. *Mach. Learn.* **8**(3), 279–292 (1992)
14. Watkins, C.J.C.H.: Learning from delayed rewards (1989)
15. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized Bayesian knowledge tracing models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) *AIED 2013. LNCS (LNAI)*, vol. 7926, pp. 171–180. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39112-5\\_18](https://doi.org/10.1007/978-3-642-39112-5_18)



# Stargazing Live! Inspiring with Real Data in a Mobile Planetarium and Learning Through Conceptual Modelling

Bert Bredeweg<sup>1,2</sup>(✉) , Marco Kragten<sup>1</sup> , Joanna Holt<sup>1,3</sup>, Dennis Vaendel<sup>3</sup>,  
Joris Hanse<sup>3</sup>, and Steven Bloemen<sup>4</sup>

<sup>1</sup> Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

B.Bredeweg@hva.nl

<sup>2</sup> Faculty of Science, Informatics Institute,

University of Amsterdam, Amsterdam, The Netherlands

<sup>3</sup> Faculty of Science, API/NOVA, University of Amsterdam, Amsterdam, The Netherlands

<sup>4</sup> Department of Astrophysics/IMAPP,  
Radboud University Nijmegen, Nijmegen, The Netherlands

**Abstract.** We present the Stargazing Live! program comprising a planetarium experience and supporting lesson activities for pre-university physics education. The mobile planetarium aims to inspire and motivate learners using real telescope data during the experience. Learners then consolidate their learning by creating conceptual models in the DynaLearn software. During development of the program, content experts and stakeholders were consulted. Three conceptual model lesson activities have been created: star properties, star states and the fusion-gravity balance. The present paper evaluates the planetarium experience plus the star properties lesson activity in nine grade 11 and 12 classes across three secondary schools in the Netherlands. Learners are very positive about the planetarium experience, but they are less able to link the topics in the planetarium to the curriculum. The conceptual modelling activity improves the learners understanding of the causal relationship between the various stellar properties. Future work includes classroom testing of the star states and fusion-gravity balance lessons.

**Keywords:** Education · Physics education · Planetarium · Conceptual modelling

## 1 Introduction

Star formation, stellar properties and the underlying physical laws are fundamental topics in pre-university physics education. However, learning about stars can be challenging for learners, due to a variety of pre-instructional conceptions and learning difficulties. For example, learners often do not know that nuclear fusion provides stars with their energy, allowing them to generate light [1, 3]. In addition, they have an incomplete understanding of how stars are formed. When asked how stars differ from each other, learners often mention properties such as size or composition, but less often luminosity,

temperature or lifespan. For example, in Bailey and colleagues' study [2], only 21 of 381 learners named mass as a property that distinguishes stars. Previous research shows that traditional instruction in astrophysics courses is not always sufficiently effective and that there is a need for interventions that stimulate conceptual understanding [3].

The Stargazing Live! project [12] uses a mobile planetarium to bring semi-live real scientific astronomy data into the classroom. Planetariums have played a role in the learning of astronomical concepts since their inception [4]. They can provide a unique and enriching learning experience [14] and spark learners interest and excitement for astronomy [13, 16] and help improving retention [19]. Planetariums can be used to observe different parts of the sky at different times of the year, to view the sky from different perspectives (e.g. from the equator, southern hemisphere, poles, etc.), and to simulate the appearance of the sky at different times in history. They can also show 3D-models of all kinds of celestial objects from up close, show full-dome videos and present complex simulations. Together this allows learners to gain a better understanding of complex phenomena and their dynamics [15].

Key requirements for an effective learning experience in a planetarium is that viewers are allowed and encouraged to ask questions, participate in simulations, and engage in hands-on activities to deepen their understanding of the concepts [13, 16, 17]. The combination of planetarium and traditional classroom lessons can provide a well-rounded education experience that complements and reinforces each other [14, 15].

Key aims for the Stargazing Live! are to inspire learners and to support the teaching and learning of pre-university physics. Furthermore, shortly after the project kick-off in early 2021, evaluation with stakeholders highlighted the need to link the planetarium experience more closely to the curriculum and the need for related lesson activities [11].

In order to address these issues, the Stargazing Live! program now comprises two parts. First, learners are introduced to the idea of the changing universe and associated astronomy concepts during a live and interactive planetarium experience (Sect. 2.1). Shortly thereafter, learners further develop and consolidate their knowledge with a lesson activity during which they create and simulate a cause-and-effect model using computer-supported modelling software (Sect. 2.2). By constructing a model of a system, learners develop a deeper understanding of its underlying principles and relationships between components. This process helps to build and refine their conceptual model, providing a clearer and more comprehensive understanding of the system [8, 9]. Moreover, constructing a model requires active engagement, as learners think deeply about the information and make connections to their prior knowledge. This form of active learning, where learners are actively involved in the learning process, has been shown to be more effective than passive forms of learning [18].

This paper first discusses the development of the planetarium experience and the conceptual modelling lesson activities of the Stargazing Live! program. Next, we report the results of the evaluation study to assess (*i*) how learners experience the live and interactive planetarium, and (*ii*) how learners learn by creating the cause-effect model.

## 2 Development of the Stargazing Live! Program

### 2.1 Planetarium Experience

The planetarium experience has been developed by NOVA (Netherlands Research School for Astronomy) using a Mobile Planetarium (Fig. 1). The semi-live real data are taken from the small optical telescopes MeerLICHT ([www.meerlicht.org](http://www.meerlicht.org)) and BlackGEM ([www.blackgem.org](http://www.blackgem.org)), both operated by Radboud University in the Netherlands. MeerLICHT is stationed in South Africa and performs optical follow-up for the MeerKAT radio telescope. The BlackGEM array is in La Silla, Chile and currently comprises three telescopes. Data from the telescopes are uploaded each night, processed automatically, and made available for use within 20 min. For the lessons described in Sect. 3, the planetarium used a standard dataset and customized scripts in the Digistar 6 software.

The topic is *the changing universe* and discusses a range of transient phenomena in the night sky including (near Earth) asteroids, variable stars, (super)novae and gravitational wave events, such as kilonovae. Each topic is introduced with a discussion around a data set from the telescopes projected onto the correct region of the sky in the planetarium software. Learners are asked to identify changing features in the images and think about what they might be seeing. The various physical processes at work are then explained using custom-made 3D-visualisations and animations. Key curriculum topics for pre-university level astrophysics are also included such as an explanation of how Wien's law connects stellar surface temperature to the observed color of an object and how the luminosity of a star is related to other measurable parameters.



**Fig. 1.** During the planetarium experience, learners learn about the transient universe.

### 2.2 Conceptual Modelling Lesson Activities

Lesson activities were developed to extend the planetarium experience, focusing on key concepts in the Dutch secondary school physics curriculum. A specific request

was to focus on conceptual understanding of star formation and star properties and the associated laws (e.g. Wien's law and the Stefan-Boltzmann law).

Three lesson activities were created. The *star properties* activity focuses on learners identifying the key quantities that characterize stars and establishing the causal dependencies between those quantities. The *star states* activity follows on from the star properties assignment by adding ranges of qualitative values (known as quantity spaces) to the quantities (e.g. temperature, peak wavelength and life-span). During this assignment learners learn how stellar mass relates to the different stellar properties. The *fusion-gravity balance* model focusses on the birth of stars and how a balance emerges between the gravitational force (inwards) and the nuclear fusion force (outwards).

The lesson activities for the Stargazing Live! project use the DynaLearn software (<https://dynalearn.nl>) [6]. This software provides a qualitative vocabulary to represent conceptual models [10]. No quantitative information is used. Instead, logic-based algorithms are used to generate simulations [5].

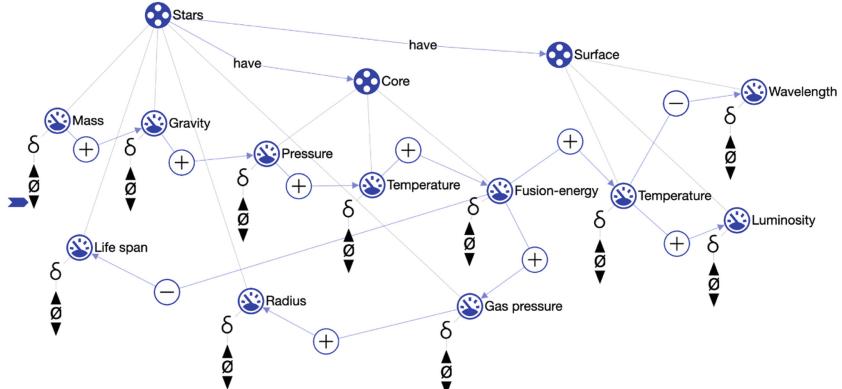
Models built in DynaLearn can be represented at four levels of complexity [6]. Higher levels use a richer vocabulary to express the system and its behavior. At each level, the software has scaffolds to support learners during their knowledge creating effort. The *norm-based feedback* pinpoints errors made by learners (solving these remains a task of the learner). The *scenario advisor* inspects the status of the model before starting a simulation and automatically highlights missing initial settings as well inconsistent settings. The *progress bar* shows how many ingredients have already been created and how many still need to be created. The working of the software is partly explained in the workbook which guides learners through the assignments, but it is also provided from within the software [7]. The following section discusses the *star properties* model which, along with the planetarium experience, is the focus of this article.

### 2.3 Star Properties

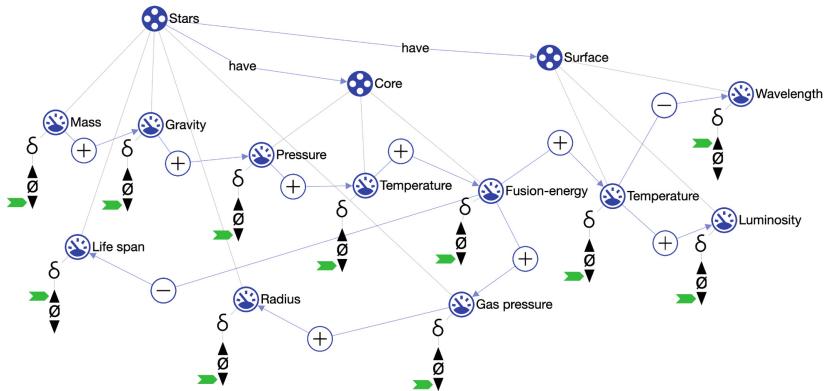
The *star properties* model is shown in Fig. 2. The model is created at level 2, which is relatively simple for learners in pre-university education. The complexity arises from the number of ingredients that need to be created and connected (26 modelling steps) combined with running various intermediate simulations with various initial values.

Entities are used to represent the objects (or parts) that together form the system. This model comprises three entities, Stars (the overarching object), the inner Core and the outer Surface. Two configurations specify that Stars *have a* Core and Stars *have a* Surface. Quantities represent the dynamic and measurable properties that characterize the stars and their behavior. Eleven quantities are defined, such as Mass, Gravity, Fusion-energy, etc. Causal dependencies specify how the change of one quantity influences the change of another quantity. They can be positive, e.g. more Mass results in more Gravity, or negative, e.g. higher Fusion-energy results in a shorter Life-span.

Initial values are required to run a simulation. Mass is the quantity at the beginning of the causal chain and thus the only quantity for which an initial change must be specified. When the Mass is set to change, the simulation shows how the remaining quantities will change (green arrows in Fig. 3). As can be seen in Fig. 3, when the Mass decreases, all intermediate quantities also decrease, and at the end of the causal chain, Radius and Luminosity also decrease while Wavelength and Life-span increase.



**Fig. 2.** Star properties model with three entities (Stars, Core & Surface), two configurations (2x have), eleven quantities (Mass, Gravity, Pressure, Temperature (of the Core), Fusion-energy, Temperature (of the Surface), Wavelength, Luminosity, Gas pressure, Radius & Life span), and ten causal dependencies (2 negative & 8 positive). Mass is set to initially decrease (blue arrow).



**Fig. 3.** Simulation result for the star properties model shown in Fig. 2. Each quantity has a  $\delta$  which can be decreasing (arrow down), steady ( $\emptyset$ ), or increasing (arrow up). Starting with Mass decreasing, the simulation shows how other quantities change depending on their proportional relationship with the preceding quantity.

**Developed with Experts.** Three astrophysics experts contributed to creating the *star properties* model, which took 6 one-hour meetings. During each meeting an improved model was presented to the experts for critical reflection. After consensus was reached, the model was reviewed by two further experts, each taking a two-hour session. In addition to clarifying terms and agreeing on the basic mechanism, most discussion concerned the notion of temperature and pressure before and after the start of nuclear fusion. Two postulates were formulated to reach consensus. Firstly, the model represents a family of stars, those in the *main sequence*, and not the specific behavior of a single star. Hence, ‘changing the Mass of a star’ refers to comparing stars of different mass in

the main sequence. Secondly, the quantities may refer to features at different moments during the life span of stars. As such, Pressure and Temperature (of the Core) refer to the features that led to the nuclear fusion starting, while Temperature and Gas pressure (of the Surface) refer to features that result from the nuclear fusion being active.

**Pilot.** A pilot version of the lesson activity *star properties* was tested by two astrophysics master students, taking about 1 h to complete. Students reflected on the activity and suggested improvements to the workbooks. The models remained unchanged.

**Model Construction.** During the lesson activity each learner starts with an empty canvas and creates the complete model. The task is divided into four steps: (*i*) the Star with Mass and Gravity, (*ii*) the Core with Pressure, Temperature and Fusion-Energy, and their relationships with the ingredients in the previous step, (*iii*) the Surface with Temperature, Wavelength and Luminosity and their relationships with the ingredients in the previous steps, and (*iv*) discovering the still missing quantities (Gas pressure, Radius and Life span), add those and relate them to the rest of the model.

## 2.4 Star States Model and Fusion-Gravity Balance Model

The *star states* model is created at level 3 and augments six key quantities from the star properties model with a quantity space, notably Mass, Temperature (of the Surface), Wavelength, Luminosity, Radius & Life-span. Each quantity space holds five values (three intervals separated by two points), and specific values correspond to quantities across the model. For instance, stars with Mass in the red dwarf region (0.5 or less times the mass of the sun), have a (Surface) Temperature of less than 4000 K, a Life span of more than  $10^{11}$  years and a peak Wavelength of more than 720 nm, etc. Learners build the quantity space for each of the key quantities and specify how these values correspond across the model. The hypothesis is that by creating this model, learners learn about typical star types.

The *fusion-gravity balance* model is created at level 4 and is the most complex. It represents how the accretion process builds the Mass of a star, how this leads to increased Gravity, which at some point ignites the fusion process, that counteracts Gravity resulting in a balance between these two forces (gravitation inwards and fusion outwards). Learners create the model stepwise from scratch interleaved with simulations.

## 3 Method

### 3.1 Planetarium Experience

The planetarium experience was provided as a live event by astronomers or university astronomy students and lasted around 50 min.

**Questionnaire.** The experience was evaluated using a questionnaire including Likert-scale statements and open-ended questions. The questions focused on the student experience including questions such as ‘*What is your first reaction to seeing the planetarium?*

(answer options on a 7-point Likert-scale; 1 = negative and 7 = positive). We also wanted to know if the key message that the sky is ever changing came clearly across (*'I did not know that so many things changed in the night sky'*: answer options on a 5-point Likert-scale; 1 = strongly disagree and 5 = strongly agree). Learners were also asked to comment on the relevance of the telescope data (*'The live data added value to the planetarium lesson'*; answers using the 5-point Likert-scale described above). Both 5-point Likert scale statements and open questions also asked about the link between the planetarium experience and curriculum topics.

### 3.2 Conceptual Modelling Lesson

**Teachers.** During a 90-min teacher-training, physics teachers from the participating schools were informed about the lesson activities and the evaluation study. Teachers agreed to reserve 90 min for this lesson, including a pre- and post-test.

**Pre-test and Post-test.** The tests consist of two parts. The first part of the test measures the extent to which learners understand the qualitative vocabulary of conceptual modelling. This part of the test has six multiple-choice and -response items (Table 1). Simple contexts were used to avoid the effect of prior domain knowledge during this part of the test. For example, the fourth item required learners to choose the representation that correctly describes the effect of ocean pollution on coral reef mortality, i.e., this system has two entities (ocean and coral reef), two quantities (pollution and mortality) and one causal relationship (pollution causes mortality). In the fifth item learners were given a causal chain of three quantities. The first quantity is positively proportionally related (+) to the second quantity which itself is negatively proportionally related (−) to the third quantity. Learners were required to describe the behavior of the third quantity when the first quantity increases or remains stable.

The second part of the test (Table 1) measures learners' understanding of the causal relationships between quantities of stars in the main sequence. Firstly, learners were required to answer an open question: *'Explain, step-by-step, the effect of a high stellar mass on the luminosity of a star'*. Next, learners were required to answer a multiple-choice item regarding the effect of high stellar mass on a star's fusion, life-span, radius, luminosity and wavelength. Learners were not able to return to the open question after answering the multiple-choice question to avoid answers being adjusted.

**Data-Analysis.** The score of the multiple-response items is calculated as correct minus incorrect and then normalized to a number between 0 and 1. The multiple-choice items are scored as 1 if correct. The first part of the test has a maximum score of 6 (i.e., six items) and the second part of the test has a maximum score of 5.

The data were first pre-analyzed to check for potential differences between schools and classes, however the pre-analysis showed no significant differences. We therefore made no further distinction between participating groups and analyzed the data as one data set. Firstly, a paired t-test was used to test if there is a significant difference between the pre-test and post-test scores per item. Secondly, the total score for both parts of the test was calculated and analyzed using a paired t-test to test if there is a difference between total pre-test and post-test scores. The open question of part 2 was analyzed for the number of causal relationships that learners describe between quantities.

**Table 1.** Items of the pre-test and post-test.

Part 1. Qualitative vocabulary	Part 2. Causal relationships in a star
What is an entity (MR)?	Effect of high mass on luminosity of star (O)
What is a quantity (MR)?	Stars with high mass have <i>high/low</i> (MC):
What is a causal relationship (MR)?	<ul style="list-style-type: none"> <li>o Fusion-energy (Fe)</li> </ul>
What is a system (MC)?	<ul style="list-style-type: none"> <li>o Life span (Ls)</li> </ul>
Propagation in a causal chain (MR)	<ul style="list-style-type: none"> <li>o Radius (R)</li> </ul>
Simulation results (MC)	<ul style="list-style-type: none"> <li>o Luminosity (L)</li> <li>o Wavelength (W)</li> </ul>

Note. MR = multiple-response, MC = multiple-choice, O = open.

## 4 Results

One hundred and fifty-two learners from three secondary schools (across the Netherlands) participated in the study. From the first school three grade-12 classes participated ( $n = 56$ ), from the second school one grade-12 class participated ( $n = 15$ ) and from the third school two grade-12 and three grade-11 classes ( $n = 81$ ) participated. Learners had no previous experience with learning by constructing qualitative representations.

### 4.1 Planetarium Experience

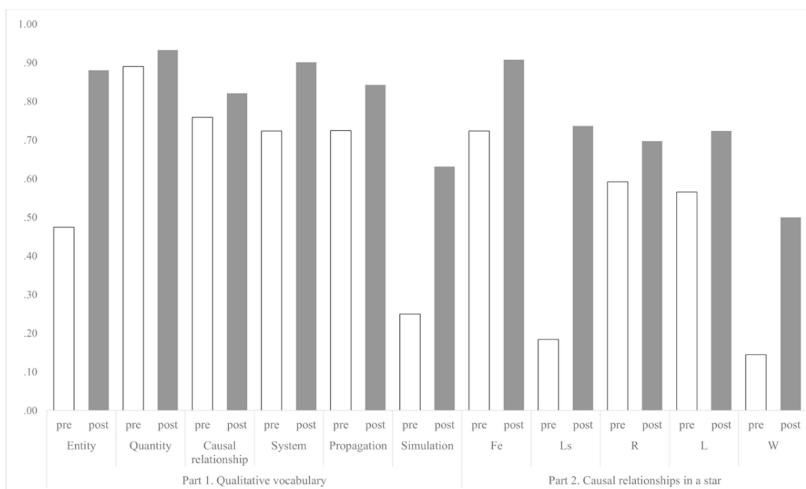
The overall reaction to the planetarium was positive, scoring on average 5.8 ( $SD = .7$ ) and 6.6 ( $SD = .6$ ) in the digital survey. The question '*Did the physics content in the planetarium match with what you had learned in your physics lessons?*' returned an average of 2.9 ( $SD = 1.3$ ). So, whilst the planetarium experience aimed to make clear links with the curriculum, most learners recognized few curriculum topics in the planetarium lessons. However, learners did not find this a problem and the majority agreed that the planetarium experience taught them new things ( $M = 4.2$ ,  $SD = .8$ ). Topics which many learners easily recognized included both Wien's Law and Stefan Boltzmann's Law (open question), two key curriculum concepts which were included after the stakeholder evaluation [11]. The learners were very positive about the use of real astronomical data in the classroom ( $M = 4.6$ ,  $SD = .6$ ). The majority of learners also agreed that they were unaware that the universe contained so many transient phenomena, a key messages in the planetarium experience ( $M = 3.8$ ,  $SD = 1.0$ ).

### 4.2 Conceptual Modelling Lesson

Figure 4 shows the pre-test and post-test scores per item of part 1 (qualitative vocabulary) and part 2 (causal relationships between quantities of stars). Paired t-tests showed that there is a significant difference between the score on all pre-test and post-test items ( $p < .05$ ). In part 1, the first ('*What is an entity?*') and sixth items ('*Simulation results.*') showed the highest increase in score from pre-test to post-test (.41 and .38 points,

respectively). Regarding part 2, the highest increase was found for the effect of mass on life span ( $L_s$ , .55 points increase) and on the peak wavelength ( $W$ , .36 points increase).

There is a significant difference in total pre-test ( $M = 3.82, SD = .96$ ) and post-test score ( $M = 5.01, SD = .86$ ) of part 1 (paired t-test:  $t = 13.20, df = 151, p = < .001$ , Cohens'  $d = 1.07$ ). We can therefore conclude that the lesson significantly increases the understanding of the qualitative vocabulary associated with concept modelling. We also found a significant difference in total pre-test ( $M = 3.61, SD = 1.12$ ) and post-test score ( $M = 4.28, SD = 1.03$ ) related to the understanding of the causal relationships between the properties of stars ( $t = 6.26, df = 151, p = < .001$ , Cohens'  $d = .51$ ).



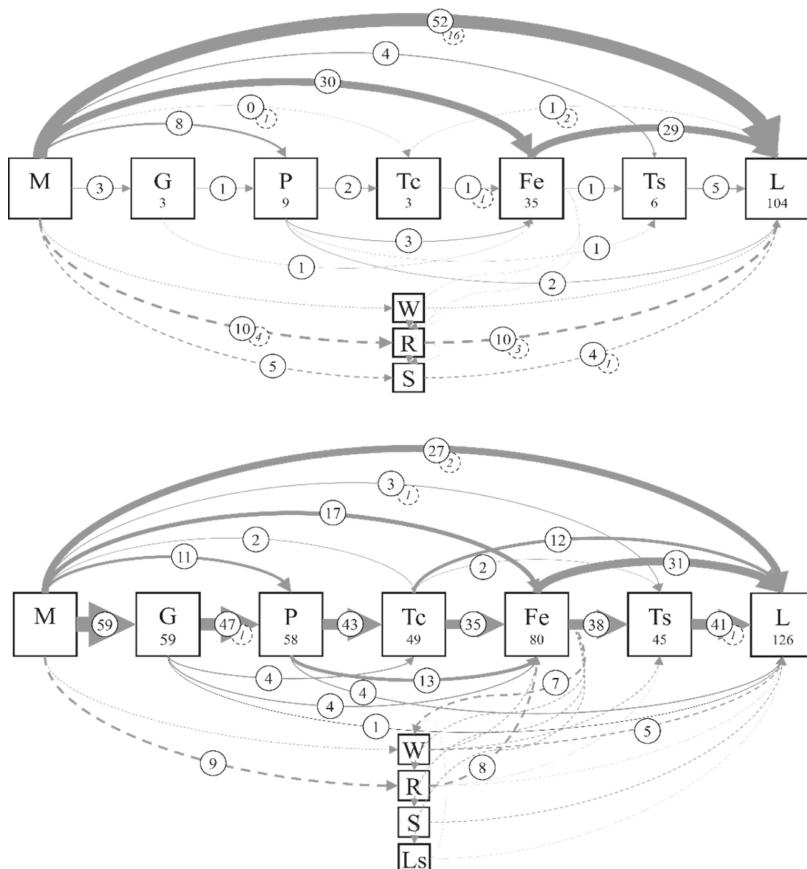
**Fig. 4.** Pre-test and post-test scores per item.  $Fe$  = Fusion-energy,  $Ls$  = Life span,  $R$  = Radius,  $L$  = Luminosity,  $W$  = Wavelength.

Figure 5 shows the number of causal relationships that learners describe between quantities in their answer to the open question. In the pre-test, 52 learners describe a direct causal relationship between mass ( $M$ ) and luminosity ( $L$ ), e.g., ‘Stars with a high mass have a high luminosity...’. Sixteen learners describe this relationship incorrectly, i.e., they describe a negative proportional relationship. Thirty learners describe the causal relationship between mass ( $M$ ) and fusion-energy ( $Fe$ ) and 29 learners describe the relationship between fusion-energy ( $Fe$ ) and luminosity ( $L$ ). Eight learners describe the relationship between mass ( $M$ ) and pressure ( $P$ ), and one student describes the relationship between gravity ( $G$ ) and pressure ( $P$ ). Learners mentioning pressure in their answer also continue to describe the relationship between pressure with temperature of the core ( $T_c$ ,  $n = 2$ ), fusion energy ( $Fe$ ,  $n = 2$ ) and luminosity ( $L$ ,  $n = 3$ ). A small number of learners describe causal relationships other than those mentioned.

Some learners also describe relationships between quantities that are not part of the correct causal path. Ten learners describe a causal relationship between mass ( $M$ ) and radius ( $R$ ) and then state that radius ( $R$ ) causes luminosity ( $L$ ). Five learners describe a causal relationship between surface area ( $S$ ) and luminosity ( $L$ ).

In total, 104 learners correctly describe the relationship between mass and luminosity. Note that the latter number also includes learners that include quantities in their answer that are not part of the correct causal path.

In the post-test there is an increase in the number of learners that describe causal relationships between quantities that are part of the correct causal path. 59 learners describe the causal relationship between mass (M) and gravity (G), 47 between gravity (G) and pressure (P), 43 between pressure (P) and temperature of the core (Tc), 35 between temperature of the core (Tc) and fusion-energy (Fe), 38 between temperature of the core (Tc) and fusion-energy (Fe), 38 between fusion-energy (Fe) and temperature surface (Ts), 5 between temperature surface (Ts) and luminosity (L), and 104 between mass (M) and luminosity (L).



**Fig. 5.** Causal relationships that learners describe between quantities in the open question in the pre-test (upper figure) and post-test (lower figure). Quantities G, P, Tc, Fe and Ts are part of the correct and complete causal path and are aligned horizontally between mass (M) and luminosity (L). Dashed lines show the causal relationships that are not part of the correct path (W, R, S and Ls). To maintain readability, only numbers greater than 3 are shown. The number of learners that describe causal relationships of the wrong type (positive instead of negative or vice versa) are indicated by a dashed circle bottom right of the number of correct causal relationships. M = Mass, G = Gravity, P = Pressure, Tc = Temperature core, Fe = Fusion energy, Ts = Temperature surface, L = Luminosity, W = Wavelength, R = Radius, S = Surface area, Ls = Life span.

(Fe) and temperature of the surface (Ts) and 41 between temperature of the surface (Ts) and luminosity (L).

Twenty-seven learners describe a direct causal relationship between mass (M) and luminosity (L), and two learners describe this relationship incorrectly with a negative proportional relationship. Seventeen learners describe the causal relationship between mass (M) and fusion-energy (Fe) and 31 learners that describe the relationship between fusion-energy (Fe) and luminosity (L).

Several learners include quantities in their answer that are not part of the correct causal path. For example, eight learners describe a relationship between radius (R) and fusion-energy (Fe), e.g., ‘If mass is higher than the radius increases and then the fusion-energy will increase...’. Finally, 126 learners correctly describe the relationship between mass and luminosity.

## 5 Discussion and Conclusion

In this contribution we report on the Stargazing Live! program. This program has two parts: a planetarium experience and a conceptual modelling lesson activity. The planetarium experience is given live in a mobile planetarium that projects semi-live real data. The goal is to inspire learners and make them aware of the continuously changing universe. It touches on curriculum topics which also return in more depth in the modelling lesson activity. The conceptual modelling lesson activity takes place in the classroom. Learners individually build a conceptual model using interactive modelling software.

Three assignments and corresponding models have been developed to extend the planetarium experience with lesson activities that relate to the Dutch secondary school physics curriculum. From these assignments, the cause-effect model on *star properties* was selected by the participating teachers to be used and evaluated in their classes.

The results from the evaluation study show that learners are very positive regarding the planetarium experience. They find it engaging and it creates an awareness of the many transient phenomena in the universe. However, learners are less able to see how the provided information relates to the topics in the school curriculum. The latter supports the need for the additional lesson activities that we have created to further support learners’ development and consolidate their knowledge concerning star formation, star properties and the underlying physical concepts and laws.

Results obtained during the conceptual modelling lesson show that there is a positive effect of conceptual modelling on learners’ understanding of the causal relationships between quantities of stars in the main sequence and the qualitative vocabulary. In the pre-test, 104 learners correctly described the relationship between mass and luminosity, while in the post-test, this number increased to 126. There were also improvements in the number of learners correctly describing other causal relationships, such as between mass and gravity, pressure and temperature of the core, and fusion-energy and temperature of the surface. However, some learners still described relationships incorrectly or included quantities that were not part of the correct causal path. Notable is that none of the answers (i.e., both in the pre- and post-test) contain non-scientific descriptions [2].

As future research we plan to evaluate the assignments on *star states* and on *fusion-gravity balance*, in schools participating in the project Denker (<https://denker.nu>). Learners in these schools have multiple years of experience with conceptual modeling using

interactive qualitative representations and will more easily cope with the higher-level model representations (level 3 and 4). Furthermore, we intend to expand the set of conceptual modelling lessons to include other phenomena discussed in the planetarium lesson. For instance, we are currently developing conceptual modeling lessons relating to circular and elliptical orbits of celestial bodies.

## References

1. Agan, L.: Stellar ideas: Exploring students' understanding of stars. *Astro. Educ. Rev.* **3**(1), 77–97 (2004)
2. Bailey, J.M., Prather, E.E., Johnson, B., Slater, T.F.: College students' pre-instructional ideas about stars and star formation. *Astron. Educ. Rev.* **8**(1), 010110–010111 (2009)
3. Bailey, J.M., Johnson, B., Prather, E.E., Slater, T.F.: Development and validation of the star properties concept inventory. *Int. J. Sci. Educ.* **34**(14), 2257–2286 (2012)
4. Branley, F.M.: Planetariums – an evaluation. *Sci. Teach.* **31**(6), 18–19 (1964)
5. Bredeweg, B., Linnebank, F., Bouwer, A., Liem, J.: Garp3 – workbench for qualitative modelling and simulation. *Eco. Inform.* **4**(5–6), 263–281 (2009)
6. Bredeweg, B., et al.: Dynalearn - an intelligent learning environment for learning conceptual knowledge. *AI Mag.* **34**(4), 46–65 (2013)
7. Bredeweg, B., Kragten, M., Spitz, L.: Qualitative representations for systems thinking in secondary education. In: Proceedings of the 34th International QR Workshop. Montreal, Canada (2021)
8. Doyle, J.K., Ford, D.N.: Mental models concepts for system dynamics research. *Syst. Dyn. Rev.* **14**(1), 3–29 (1998)
9. Greca, I.M., Moreira, M.A.: Mental, physical, and mathematical models in the teaching and learning of physics. *Sci. Educ.* **86**(1), 106–121 (2002)
10. Forbus, K.D.: Qualitative Representations. How People Reason and Learn About the Continuous World. The MIT Press, Cambridge, MA (2018)
11. Holt, J., Hanse, J., Baan, M., Groot, P., Bloemen, S.: Co-creation 3.0: taking the development of astronomy education resources to the next level in the project Star-gazing Live! In: Proceedings of the 3rd Shaw IAU Workshop on Astronomy for Education International Astronomical Union (2021)
12. Holt, J., et al.: Stargazing Live! Inspiring with semi-live astronomy data; teaching curriculum topics using smart education tools. In: Proceedings of the 4th Shaw IAU Workshop on Astronomy for Educational International Astronomical Union (2022)
13. Plummer, J.D., Small, K.J.: Informal science educators' pedagogical choices and goals for learners: The case of planetarium professionals. *Astron. Educ. Rev.* **12**(1), 1–16 (2013)
14. Plummer, J.D., Kocareli, A., Slagle, C.: Learning to explain astronomy across moving frames of reference: exploring the role of classroom and planetarium-based instructional contexts. *Int. J. Sci. Educ.* **36**(7), 1083–1106 (2014)
15. Plummer, J.D., Small, K.J.: Using a planetarium fieldtrip to engage young children in three-dimensional learning through representations, patterns, and lunar phenomena. *Int. J. Sci. Edu.* **8**(3), 193–212 (2018)
16. Schmoll, S.E.: Toward a framework for integrating planetarium and classroom learning. Doctoral Dissertation (2013)
17. Schultz, S.K., Slater, T.F.: Use of formative assessment-based active learning by astronomy educators teaching in live planetarium learning environments. *J. Astron. Earth Sci. Educ.* **8**(1), 27–38 (2021)

18. Tippett, C.D.: What recent research on diagrams suggests about learning with rather than learning from visual representations in science. *Int. J. Sci. Educ.* **38**(5), 725–746 (2016)
19. Zimmerman, L., Spillane, S., Reiff, P., Sumners, C.: Comparison of student learning about space in immersive and computer environments. *J. Rev. Astron. Educ. Outreach* **V1**, A5–A20 (2014)



# A GPT-Based Vocabulary Tutor

George Zografas<sup>(✉)</sup> and Lefteris Moussiades

International Hellenic University, 65404 Kavala, Greece  
`{geozozra,lmous}@cs.ihu.gr`

**Abstract.** Students of a foreign language need help to study vocabulary usually housed in foreign language textbooks. This need can be supported by an artificial vocabulary learning assistant, i.e., software that can ask relevant questions and reliably evaluate the answers. In this work, we present an assistant based on a combination of traditional dialogue generation technology with the latest technology of Generative Pre-trained Transformers. Moreover, we test our software and get admirable results that encourage us to continue our research in this direction.

**Keywords:** Vocabulary tutor · GPT-3 · dialogue system · natural language interface

## 1 Introduction

In foreign language learning books, vocabulary sections are often included, i.e., areas that have the structure of a dictionary. More specifically, these sections include a series of terms. For each term, they give its definition, usage examples, explanation, translation in the student's language and other information, such as what part of speech the term is, etc.

When studying such a vocabulary, students often need someone to help them. Typically, the assistant's role is to hold the book and ask questions of the type: Give me the definition of the term X or an example of its use. After each question, the assistant assesses the student's answers.

In this work, we present such an assistant, which is nevertheless artificial and performs its tasks with admirable reliability. The assistant we propose is based on various technologies with the GPT-3 language model prominent [1], and we call it GPT Vocabulary Tutor or GPT-VT for short.

GPT-VT communicates with the learner through natural language and directs the dialogue towards learning specific vocabulary, which can be defined by the learner himself or the learner's human teacher. GPT-VT also follows basic pedagogical principles.

Our contribution consists of the following:

1. In the construction and delivery of a novel artificial vocabulary-learning assistant. The GPT-VT differs substantially from any previous construct intended to aid vocabulary learning.

2. We present an architecture that combines the traditional rule-based and pattern-matching technology of natural dialogues with modern technology based on the sophisticated transformer neural network [2] and, more specifically, on the GPT-3 language model. We manage with this architecture to limit GPT-3, which is an open-domain dialogue system, exclusively to vocabulary learning.
3. We study the effectiveness and weaknesses of GPT-VT, which is at the same time a study of the efficacy and deficiencies of GPT-3 underline technology about the specific tasks of evaluating definitions of terms and examples of their use.

## 2 Intelligent Vocabulary Tutors

Intelligent tutoring systems are pretty widespread. You can find a lot of information about them in reviews [3, 4]. Some of them deal with helping students learn vocabulary.

Paper [5] presents a small French vocabulary tutor for the World Wide Web, which helps students learn new French words and their pronunciation through words, pictures, and sounds. This work does not support a natural language interface.

The work in [6] presents the Web-based Intelligent Learning Environment (WILE) framework as a solution. WILE includes four essential components: a communication module, a pedagogical module, a student model and an expert module, deployed over a client-server architecture. The article also mentions the emergence of mobile learning (mLearning) as an alternative to eLearning, and its advantages and constraints, such as the display of large content. It also does not support a natural language interface.

Paper [7] proposes a mobile-based vocabulary tutor and examines its use. The presented tutor allows learners to complete vocabulary activities through mobile phones or computers and tracks their progress. The system creates a profile of each learner based on the vocabulary they have difficulty with and provides more practice for those words. The study analysed the learners' usage patterns and surveyed them at the end of the project to gather their perspectives on the system. It also does not support a natural language interface.

A web-based system aiming at learning vocabulary through personalised word association games is presented by [8]. This work focuses on personalising the learning process and game-based learning to make language learning more attractive to students. It does not support a natural language interface.

The study [9] examines the impact of sociocultural factors on using a web-based tool for learning English vocabulary.

Other works relative to vocabulary acquisition include:

STI-DICO [10] is an Intelligent Tutoring System (ITS) that incorporates a Learning Management System (LMS) as its learning interface. It aims to help future French primary school teachers acquire the knowledge and skills to use the French dictionary. It uses authentic learning scenarios to allow learners to develop the skills and knowledge needed for the classroom.

A web-based training platform for those who struggle with acquiring German orthography is presented by [11]. On this platform, students can write, publish and correct texts supported by a feature that gives feedback automatically and is called the intelligent dictionary.

A systematic review of AI-based language learning tools is presented in [12] for more information.

Finally, specifications for a vocabulary learning assistant based on a natural language interface are given in [13], while in [14], vocabulary structure recognition is attempted to be utilised in such an artificial teacher.

### 3 Architecture, Functionality and Technology

None of the efforts mentioned above yielded a satisfactory solution to the following question. How can we reliably evaluate the student's response that will be given in natural language?

Then, transformer neural networks appeared, and GPT-3, a transformer-based language model, solved the question. GPT-3 (Generative Pre-trained Transformer 3) is a state-of-the-art language generation model developed by OpenAI. It is trained on a massive dataset of text data and can generate human-like text on a wide range of topics. It can perform various natural language processing tasks, such as language translation, text summarisation, question-answering and text generation. It is considered one of the most advanced language models currently available, with the ability to generate highly coherent and fluent text, and can be fine-tuned for specific tasks. GPT-3 is also notable for its ability to generate text independently without being prompted with a particular task, making it a powerful tool for chatbots implementing natural language application interfaces.

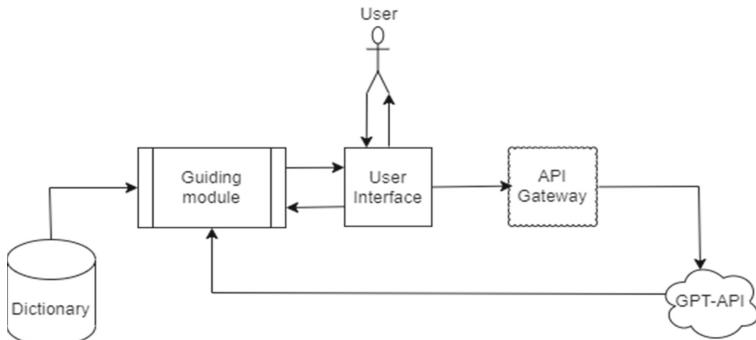
However, GPT-3 cannot be used directly to play the role of a vocabulary assistant. This is because GPT-3 is an open-domain system, i.e. it can enter into dialogues on various issues. This functionality does not serve the GPT-VT, which requires exchanges to be directed solely at learning a specific vocabulary at a time. Therefore, the solution we propose is the cooperation of Chatscript with GPT-3.

ChatScript is a free, open-source platform for creating natural language processing (NLP) tasks [15]. It uses a combination of pattern matching and rule-based processing to understand the context and respond accordingly. The platform features a pattern-matching system that allows it to understand a wide range of inputs, a scripting language for creating custom rules, and a built-in dialogue management system that enables the creation of complex multi-turn conversations.

In Fig. 1, we present GPT-VT's abstract architecture. Then we analyse each component.

- **Dictionary:** The Dictionary is a data pool where vocabulary entries are stored as JSON records. Currently, two types of such records are supported: definition and usage. The definition type causes the guiding module to ask the user for an explanation of the term, while the usage type questions for a usage example. An example of each type follows.

- {term: "chair", type:"definition", response:"a piece of furniture with a raised surface supported by legs, commonly used to seat a single person"}
- {term: "chair", type:"usage", response:"he pulled out a chair from the dining table and sat down for dinner"}



**Fig. 1.** GPT-VT's abstract architecture

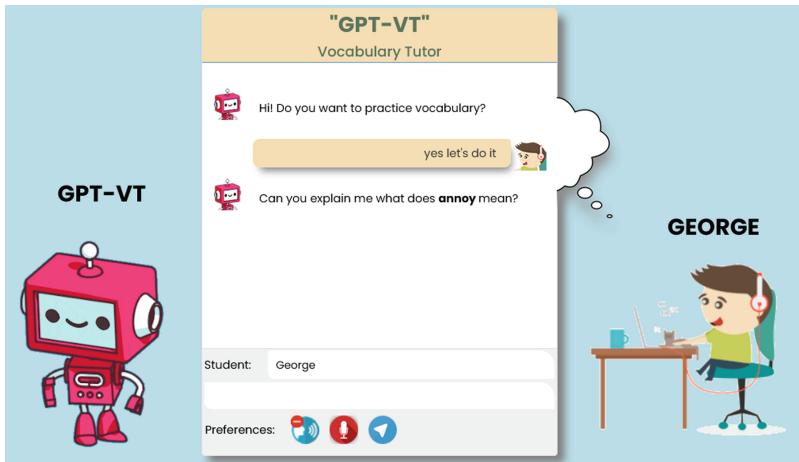
This arrangement enables the human teacher to precisely determine the type of questions to be asked of the student and the terms these questions will include by simply loading an appropriate JSON file. However, a facility is also provided to load definition entries automatically from the Cambridge Dictionary<sup>1</sup>.

- **Guiding module:** Based on an entry from the Dictionary, the guiding module constructs the question for the user. Within the guiding module, Chatscript operates, which combines data from the Dictionary with its pattern-matching rules to differentiate the questions to the user making the dialogue more human-like. So, we have two main types of questions, one on the definition and one on usage. However, its role is not limited to query formulation based on the Dictionary. Its remaining functions are described below when describing the functionality of the remaining components.
- **User Interface:** The user interface design is inspired by the interface of bot LIA [16]. Figure 2 provides a view of the user interface design. It features a student's sketch on the right side communicating with a computer, a moving robot on the left side, and a dialogue box between them. The robot is labelled as the GPT-VT, and when the student gives their name, it appears as a tag on the student's sketch. The purpose is to identify the student with the corresponding image and create the perception that the student is talking to a robot. The interface supports text-based dialogues but also includes ASR (Automatic Speech Recognition) and TTS (Text-to-Speech) interfaces that can be controlled by voice and microphone icons. The microphone can be activated automatically when the GPT-VT expects a student's response, and it will be maintained for a short time or until the student completes their phrase. If the student does not speak, the microphone can be turned on by clicking the microphone icon. This practice aims to achieve seamless voice communication between the student and the GPT-VT.

Its operation consists of the following: It sends the question to the user and receives the answer. Often, in studying a particular vocabulary, students know the answer as it is in the response field. If the student's answer is identical to JSON's response field, the

<sup>1</sup> <https://dictionary.cambridge.org/>.

User Interface updates the Guiding module backwards. Thus, GPT -VT saves time; otherwise, it passes the question along with the user's response to the API gateway.



**Fig. 2.** GTP-VT user interface

- **API gateway:** The API gateway performs the following tasks. Configures the phrase to be sent to GPT-3. For example, suppose the definition of the term chair has been requested, and the user answered, ‘A chair is a piece of furniture with a raised surface supported by legs, commonly used to seat a single person’. In that case, the sent sentence is formatted as follows: Is the phrase “A chair is a piece of furniture with a raised surface supported by legs, commonly used to seat a single person” a correct definition for ‘chair’? Start your answer with yes or no and include a proper explanation. In general, phrases send to GPT-3 formulate according to the two types of JSON entries:
  - Is the phrase [“user answer”] a correct definition for [“JSON’s term”]? Start your answer with yes or no and include a proper explanation.
  - Is the phrase [“user answer”] a representative example of [“JSON’s term”]? Start your answer with yes or no and include a good usage example.
 API gateway also includes a PHP handler acting as an interface between the user interface and the GPT-3 API.
- **GPT-API:** The GPT-API sends the appropriate phrase to GPT neural network and receives GPT’s evaluation, which it forwards to the Guiding module.

The Guiding module analyses the assessment to understand whether the student's answer is correct. This analysis for the definition and usage types is based on the fact that the GPT evaluation starts with Yes or No.

If the user's answer is identified to be correct, the Guiding module proceeds to the next JSON record. Suppose the user's response is recognised as incorrect. In that case, the Guiding module informs the student of the GPT's evaluation and asks if they would like a second attempt with the same term to facilitate the student to consolidate the term.

The Guiding module also recognises students based on their name and IP address. These credentials were chosen so the system can be accessed by young students who may be troubled by the login process. However, this is an option that can easily be changed.

In addition to these functions, the Guiding module keeps the history of GPT-VT interaction with the student. Therefore, it enables the system to continue the study of a vocabulary even in different sessions while maintaining information that can be used to evaluate the student. Besides, when a cycle with the dictionary terms is completed, the Guiding module repeats questions that only address the words that were not answered correctly. The vocabulary study is completed when the student has answered correctly at least once for each Dictionary term.

The technology that supports the described architecture includes Chatscript for the Guiding module, HTML, CSS, Javascript and Ajax for the User Interface, PHP for the API gateway and Python for the GPT3 API.

It follows an example dialogue between GPT-VT and a human reviewer.

```
GPT-VT : Hello George! Do you want to practice vocabulary?
User   : yes
GPT-VT : Can you explain to me what chair mean?
User   : a piece of furniture made of steel with two legs
GPT-VT : No, the correct definition for the term chair is a piece of furniture for one
         person to sit on, typically with four legs and a back. Your example was not
         quite accurate. Do you want to try again with the same lemma?
User   : yes
GPT-VT : Okay, let's try again. What's your definition?
User   : a furniture you can sit on
GPT-VT : Yes, a chair is a piece of furniture designed for a person to sit on. Ok, you
         are good so let's move to the next one! Give me a usage example of the term
         chair.
User   : I got tired of standing for so many hours. I urgently need a chair.
GPT-VT : Yes, this is a correct usage example for the term "chair". Another correct
         usage example is: "I went to the store to buy a chair for my office." Ok, so
         let's move to the next one! What does record mean?
```

## 4 Pedagogical Principles

The GPT-VT was designed from the author's experience as their children often give them a foreign language book and ask for help practising vocabulary. However, it follows some basic pedagogical principles:

- **Active learning** [17]: The use of GPT-VT demands interaction with the student. Furthermore, it supports self-directed learning allowing the student to load vocabulary of their liking using an appropriate JSON file.

- **Instructional scaffolding:** It supports the essential features of scaffolding [18]. It supports the interaction between the learner and the expert, where the expert is the GPT-VT. It allows the human tutor to set the appropriate vocabulary such that learning takes place in the learner’s zone of proximal development [19]. As it keeps the history of the interactions, it can easily be extended to adapt to the appropriate vocabulary automatically.
- **Content differentiation** is also supported by selecting the appropriate JSON file, and feedback is immediate.

## 5 Experimentation

To evaluate the performance of GPT-VT, we used 35 terms from the Cambridge dictionary. For 18 of them, we checked whether GPT-VT correctly evaluates their definition. For each of the 18 terms, we submitted a correct definition from the Cambridge dictionary and an incorrect definition according to a human reviewer. Because GPT-VT results are not deterministic, we repeated this test 50 times. Therefore, to check the definition evaluation, we asked a total of  $18 \times 2 \times 50 = 1800$  questions to GPT-VT.

We did precisely the same for testing the usage examples. So here we were asking a total of  $17 \times 2 \times 50 = 1700$  questions.

The total number of definition and usage example questions is 3500.

For this test, we have used text-davinci-003, the most advanced model of the GPT-3 series (except ChatGPT) [20], setting the maximum length to 256 and temperature to 0.5. The maximum length parameter represents the maximum number of tokens to generate. Its highest accepted value is 4000. Temperature controls the degree of randomness of the generated responses. It takes values from 0 to 1. If the temperature equals 0, the model becomes deterministic [21].

In the following, we present the aggregated results and then analyse them with particular emphasis on GPT-VT errors.

### 5.1 Results

Table 1 presents the aggregated results.

**Table 1.** Aggregated results

		Prediction	
		Positive	Negative
User Response	Correct	TP: 1648	FN: 102
	False	FP: 337	TN: 1413

Based on Table 1, we calculated the measures presented in Table 2.

The high sensitivity value indicates the GPT-VT’s ability to correctly identify correct user responses.

**Table 2.** Performance Measures

Measure	Formula	Value
Sensitivity	$TPR = \frac{TP}{(TP+FN)}$	<b>0.9417</b>
Specificity	$SPC = \frac{TN}{(FP+TN)}$	<b>0.8385</b>
Precision	$PPV = \frac{TP}{(TP+FP)}$	<b>0.8302</b>
Accuracy	$ACC = \frac{(TP+TN)}{(P+N)}$	<b>0.8856</b>
F1 Score	$F1 = \frac{2TP}{(2TP+FP+FN)}$	<b>0.8825</b>
Matthews Correlation Coefficient	$\frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$	<b>0.7777</b>

The specificity that reflects the GPT-VT’s ability to identify incorrect user responses correctly is also relatively high. However, it is lower than sensitivity, which means that GPT-VT is better at identifying correct than wrong user answers.

Precision reflects how many of the answers detected as correct are actually correct.

Accuracy is considered a measure of the overall performance of a model.

F1 Score measures a model’s performance and balances precision and recall. It is calculated as the harmonic mean of precision and recall.

All the above measures take values from 0 to 1. Matthews Correlation Coefficient differs; its lowest value is  $-1$ , and its highest is  $1$ . It computes the correlation between observed and predicted binary classifications and is considered a robust metric of binary classification performance.

The general picture from the measurements in Table 2 is that GPT-VT performs significantly well. However, the deviation from optimal performance raises questions analysed in Sect. 5.2.

## 5.2 Analysis of the Results

Analysing the results, we identified three different types of GPT-VT errors. The first type concerns the definitions of the terms. In this case, they were evaluated as incorrect user responses, whereas they were correct. Both terms for which we had such a result were homonyms. So GPT-VT compared the definition we gave it with another meaning. We call this type of error the homonym error. An example of a homonym error is shown next.

- Query: Is the phrase “the power and ability to be physically and mentally active “ a correct definition for “energy”? Start your answer with yes or no and include a proper explanation.
- Response: No, this is not a correct definition for “energy.” Energy is defined as the ability to do work or cause change. It is the capacity to cause motion or create change.

This phrase describes the ability to be physically and mentally active, but does not address the capacity to cause motion or create change.

The remaining two error types concern the usage examples, and we call them conceptual and metaphorical errors. Conceptual errors refer to phrases that are syntactically correct but could not be considered appropriate examples of the use of the term. For example,

- Query: Is the phrase “What is your car’s opinion?” a representative example of “opinion”? Start your answer with yes or no and include a good usage example.
- Response: Yes, “What is your car’s opinion?” is a representative example of “opinion”. For example, you could ask a friend, “What is your opinion on the new movie?”

The metaphorical errors concern usage examples that are linguistically correct but do not adequately convey the term’s literal meaning. At the same time, however, they could be considered useful examples in a metaphorical sense of the term. Here is an example of a metaphorical error.

- Query: Is the phrase “The conclusion of the story was its beginning.” a representative example of “conclusion”? Start your answer with yes or no and include a good usage example.
- Response: Yes, this phrase is a representative example of “conclusion”. For example, the conclusion of the novel was that the protagonist had to go back to the moment when the story began in order to resolve the conflict.

Table 3 presents statistics for the three types of errors.

**Table 3.** Statistics of three types of Errors.

	Homonym Error	Conceptual Error	Metaphorical Error
Number of phrases	2	3	4
FP	–	78	200
FN	100	–	–
TN	–	72	–

According to Table 3, we had a Homonym error in two phrases. This happened all fifty times that each phrase was submitted as a question. We had conceptual errors in three phrases. In this case, GPT-VT considered about half of the user responses correct and the remaining incorrect. Finally, four phrases were detected with a metaphorical error. All questions related to them were rejected.

## 6 Conclusions

This paper presents an artificial vocabulary tutor based on GPT-3. Its overall performance is shown by the measurements in Table 2. The cases where GPT-VT fails to evaluate correctly, we believe, can be largely mitigated. As for homonym errors, the question to GPT-3 could be formatted to include the term's meaning along with the user's answer. This may limit, perhaps even eliminate, homonym errors. For the other types of errors, more investigation is needed. Apparently, GPT-3 takes examples that make metaphorical use of a term primarily as correct ones.

On the other hand, conceptual errors are sometimes considered correct and sometimes incorrect. Also, it should be noted that the temperature parameter plays a significant role in shaping the results. We intend to do experiments with more conditions, with varying temperature values and, of course, with better statistical processing of the results. Also, we intend to compare our model using GPT-3 and ChatGPT. Moreover, our work should be extended to investigate the effect of the vocabulary tutor on the student's knowledge acquisition. Although there is a lot to do, the results so far show us that the road to such reliable applications is wide open as long as we take it.

**Acknowledgements.** This work was supported by the MPhil program “Advanced Technologies in Informatics and Computers”, hosted by the Department of Computer Science, International Hellenic University, Greece.

## References

1. Radford, A., Narasimhan, K.: Improving Language Understanding by Generative Pre-Training. <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035> (2018). Accessed: 16 Jan. 2023
2. Vaswani, A., et al.: Attention Is All You Need. arXiv, 05 Dec 2017
3. Paladines, J., Ramirez, J.: A systematic literature review of intelligent tutoring systems with dialogue in natural language. IEEE Access **8**, 164246–164267 (2020)
4. Mousavinasab, E., Zarfsanaiey, N., Niakan Kalhorri, S.R., Rakhshan, M., Keikha, L., Ghazi Saedi, M.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interact. Learn. Environ. **29**(1), 142–163 (2021)
5. Labrie, G.: A french vocabulary tutor for the web. CALICO J. **17**(3), 475–499 (2000)
6. Kazi, S.A.: VocaTest : an intelligent tutoring system for vocabulary learning using the ‘mLearning’ approach. <https://repository.nie.edu.sg/handle/10497/217> (2005). Accessed: 15 Jan 2023
7. Stockwell, G.: Vocabulary on the move: Investigating an intelligent mobile phone-based vocabulary tutor. Comput. Assist. Lang. Learn. **20**(4), 365–383 (2007)
8. Jung, J., Graf, S.: An approach for personalised web-based vocabulary learning through word association games. In: 2008 International Symposium on Applications and the Internet, pp. 325–328 (2008)
9. Juffts, A., Friedline, B.E.: Sociocultural influences on the use of a web-based tool for learning English vocabulary. System **42**, 48–59 (2014)

10. Lucioni, A., Nkambou, R., Massardi, J., Bourdeau, J., Coulombe, C.: STI-DICO: a web-based system for intelligent tutoring of dictionary skills. In: Proceedings of the 25th International Conference Companion on World Wide Web, Republic and Canton of Geneva, CHE, pp. 923–928 (2016)
11. Ebner, M., Edtstadler, K., Ebner, M.: Tutoring writing spelling skills within a web-based platform for children. *Univ. Access Inf. Soc.* **17**, 305–323 (2017). <https://doi.org/10.1007/s10209-017-0564-6>
12. Woo, J.H., Choi, H.: Systematic Review for AI-based Language Learning Tools. arXiv, 29 Oct 2021
13. Towards a robot assistant in vocabulary learning - International Conference on Robotics, Mechatronics and Social Implementation 2018, Varna, Bulgaria, In: HUMAIN Lab. <http://humain-lab.cs.ihu.gr/index.php/2018/09/01/towards-a-robot-assistant-in-vocabulary-learning/?lang=en> (2018). Accessed 16 Jan 2023
14. Moussiades, L., Tsimerperidis, I., Karanakis, S.: Dictionary structure identification. *Eng. Sci. Technol. Rev.* **14**(6), 28–34 (2021)
15. ChatScript: SourceForge. <https://sourceforge.net/projects/chatscript/>. Accessed 16 Jan 2023
16. Moussiades, L., Zografos, G.: LIA: a mathbot assistant for first-grade students. *Eur. J. Eng. Technol. Res.* **6**(5), 89–93 (2021)
17. Bonwell, C.C.: Active Learning : Creating Excitement in the Classroom. School of Education and Human Development, George Washington University, Washington, D.C. <http://archive.org/details/activelearningcr0000bonw> (1991). Accessed 18 Jan 2023
18. Beed, P.: Moving Learners toward Independence: The Power of Scaffolded Instruction. *The Reading Teacher*. <https://www.semanticscholar.org/paper/Moving-Learners-toward-Independence%3A-The-Power-of-Beed/a7ff77e4215f373570407892ac9fc77ba3f6399> (1991). Accessed 18 Jan 2023
19. McLeod, S.A.: What Is the Zone of Proximal Development? <https://www.simplypsychology.org/Zone-of-Proximal-Development.html> (2012). Accessed 18 Jan 2023
20. How do text-davinci-002 and text-davinci-003 differ?. <https://help.openai.com/en/articles/6779149-how-do-text-davinci-002-and-text-davinci-003-differ>. Accessed 21 Jan 2023
21. Zong, M., Krishnamachari, B.: A survey on GPT-3 (2022)



# ETHOSCHOOL: An Artificial Moral Agent Model for Collaborative Learning

Paulo Roberto Córdova<sup>1</sup> , Iderli Pereira de Souza Filho<sup>2</sup> ,  
Gilleannes Thorwald Guedes<sup>2</sup> , and Rosa Maria Vicari<sup>1</sup>

<sup>1</sup> Federal University of Rio Grande do Sul-UFRGS, Porto Alegre, Brazil

[paulo.cordova@ifsc.edu.br](mailto:paulo.cordova@ifsc.edu.br), [rosa@inf.ufrgs.br](mailto:rosa@inf.ufrgs.br)

<sup>2</sup> Federal University of Pampa-UNIPAMPA, Alegrete, Brazil

[gilleannesguedes@unipampa.edu.br](mailto:gilleannesguedes@unipampa.edu.br)

**Abstract.** The growing popularization of Artificial Intelligence (AI) technologies to support most different human activities, including education, has been a subject of research and concern among researchers in the area. In this sense, how to translate ethical principles into computational models; how to avoid biases; how to make intelligent systems accountable; and how to make them more predictable and able to explain their actions are among the main issues raised. Such questions also extend to the Education area, where it is necessary to address this matter with the same seriousness. Despite this, few proposals in ethics by design have been observed for the Education area. To help fill this gap, this work presents a proposal for an Artificial Moral Agent (AMA) to aid in collaborative learning. The proposed AMA implements a hybrid approach between deontological and utilitarian ethical frameworks, using the Belief, Desire, and Intention (BDI) architecture. For this, through Agent-Oriented Software Engineering (AOSE) techniques, which is the focus of this paper, the functional requirements of the proposed AMA will be described, as well as its structural model.

**Keywords:** Artificial Intelligence · Ethics · Education · Software Design

## 1 Introduction

The use of AI technologies in the Education field have rising some reflections on the risks and effects that the use of this kind of solutions for teaching and learning represents for future generations [3]. In this sense, the ethical implications of using intelligent systems capable of supporting, driving, or influencing learning need to be considered with the same seriousness as they have been in other areas.

The present work proposes an investigation in ethics by design, a research area which seeks to develop algorithmic solutions to endow autonomous artificial systems with the capacity for ethical reasoning. In this context, the set of efforts undertaken by the research community, to produce intelligent systems aligned with human values, is organized in a field called Value Alignment (VA) in AI [9]. Researchers in this field have sought answers to questions related to how to translate ethical principles into computational

models, how to avoid data bias, how to make intelligent systems accountable for their decisions, among others [9]. This set of questions gave rise to the development of a special type of artificial agent capable of demonstrating behaviors considered ethical: the Artificial Moral Agent. This kind of agents must be capable of demonstrating moral behavior or preventing immoral behavior [4].

Despite the importance and the clear need to search for solutions in ethics by design, few initiatives have been observed in Education area. To help fill this gap, this article proposes and describes an AMA model to support collaborative learning using Agent-Oriented Software Engineering (AOSE) techniques. About this proposal, one can see previous results in [5] and [6].

The aim of this research is to investigate how to make more ethical use of AI in education area. To test our proposal, presented in more details in [5] and [6], we chose a collaborative learning scenario. In this sense, to understand how computer systems have been used to support collaborative learning, one can resort to Computer Supported Collaborative Learning (CSCL), a branch of learning science concerned about investigating how people can learn together by being aided by computers [11]. We have addressed this topic and its relation to this work and presented a conceptual outline of the AMA described in this paper, including scenarios and examples of application in [5].

Finally, to allow understanding of the proposed model, the functional requirements of the proposed AMA will first be described and, subsequently, its structural model will be presented.

## 2 Methodology

In this work, different methodologies and modeling languages are used for different stages of the proposed MAS project. This way, for the specification of requirements, the Multi-Agent Systems Requirements Modeling Language (MASRML), a domain-specific language based on the Unified Modeling Language (UML) for modeling requirements in Multi-Agent Systems (MAS) projects [8], was applied.

Regarding the structural aspect of the proposed MAS, the Multi-Agent System Modeling Language MAS-ML will be applied. This language can represent the essential static and dynamic aspects of the MAS, emphasizing a clear representation of their concepts and relationships [10].

## 3 Results and Discussion

The present work envisages the modeling of the Ethoschool, a MAS designed to guide collaborative learning groups through discussion forums considering ethical criteria. This solution has an AMA based on the BDI architecture, capable of making decisions based on predetermined deontological moral principles and dealing with ethical dilemmas using utilitarian reasoning [6].

Ethoschool was idealized to act as a CSCL, interacting with students to maintain the team of work cohesive. Besides, this CSCL ought to operate so that all students participate in an equitably way during the activities time.

In this way, the proposed solution is constituted by an environment in which there are two agent roles named Monitor and Tutor. The first one aims to fetch and organize the interactions carried out by the students and registered in the forum's database. Such a fetch will temporally take place and once the Monitor has the interactions data, it will have as its plan to distribute the information, sending it to the Tutor agent.

The Tutor agent, by its turn, is an AMA that follows the top-down approach [1] and aims to analyze the need for intervention by the agent in the students' interactions in the forum. This approach was chosen, as it is understood that learning ethical behavior is inappropriate for the classroom environment [6].

Therefore, the Ethoschool needs to follow rules derived from some of the ethical principles established by [12], in some characteristics of CSCL systems, and in some collaborative learning assumptions. These rules and their relationship with the principles proposed by [12] were described in [5]. Table 1 shows the rules that the Tutor ought to follow.

**Table 1.** Rules the Tutor Ought to Follow.

ID	Priority	Rules
1	1	Some student in the group must interact
2	2	All students must interact
3	3	All students must interact equitably
4	4	All students must continue to interact through-out the activity
5	1	All groups must complete activities in time
6	1	All students must be informed of their peer interactions
7	1	Do not interfere in students' decisions
8	1	Avoid disproportionate levels of demand
9	1	Avoid sending discouraging messages
10	1	Avoid exposing the student to peers in the group

The imposing character of the rules presented in Table 1, as described in more detail in [5], is directed to the Tutor agent, who must seek to satisfy them, even if it is not always possible, given the environmental variables. The contradictions, as well as the prioritization between the presented rules in order to mitigate possible conflicts are further explained in [5].

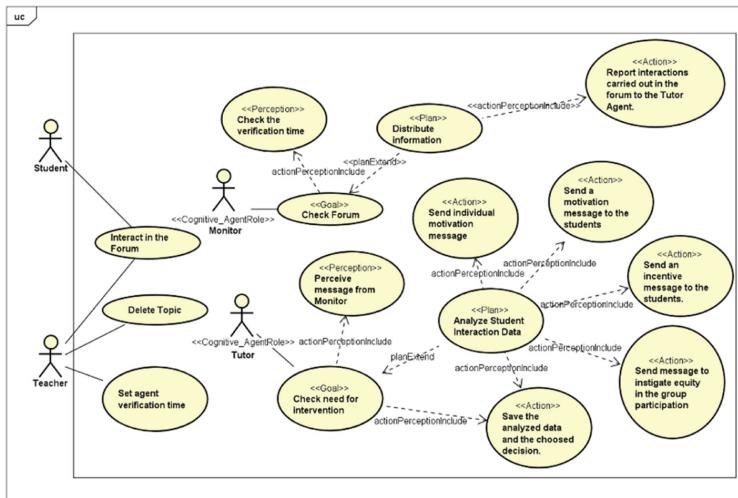
Regarding the other functional requirements of the Tutor agent, its perception is limited to perceiving the messages sent by the Monitor agent containing the compiled data from the interactions carried out by the students in the forum. Inside the Tutor's plan is included the analysis of this data, and the actions resulting from this analysis must be executed under environmental conditions. In this sense, in [5] are described the complexity of the environment, as well as the fourth scenario modeled to explain in more detail the behavior of the proposed MAS.

As for the utilitarian dimension of the AMA, the algorithm known as Hedonistic Act Utilitarianism (HAU) will be used, also as described in [5]. The HAU formula is presented below.

$$\text{Total Net Pleasure} = \sum_{i=1}^n (\text{Intensity}_i \cdot \text{Duration}_i \cdot \text{Probability}_i)$$

### 3.1 Requirements Specification

During requirements engineering, we used the MASRML modeling language [8] to help to identify specific functional requirements for the Ethoschool Multi-agent system. Thus, an Use Case Diagram (UCD) was produced, as shown in Fig. 1.



**Fig. 1.** Representation of Functional Requirements Using MASRML.

The requirements modeled in this UCD represent system behaviors. Therefore, when we examine Fig. 1, we can see that the Student and Teacher actors are normal UML actors, representing roles assumed by external users who operate the system. While Monitor and Tutor are AgentRole\_Actors, as shown by their stereotypes, representing roles that can be assumed by agents and commit to their goals.

Still examining Fig. 1, we can see that the use cases “Interact in the forum”, “Delete topic”, and “Set agent verification time” are normal use cases, that is, behaviors that can be performed by users who assume the role of teacher or student in the system.

The Internal Use Cases (IUC) “Check Forum” and “Check need for intervention” are goals, as can be seen from their stereotypes. These IUCs are associated with AgentRole\_Actors Monitor and Tutor respectively. This means that the agents who assume these roles must verify whether these goals can be achieved and act on them if they do.

Since for these desires (goals) to become intentions, it is often needed for agents to receive perceptions from the environment that make them believe that the goals can be achieved, there is an IUC representing a perception (stereotype Perception) associated with each IUC Goal (“Check the verification time” and “Perceive message from monitor”). It should be noted that these Perception IUCs are associated with Goal IUCs through actionPerceptionInclude associations, which is similar to the standard UML inclusion association in which the execution of the IUC Goal implies the execution of the IUC Perception associated with it. Thus, whenever goals are executed by agents, their perceptions must also be executed to determine if the goal can become an intention.

It can also be observed that there are IUCs with the Plan stereotype associated with the IUCs Goals. These IUCs represent the plans that will be executed by an agent if he believes that the goal can be achieved and has become an intention. Plan IUCs are associated with Goals IUCs through PlanExtend relationships. This means that plans will only be executed upon satisfaction of a certain conditions, in the case where the agent believes that the goal in the matter has become an intention. To exemplify the functioning of the MASRML documentation, Table 2, Table 3, and Table 4 present the documentation related to the Tutor Agent.

**Table 2.** Goal – Check need for intervention

Internal Use Case Name	Check need for intervention
Stereotype	Goal
AgentRoleActor	Tutor
Description	Describes the steps followed by the Tutor Agent to verify the need and the type of intervention to each student or group.
Initial Beliefs	Students interacting properly
Perceptions	Polling for new messages from the Monitor
Main Scenario	
AgentRoleActor Actions	
1. Execute the internal use case "Perceiving message from Monitor".	
Alternative Scenario - Students not interacting properly.	
AgentRoleActor Actions	
1. Evidence that there are students not interacting.	
2. Change belief to need to verify intervention = true.	
3. Solve ethical dilemma using the Hedonistic Act Utilitarianism (HAU).	
4. Make the goal into an intention.	
Alternative Scenario 2 - Students are interacting properly.	
AgentRoleActor Actions	
1. Evidence that students are interacting properly.	
2. Tutor agent saves the decision made and the data used to make it.	

The Tables 2, 3 and 4 describe the behavioral steps relative to the perception and plan associated to the goal “Check need for intervention”. Next, more details about the internal functionalities, which cannot be represented by the UCD, will be discussed.

**Table 3.** Perception – Perceive message from Monitor.

Internal Use Case Name	Perceive message from Monitor
Stereotype	Perception
AgentRoleActor	Tutor
Description	Describes the probing for possible Monitor Agent's messages.
Preconditions	The internal use case Goal "Check the need for intervention" must be in execution.
Initial Beliefs	Message received from Monitor = false. Main Scenario AgentRoleActor Actions
	1. Probe the environment to verify the occurrence of a new message sent from the Monitor containing the data of the students' interactions in the forum.
	Alternative Scenario - Monitor message sent AgentRoleActor Actions
	1. Come to believe that Student data was sent. 2. Receive Student Data.

### 3.2 Structural Model

During the design phase, we used the MAS-ML language [7] to model the structure of the Ethoschool. As one can see, MASRML [8] and MAS-ML complement each other in describing software characteristics.

Structural models help to illustrate the physical or logical composition of software from its components. In this sense, structural modeling establishes the boundary between the software being modeled and the environment in which it must operate [2].

This way, Fig. 2 was produced, presenting a structural model in the form of a MAS-ML organizational diagram, which allows identifying the types of agents supported by the system and the roles that they can assume in its organization.

In this diagram, there are two types of agents and two agent roles that can be assumed. Thus, when looking at Fig. 2, we can see that Ethoschool supports the Monitor and Tutor agent roles, which are interpreted by the Assistant and Ethical agents, which, in turn, are agent classes.

The Assistant AgentRole has the purpose of checking the forum to verify students' interactions. In addition, it also has a protocol to communicate the interactions performed by the students to the tutor. The Monitor agent has the plan to distribute information to the agent that assumes the Tutor role. This plan will only be executed when the check forum goal becomes an intention. The Assistant agent also has the beliefs through which it believes or not that the students are interacting properly and whether or not the students have started a new task on the forum.

Tutor AgentRole, in turn, aims to verify the need for intervention to encourage students' participation by analyzing the forum interaction data sent by the agent that assumes the Monitor role and perceiving the receipt of new messages sent by him. This role has also protocols for sending encouraging messages to the students, as well as the duty of saving the decision made and the data used to make it.

For this, the analysis of the students' interactions data is materialized in the form of an agent's plan, which is triggered when the goal of verifying the need for intervention

**Table 4.** Plan – Analyze Student Interaction Data

Internal Use Case Name	Analyze Student Interaction Data
Stereotype	Plan
AgentRoleActor	Tutor
Description	Describes the steps followed by the Tutor Agent when analyzing data from student interactions in the forum.
	Main Scenario
	AgentRoleActor Actions
1. To determine the number of students who are not interacting.	
	Alternative Scenario - Identified student not interacting
	AgentRoleActor Actions
1. Send individual motivating message.	
2. The Tutor agent saves the decision made and the data used to make it.	
	Alternative Scenario - Identified students who are not interacting equitably
	AgentRoleActor Actions
1. Send message to instigate equity in the group participation.	
2. Tutor saves the decision made and the data used to make it.	
	Alternative Scenario - Identified team of students not interacting
	AgentRoleActor Actions
1. Send a motivating message to the students of the team.	
2. Tutor agent saves the decision made and the data used to make it.	
	Alternative Scenario - Identified a team of students with large gap since last interaction
	AgentRoleActor Actions
1. Send a message to the group reminding about the task.	
2. Tutor agent saves the decision made and the data used to make it.	
	Alternative Scenario - Utilitarian reasoning deliberates for non-intervention
	AgentRoleActor Actions
1. Determine the inadequacy of the intervention.	
2. Tutor agent saves the decision made and the data used to make it.	

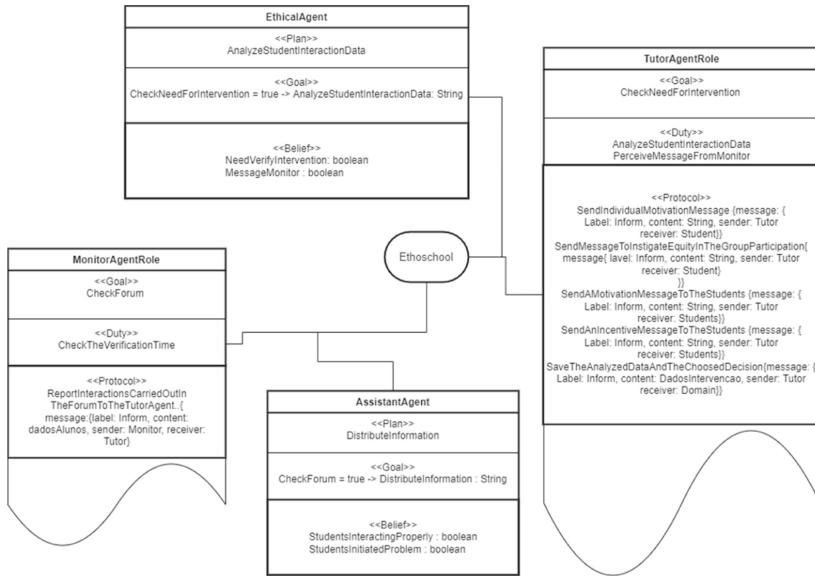
of the Tutor AgentRole becomes an intention. This agent also has beliefs determining whether or not it is necessary to execute the need for intervention analysis and whether or not a message has been received from the Monitor.

## 4 Final Considerations

Solutions for a more responsible, explainable, predictable, and therefore, reliable AI are an important contemporary demand. In this sense, one can say that in the teaching and learning context, it is an even more sensitive issue, as AI technologies can direct, with greater or lesser autonomy, the directions of a classroom.

Aiming to help fill some gaps in this research area, this work proposes a hybrid solution based on the deontological and utilitarian ethical frameworks. Such a solution uses the BDI model to construct a MAS capable to guide students in a collaborative learning group interacting through a forum.

During the requirements engineering phase of the proposed solution, we applied the MASRML language to identify and represent the functional requirements of the



**Fig. 2.** MASML Class Diagram

Ethoschool. In addition, the structural design was produced by applying the MASML language. The model was designed to guide its decisions on ethical principles based on UNESCO's proposal for ethics in AI and to be able to deal with ethical dilemmas.

This is possible due to its structure that embraces a Monitor Agent, responsible for monitoring the forum searching for students' interaction data and sending them to the Tutor Agent. The late is an AMA, responsible for processing data from students' interactions in the forum and deciding if, how, and when to intervene in student interactions.

Such a solution can bring at least two advantages: as agents based on the BDI model are guided by intentions, it is more intuitive for the end-user to understand why an agent made a certain decision; and the use of ethical frameworks already validated by a community of research, such as ethical principles and the use of reasoning capabilities to solve ethical dilemmas can make the decisions made by intelligent systems more predictable [5]. These features help to construct an AI solution able to meet an important requirement for ethics in AI, named explainability.

Thus, we believe that this solution can help make AI more reliable, especially in the education field, in which, more than relating to people, AI can drive learning processes playing an important impact on human formation and in society. For these reasons, the model proposed in this work can contribute to a more ethical AI in Education area.

**Acknowledgements.** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## References

1. Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics Inf. Technol.* **7**(3), 149–155 (2005)
2. Bourque, P., Farley, R.E.: Guide to the Software Engineering Body of Knowledge (SWEBOK (R)): Version 3.0. IEEE Computer Society Press (2014)
3. Casas-Roma, J., Arnedo-Moreno, J.: From Games to Moral Agents: Towards a Model for Moral Actions, pp. 19–28. IOS Press (2019)
4. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: a survey of the current status. *Sci. Eng. Ethics* **26**(2), 501–532 (2020)
5. Córdova, P.R., Vicari, R.M.: Practical ethical issues for artificial intelligence in education. In: Reis, A., Barroso, J., Martins, P., Jimoyiannis, A., Huang, R.Y.M., Henriques, R. (eds.) *Technology and Innovation in Learning, Teaching and Education*, pp. 437–445. Springer Nature Switzerland, Cham (2022)
6. Córdova, P.R., Vicari, R.M., Brusius, C., Coelho, H.: A proposal for artificial moral pedagogical agents. In: Rocha, A., Adeli, H., Dzemyda, G., Moreira, F., Ramalho Correia, A.M. (eds.) *Trends and Applications in Information Systems and Technologies*, pp. 396–401. Springer International Publishing (2021)
7. Da Silva, V.T., de Lucena, C.J.: Mas-ml: a multiagent system modelling language. *Int. J. Agent-Oriented Softw.* **2**, 382–421 (2008)
8. Guedes, G.T.A., de Souza Filho, I.P., Gaedicke, L.F., Mendonça, G.D., Vicari, R.M., Brusius, C.: MASRML-a domain-specific modeling language for multi-agent systems requirements. *Int. J. Softw. Eng. Appl. (IJSEA)* **11**(5), 25–45 (2020)
9. Kim, T.W., Donaldson, T., Hooker, J.: Grounding value alignment with ethical principles (2019)
10. da Silva, V.T., Lucena, C.J.: MAS-ML: a multi-agent system modeling language. In: Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, pp. 126–127 (2003)
11. Stahl, G., Koschmann, T., Suthers, D.D.: Computer-supported collaborative learning. In: Keith Sawyer, R. (ed.) *The Cambridge Handbook of the Learning Sciences*, pp. 409–426. Cambridge University Press (2012). <https://doi.org/10.1017/CBO9780511816833.025>
12. UNESCO: first draft of the recommendation on the ethics of artificial intelligence (2020)



# Toward a Smart Tool for Supporting Programming Lab Work

Anis Bey<sup>1,2(✉)</sup> and Ronan Champagnat<sup>1</sup>

<sup>1</sup> La Rochelle University, La Rochelle, France

{anis.bey,ronan.champagnat}@univ-lr.fr

<sup>2</sup> Higher School of Management Sciences, Annaba, Algeria

**Abstract.** In a lab work session, students may spend an inordinate amount of time attempting to correct simple errors, repeatedly revisiting and repeating the same errors due to the limited access to the tutor. Instructors, on the other hand, frequently find themselves explaining the same errors. This tool is designed to improve the educational environment in the computer lab for both students and instructors. The first experiment did not show that the tool improves learning programming, but it showed some insights about students that use the tool only in face-to-face sessions and students that use it only with the training mode compared to a control group.

**Keywords:** Smart lab · Introductory programming · Intelligent Tutoring System · Feedback · Learning Analytics

## 1 Introduction

One way to help students overcome difficulties in learning to program is to give them programming tasks to complete in the computer lab with the assistance of a tutor. A tutor can inspect the students' code, investigate the issue, and then provide the students with the necessary assistance. However, communication between students and instructors in the lab is usually limited. The amount of time the instructor can devote to each student decreases as the number of students in the lab increases. With limited access to the tutor, students may spend undue time attempting to correct simple errors, repeatedly revisiting and repeating the same. They may need help comprehending compiler messages, making code correction a frustrating experience.

Many researchers have studied diagnostic messages generated by compilers and interpreters during the last decade, such as syntactical errors. [1] states that these messages, which include error, warning, and run-time messages, are difficult to understand, especially for beginners.

In recent years, various tools have been developed to help instructors provide feedback to students on their work. All the tools of the state of the art were intended to provide feedback on students' assignments after completing and

submitting their work rather than for their work on programming activities in lab classes.

This paper presents the main functionalities of Smart Lab technology for introductory programming computer laboratories and the early results of the first experiment. We addressed two research questions :

**RQ1:** *Are there any significant differences in the final course score between groups of students that did not use the tool, students who used only the smart lab with a teacher, and students that have been logged in, and practicing by solving some exercises freely?*

**RQ2:** *Is there any significant difference in the final course score between students in terms of the number of solved exercises?*

This paper is organized as follows. Section 2 shows some solutions to address the difficulties faced by programming students and the Smart Lab environment, with a presentation of a new Smart Lab described in Sect. 3, followed by Sect. 4, which presents a general workflow of how the tool works. Section 5 gives the leading indicators of students' and teachers' dashboard. In Sect. 6, details of the result of the first experiment of the tool are provided, followed by a discussion, conclusion, and guidelines for future work.

## 2 Smart Lab Previous Work Survey

A Smart Lab is a lab that is outfitted with tools to improve instruction and learning [3]. The Smart Lab concept is to use technology to provide students and teachers with tools to improve their ability to communicate effectively and engage successfully with the curriculum. A Smart Lab can be implemented in a variety of ways to achieve various objectives. [4], for example, describe a lab that was built in layers using a variety of approaches and devices. The main goal of that implementation was to allow students to visualize problems and their solutions on the lab's large display screen.

Prieto-Blazquez et al. in [2] created a virtual programming laboratory known as VPLab. They identify several key components needed to ensure VPLab's success in improving students' knowledge and skills in computer programming language. These essential components are further divided into three types of resources: technological resources, pedagogical and strategic resources, and academic staff resources. The technological resources concentrate on artifacts of technology that can be used to simulate virtual laboratories and evaluate student knowledge and comprehension. Learning methodology, support documentation and other materials, and assessment are the pedagogical and strategic resources used in VPLab. The authors conducted a questionnaire-style survey to 284 distance learning students in order to assess the relevance of the suggested structure and its critical components. The main finding in the study was that although technological resources were rated highly by the students, the distance learning students appeared to place more importance on pedagogical and human factors.

In [5], authors present the Lab4CE environment, a remote laboratory for computer education. This tool aims to increase students' engagement in remote

practical activities by analyzing students' interactions. The experiment was conducted using 139 undergraduate students enrolled in the first year of a computer science degree. The results have shown that the framework has a positive effect on learners' engagement. Also, they found a significant positive correlation between students' activity and students' course achievement.

In general, smart lab technology implementations found in the literature were complex and included various components. As well as expensive hardware, they also required expensive displays, servers, and video cameras. To our knowledge, no smart lab technology reviewed dealt specifically with computer programming labs.

### 3 Software Description

The smart lab tool is a web application that aims to facilitate learning programming. The tool offers two modes.

The first mode is to be used during a lab work session with a teacher. Once logged in, students must connect to the lab work session with an access code delivered by the instructor. They find the exercises of the current session defined by the instructor. Students can start solving an exercise, propose and run a solution, and receive an immediate assessment of the correctness and feedback about some code quality rules.

The instructor can consult all submissions and give instant feedback to those struggling.

The second mode of this tool is practicing. Students can access the tool and practice by solving some exercises. They receive the same feedback as in the smart lab mode used in the lab work session, but they do not have a teacher to notify them and give them more precise and tailored feedback.

### 4 Workflow

Teachers create exercises by defining three pieces of information: the exercise statement, the aimed competencies, and the input/output tests. The IO test specifies an input to a program and the expected output. If the students' program produces the expected output, the test is passed, and they receive points. The tool user interface makes it simple to create IO tests. Multiple pairs of input and output can be used in a single IO test. The aimed competencies are selected to help students choose exercises to solve according to their weaknesses.

### 5 Dashboards for Insights and Support During Programming

We have conceived a dashboard for both students and teachers. For Students, to give some indicators about how they perform and to self-regulate their learning. For teachers, to give them some students' insights and practices that can be used to give some feedback in the face-to-face course.

## 5.1 Teachers' Dashboard

Teachers have a dashboard built with some indicators about the performance of students. We find the following indicators:

- Number of students: Total number of signed up students.
- Number of active students: Total number of students that have submit at least one submission during the last week.
- Number of submissions: Total number of submissions made by students.
- Percentage of each type of submission: Correct, Erroneous, In Progress, Compilation Error.
- Average percentage of Indentation, Comments, Expression, Variable, Repetition (see Sect. 5.3 for more information)
- Distribution of submissions' types per student.
- Number and type of submissions (correct, in-progress, correct, compilation Error) per day.

## 5.2 Students' Dashboard

Students' dashboard is described and composed of the following indicators :

- The rank of students in terms of solved exercises.
- Number of solved exercises.
- Number of remaining exercises.
- Percentage of each type of submission: Correct, Erroneous, In Progress, Compilation Errors.
- Average percentage of each feature: indentation, comments, expression, variable, repetition (see Sect. 5.3 for more information).

## 5.3 Student's Final Report

Whether in the smart lab mode with teacher or in the practicing mode, when a student submits a solution for a given exercise, the tool gives a report in the form of three parts:

**Output.** In this space, the student will see if there is a syntactical error in his codes; otherwise, the number of successful IO tests.

**Feedback.** The submitted code is evaluated in this part of the report according to five features:

- **Comments:** to check if the submission is well commented.
- **Expressions:** to evaluate expression complexity, we just look at compound expressions and long-expression lines.
- **Indentation:** to evaluate indentation/formatting consistency.
- **Repetition:** Repeating code is probably the top indicator of poor abstraction.
- **Variables:** to evaluate if some variable names have meaningful lengths and are scoped within loops.

**Quality.** Here the quality of how well the code was written is reported according to 29 rules. The rules were gathered to give guidelines for developing reliable, maintainable, clearly structured, and clearly written object-oriented programs with special emphasis on the correct use of the risky features of the C++ programming language.

## 6 Early Experiment

A first early experiment was conducted with 65 first-year students of an introductory C++ programming university course. Students did not have prior knowledge of programming, but they mastered basic operations to use a computer system. Three lab work sessions were organized with a teacher in a computer lab where students were asked to login into the tool and used an access code to connect to a session defined by the teacher. Also, we allowed the students to use the tool outside the face-to-face lab work session so that they could practice it.

Also, we used a group of students that have not participated in the experiment as a control group ( $n = 15$ ) for comparison against students who used the tool.

### 6.1 Results and Discussion

The experimental protocol was designed to test two modes. The first mode is the use of the tool during lab work sessions. In this mode, students used the tool to do the exercises of the lab work with the teacher. The teacher can use the dashboard to see all submissions and also some indicators for all students of the session and also the progression of each student individually, as described in Sect. 5.1.

**Table 1.** Mean and standard deviation of each feature (clusters of level2).

Session type	#Students	#Submissions	%Correct	%Progress	%Compil.	%Erronuous
SmartLab1	24	270	11.85	15.18	29.25	43.70
SmartLab2	29	368	17.39	4.34	42.11	36.14
SmartLab3	11	183	8.19	2.73	50.81	38.25
Training	21	275	11.63	4.36	37.45	46.54

The second mode consists of letting students free if they want to access on their own to solve some exercises in the tool when they are not in the lab session with the teacher.

Table 1 shows the distribution of students in each session with the number of submissions, the percentage of correct, in-progress, compilation errors, and erroneous submissions.

In the first session, 24 students participated and produced 270 submissions trying to solve three exercises. About 12% of these submissions were correct, 15% somewhat correct submissions, 29% of submissions had syntactical errors, and 43% were syntactically correct, but they did not provide the expected output.

In the second session, 29 students participated and produced 368 submissions trying to solve three exercises. About 17% of these submissions were correct, 4.35% were somewhat correct submissions, 42% of submissions had syntactical errors, and 36.14% were syntactically correct, but they did not provide the expected output.

In the third session, only 11 students participated and produced 183 submissions trying to solve three exercises. About 08% of these submissions were correct, 3% somewhat correct submissions, 51% of submissions had syntactical errors, and 38.25% were syntactically correct, but they did not provide the expected output.

However, in the training mode, where students are free to log in and practice by solving some exercises, only 21 students have logged in and tried to practice. They have produced 275 submissions where 11.63% were correct, 4.35% were somewhat correct submissions, 37.45% of submissions having syntactical errors, and 46.54% were syntactically correct, but they did not provide the expected output.

**RQ1:** *Are there any significant differences in the final course score between groups of students that did not use the tool (control group), students who used only the smart lab with a teacher, and students that have been logged in and practicing by solving some exercises freely?*

We found a statistically-significant difference in average final course score according to session type ( $F(3) = 3.096, p < 0.05$ ). A Tukey post-hoc test revealed significant pairwise differences between students that just train by solving some exercise and the control group with an average difference of 3.3 ( $p < 0.05$ ).

However, there is no significant difference between the other groups of students. The control group represents the group that has obtained the lowest final course score ( $mean = 7.93, sd = 2.19$ ). Students who have practiced by solving some exercises without participating in sessions with teachers have obtained the highest final course score ( $mean = 11.24, sd = 2.83$ ).

To analyze the characteristics of each group of students according to the mode of practicing, we summarized the mean and standard deviation values of each group and this for the number of submissions, the number of correct, in-progress, compilation.Error and erroneous submissions as described in Table 2.

A Kruskal-Wallis test was conducted to determine whether there is a significant difference in the medians of these groups of students.

- Number of submissions: The test statistic was  $H = 5.87$ , with  $2^{\circ}$  of freedom, and the p-value was  $< 0.05$ . Based on this result, we reject the null hypothesis and conclude that there is a significant difference in the medians of the groups. Post-hoc analysis using the Dunn's test revealed that students that were in

both SmartLab sessions and Training mode were significantly different from students that only were in Training mode ( $p = 0.04$ ), but there was no significant difference between the other groups.

- Number of correct submissions: The test statistic was  $H = 1.78$ , with  $2^{\circ}$  of freedom, and the p-value was  $> 0.05$ . Based on this result, we accept the null hypothesis and conclude that there is a no significant difference in the medians of the groups.
- Number of in-progress submissions: The test statistic was  $H = 5.07$ , with  $2^{\circ}$  of freedom, and the p-value was  $> 0.05$ . Based on this result, we accept the null hypothesis and conclude that there is a no significant difference in the medians of the groups.
- Number of submissions with compilation errors: The test statistic was  $H = 8.5$ , with  $2^{\circ}$  of freedom, and the p-value was  $< 0.05$ . Based on this result, we reject the null hypothesis and conclude that there is a significant difference in the medians of the groups. Post-hoc analysis using the Dunn's test revealed that students that were in both SmartLab sessions and Training mode were significantly different from students that only were in Training mode ( $p = 0.01$ ), but there was no significant difference between the other groups.
- Number of erroneous submissions: The test statistic was  $H = 6.35$ , with  $2^{\circ}$  of freedom, and the p-value was  $< 0.05$ . Based on this result, we reject the null hypothesis and conclude that there is a significant difference in the medians of the groups. However, Post-hoc analysis using the Dunn's test did not reveal any difference between students' groups.

According to this analysis, we found that only the number of submissions and the number of submissions with compilation errors distinguish the group of students that only practice in the training mode from the students' group that use the smartlab in face-to-face sessions with the teacher and also practice in the training mode.

**Table 2.** Mean and standard deviation of types of submissions according to practicing mode.

Session mode	N	Submissions	Correct	In-Progress	Compil.Error	Erroneous
SmartLab	43	16.58(15.35)	2.60(3.35)	2.09(5.07)	6.41(7.48)	5.46(6.20)
Training	11	17.92(9.49)	2.30(1.31)	0.92(1.93)	6.92(5.45)	7.76(6.90)
SmartLab+Training	08	33(20.16)	2.62(3.46)	0.12(0.35)	17.37(10.45)	12.87(11.25)

**RQ2:** *Is there any significant difference in the final course score between students in terms of the number of solved exercises?*

To answer this question, we have compared groups of students according to the number of solved exercises (one, two, or three exercises). The result shows that there is no significant difference between students having solved one exercise ( $M = 9.48$ ,  $SD = 2.08$ ) and with whom solved two ( $M = 8.8$ ,  $SD = 3.28$ ) or

three ( $M = 10.75$ ,  $SD = 3.38$ ) exercises in terms of final course score ( $p < 0.05$ ). However, there is a significant difference between students having solved two exercises and who solved three exercises.

## 7 Conclusion and Future Works

This research resulted in using smart classroom technology in introductory programming lab works. The Smart Lab system was successfully implemented and met its intended goal of providing instructors with specific, timely, and detailed information about their student's performance on programming tasks during lab class sessions. The Smart Lab improved the overall effectiveness of the lab session in achieving the learning objectives by facilitating communication between tutors and their students in the lab.

This tool allows teachers to understand their students' programming strategies and compiling behaviors and identify the progress made by each student in solving the programming exercises.

In terms of future works, the most important functionality that could be very helpful for both teachers and students is to implement an algorithm to capitalize on recurrent errors made by students. Also, this tool could be extended by adding automated feedback for students about the errors they make and the actions they miss.

## References

1. Becker, B. A., et al.: Compiler error messages considered unhelpful: the landscape of text-based programming error message research. In: 2019 ITiCSE Working Group Reports (ITiCSE-WGR 2019), 15–17 July 2019, Aberdeen, Scotland UK. ACM, New York, NY, USA, p. 34 (2019). <https://doi.org/10.1145/3344429.3372508>
2. Prieto-Blazquez, J., Herrera-Joancomarti, J., Guerrero-Rolda, A.: A virtual laboratory structure for developing programming labs. Int. J. Emerg. Technol. Learn. (iJET) **4**(2009), 47–52 (2009)
3. Di, C., Gang, Z., Juhong, X.: An introduction to the technology of blending-reality smart classroom. In: 2008 International Symposium on Knowledge Acquisition and Modeling, Wuhan, pp. 516–519. IEEE Computer Society (2008)
4. Tissenbaum, M., Slotta, J.D.: A new framework for smart classroom research: co-designing curriculum, research and technology. In: Proceedings of the 9th International Conference on Computer Supported Collaborative Learning, Rhodes, Greece: International Society of the Learning Sciences, pp. 91–93 (2009)
5. Broisin, J., Venant, R., Vidal, P.: Lab4CE: a remote laboratory for computer education. Int. J. Artif. Intell. Educ. **27**(1), 154–180 (2017)



# Towards Embodied Wearable Intelligent Tutoring Systems

Injila Rasul<sup>(✉)</sup>, Francisco Castro, and Ivon Arroyo

University of Massachusetts Amherst, Amherst, USA  
irasul@umass.edu, {fcastro,ivon}@cs.umass.edu

**Abstract.** Embodied cognition posits that human-environment interaction positively impacts thinking and learning, making it a valuable pedagogical tool. Technology in teaching and learning has seen tremendous maturation, such as the development of Intelligent Tutoring Systems (ITS). However, most ITS provide static learning experiences that do not incorporate embodiment, movement, and interaction with the space around the learner. This paper examines the results of using an embodied tutoring system across three case studies with different dimensions of embodiment. In all cases, we found trends highlighting how embodied tutoring systems can support learning. We also discuss different ways to incorporate embodiment into future research on ITS.

**Keywords:** Embodied Cognition · Intelligent Tutoring Systems · Math

## 1 Introduction and Motivation

*Embodied learning* has emerged as a powerful pedagogical approach for mathematics education, with mounting evidence about its essential role in developing and strengthening mathematical thinking skills [2, 7, 10]. At the core of embodied learning is the notion of *embodied cognition*, which asserts that knowledge develops from our understanding of the interactions between ourselves, our environment, and elements within the environment. This suggests that incorporating sensorimotor processes in learning activities can strengthen learning through learners' use of their bodies, reflection, and social interaction [7, 16].

Intelligent Tutoring Systems (ITS), however, particularly within mathematics education, have traditionally been screen-based (i.e., students use a computer and answer math problems on screen—e.g., MathSpring [4, 11]) and do not support learners in engaging in embodied learning practices. As new theories of learning evolve, with more significant evidence supporting the incorporation of embodiment, there is also a need for ITS to adapt. For example, ITS that can be accessed through mobile devices, enabling movement and exploration within physical environments, is one way of re-imagining embodied ITS. Developing technologies that support mathematics learning through environments that enable embodied interactions is thus a key area for future development [1, 6].

We present three case studies that analyze the effectiveness of embodiment in tutoring systems. We explore different kinds of embodied interaction in digital learning environments by comparing traditional tutoring systems with a tutoring system that incorporates embodiment.

## 2 Related Work

**Embodied Cognition and Pedagogy.** *Embodied cognition* is the idea that perception of the environment, and the body's interactions with elements within it, influence cognition [2, 16]. *Embodied learning* emphasizes that learning is supported and bolstered when the deep connection between the mind and the body [12] is employed as a pedagogical tool [16]. A key aspect of embodied cognition is *spatial cognition*, the cognitive process involved in spatial reasoning and problem-solving. Spatial cognition is significant to STEM pedagogy because concepts in mathematics and science incorporate ideas from the physical world around us, such as position, direction, relative speed, and spatial perception, among others. Prior research has shown gains in STEM performance when leveraging spatial cognition as a pedagogical tool [14, 16], which supports the case for instruction that incorporates spatial cognition, such as visuospatial skills (the ability to recognize visual and spatial relationships among objects [16]), as a means to improve conceptual understanding in STEM education.

**Making the Abstract Concrete.** The potential of embodied learning leads to the question: *how can embodiment be incorporated into ITS effectively to support STEM learning?* Weisberg and Newcombe [16] presented a framework that outlines mechanisms through which this is possible: (a) linking sensorimotor information with abstract concepts, (b) using gestures as a cognitive tool, (c) developing cognitive skills such as spatial cognition, (d) offloading onto the body/environment to free up cognitive processing, and (e) creating and understanding visual representations. *Offloading* refers to storing information in the environment/body without further burdening their mental resources, focusing cognitive capacity on problem-solving, inference-making, and communicating ideas [16]. An example is a student offloading a shape's mental rotation onto a physical object's rotation to retain a visual representation of its direction and orientation, enabling them to move on to another part of the problem. Creating visual representations is also critical in communicating ideas and seeing the bigger picture while problem-solving. Additionally, analogy-building between sensorimotor and abstract concepts (such as moving along a number line while adding numbers; a number line does not concretely exist in the natural world) and visual representations (i.e., a number line drawn on the floor) supports the learning of abstract concepts, using the physical world around learners [10, 16].

**Embodiment as Support for Mathematical Ideas.** Tools and manipulatives can illustrate, communicate, and test concepts and serve as testing grounds for emerging mathematical ideas, where the mind can impose mathematical relationships on the objects [15]. According to Walle, there are five ways to represent mathematical ideas: (1) manipulative models, (2) pictures, (3) real-world

situations, (4) written symbols, and (5) translations between and within representations. These representations help develop new mathematical concepts [15].

Prior research suggests that mathematics learning is embodied, based on perception and action, grounded in the physical environment and that different gestures suggest different types of thinking [6, 10]. Gestures with hands and body, which are “inherently spatial” because they are expressed in the space that a person occupies/perceives [16], become mechanisms through which cognition is communicated and information is encoded [2]. Prior work suggests that gestures are evidence of involvement in mathematical thinking [2] and may be a vital aspect of learning as students work with their peers and around physical objects.

**Intelligent Tutoring Systems and Embodied Cognition.** Given the impact of embodied pedagogy on STEM learning, designing ITS to incorporate embodiment is a good path forward for improving ITS [6, 7]. Technology that can support and enhance active learning environments has been studied previously, with increased learning gains compared to control groups [7]. However, most ITS are currently largely screen-based and do not incorporate movement. The ones that do often focus heavily on gesture-based movements rather than engaging the entire body [7]. It is crucial to re-conceptualize how ITS can provide embodied pedagogical support for STEM classrooms.

### 3 The WearableLearning Platform and Embodied Games

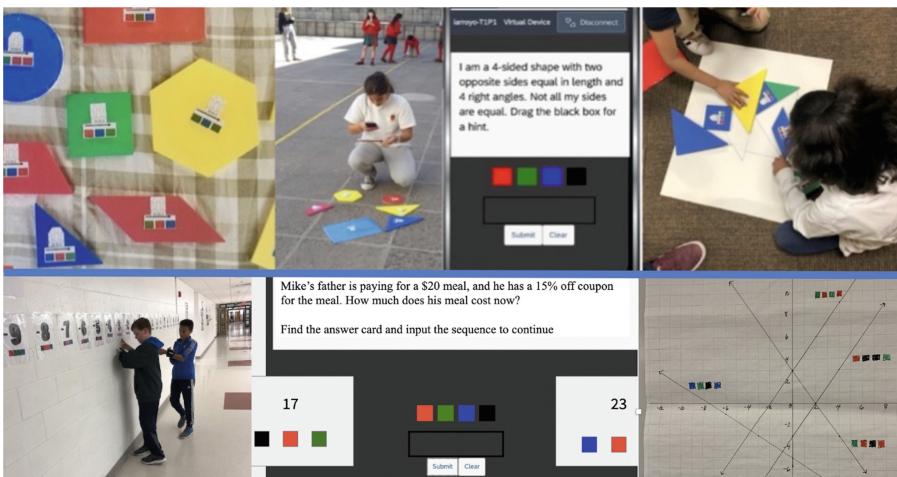
Our team developed WL<sup>1</sup> a web-based platform that enables users to play physically active, embodied, educational math games using mobile devices that students carry [3]. Thus, WL games can be situated within physical classrooms and incorporate movement, physical manipulations, and teamwork. Within the WL interface (Fig. 1), students can input answers to question prompts to receive feedback, request hints, and proceed through the game. Students can be organized into multiplayer teams, with each member taking on a specific role throughout the game, which may be complementary to other team members.

#### 3.1 WearableLearning Case Studies

**Case 1: WL Games.** Three teacher-created games, “Integer Hopscotch”, “Let’s Shop”, and “What’s my Line?”, were played with groups of 49 (6th grade), 84 (7th grade), and 102 (8th grade) students, respectively. These games focused on concepts of integer addition/subtraction operations, price percentage calculations, and graphs. *Integer Hopscotch* uses word problems on addition/subtraction of integers to prompt students to walk along a 20-foot number line to find the correct answer. *Let’s Shop* uses word problems based on shopping scenarios, where students need to find the correct price or discount for shopping items (e.g., tips, final sale price, amount of tax, see Fig. 1). Teams calculate the answer and locate it on color-coded cards scattered in the classroom, like a scavenger hunt. *What’s*

---

<sup>1</sup> WearableLearning: <http://wearablelearning.org/>.



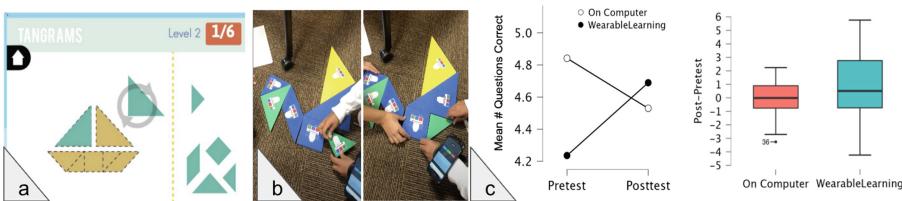
**Fig. 1.** (Top) *Tangrams Race*: shapes with color combination codes; a student retrieving a piece guided by WL on a phone; a math question in *Tangrams Race*; students building a *Tangrams* puzzle with retrieved pieces. (Bottom) Left: *Integer Hopscotch*; Middle: screenshot of the *Let's Shop* game; Right: a *What's my Line?* game object.

*My Line?* is a slope-equation and graph-matching game that involves the slope-intercept form of equations to challenge students to read and interpret equations. Teams match equations in WL with the corresponding line on a paper graph, with several other lines drawn as distractions (see Fig. 1), at one of seven different stations. Students race through all stations to correctly match each equation to its line graph and draw the correct lines.

*Data Collection.* Pre- and post-tests assessed student content knowledge. Students playing Integer Hopscotch were tested on integer addition and subtraction. Students playing Let's Shop were tested on percentages. Students playing What's my Line? were tested on equations.

*Results.* For *Integer Hopscotch*, students improved significantly from pre-test ( $M=.62$ ,  $SD=.26$ ) to post-test ( $M=.79$   $SD=.17$ ),  $t=6.39$ ,  $p<.001$ , after one class period. For *Let's Shop*, students improved significantly from pre-test ( $M=.57$ ,  $SD=.22$ ) to post-test ( $M=.67$ ,  $SD=.18$ ),  $t=3.12$ ,  $p=.002$ , after one class period. For *What's My Line?*, students improved significantly from pretest ( $M=.48$ ,  $SD=.27$ ) to posttest ( $M=.77$ ,  $SD=.27$ ),  $t=7.15$ ,  $p<.001$ , after one class period.

*Findings.* All games led to statistically significant learning gains, and while this study did not have a control condition, the results suggest that even a short exposure to embodied learning experiences may yield promising learning gains. The embodied learning experiences in these experiments provided support for teaching abstract ideas through offloading onto physical representations; for example, in *Integer Hopscotch*, the abstract number line is physically represented and students hop and count a certain number of steps to get to the correct answer. In



**Fig. 2.** (a) Control condition: Virtual manipulation of Tangrams pieces; (b) Embodied condition: Physical manipulation of Tangrams pieces by hand; (c) Mean change from students' pre- to post-tests in embodied vs. control.

*What's My Line?*, students worked on visual representations by plotting lines for equations. The students used abstract concepts, such as operations on negative numbers; we suspect that perhaps embodying the concept helped them visualize more concretely through a spatial representation (e.g., why the subtraction of two negative numbers may yield a negative number) rather than rote-learn sign conventions without developing meaningful understanding.

**Case 2: Tangrams Race.** *Tangrams Race* is a geometry-puzzle game focusing on classification based on angles/sides and parallel/perpendicular lines. In WL, it is played as a relay race for multiple three-player teams that requires each team to acquire the correct tangram pieces to create a specific shape arrangement. To acquire the pieces, student teams are positioned on one end of a room and each team player takes turns running to the opposite end of the room to retrieve tangram pieces from a container based on a prompt shown on the WL interface (e.g., "What is a shape with three sides?"). The prompts are mathematical descriptions of the desired tangram pieces (e.g., Fig. 1), that match math common core skills that students need to practice. Each piece has a color code that students enter into the WL interface (Fig. 1), which provides feedback.

Third to fifth-grade students were randomly assigned to one of two game conditions: (1) *Physical* — physically played tangrams (using WL and physical pieces, 54 students) or (2) *Digital* (control, 21 students) — digitally played tangrams. Both groups received equivalent instruction (similar math problems, help, hints, and time). The control group used a digital tutoring system where students answered problems through multiple choice or short answers. They can also ask for hints and construct the same puzzles virtually on the screen through a 'virtual Tangrams' website that allows rotating/dragging/dropping tangram pieces to form the puzzle (Fig. 2). An advantage to the control condition was that all math problems were available to students. In contrast, in the embodied condition, only a third of the problems were available to each player because of their complementary roles (i.e., each player looked for a specific tangram piece).

**Data Collection.** Pre- and post-tests assessed students' learning gains, using the pre-test as a baseline measure for students' content knowledge. The assessments contained the same ten content-related questions from the Massachusetts Comprehensive Assessment System (MCAS) 4th-grade math standardized exam.

*Results.* From pre- to post-test, *Embodied* condition students increased by 14.7% and *Digital* condition students decreased by 4.1%. Even though the learning gain of students in the Digital condition was the lowest (-0.31 learning gain for the Digital group versus 0.45 for the Embodied group), there was no significant effect for condition ( $F=0.638$ ,  $p=0.42$ ), however, Fig. 2c shows a trend, highlighting how students in the *Embodied* condition improved on average in comparison to students in the *Digital* condition.

*Findings.* Tangrams Race, as designed on WL, has many embodied affordances such as: physically manipulating pieces, constructing visual imagery, collaborative manipulation, and gesturing (Fig. 1). Before students run to the container of tangram shapes, they are given a description of the shape from the WL interface on their phones, enabling students to conduct visual imagery by visually hypothesizing in their minds what the shape could look like. If a student gets stuck and the hints are not helping, a teacher may help by gesturing to show clues, such as what parallel or perpendicular lines look like. The WL Tangrams Race game is structured to provide support for embodied learning, so it was surprising that the results we obtained were marginally significant trends.

We do not think that students “unlearned” in the *Digital* condition, but instead believe that the post-test scores in both groups in general are deflated, because, in our experience, students do not want to take post-tests, while they are very focused during pre-test time, due to the novelty of the situation.

Despite the advantage of more content, the control condition fared worse than the experimental condition, on average. In contrast, despite exposure to lesser content, the students from the *Physical* condition improved more in comparison.

**Case 3: EstimateIt!** *EstimateIt!* is a scavenger hunt where student teams search for volumes described in WL (e.g., the mobile device could show: “*Look for a sphere with a 6” radius*” or hints, see Fig. 3). Students are provided an unmarked 12-inch ruler for measurement estimations. Objects have a color code; after finding the correct object, the code is input into WL, to get feedback.

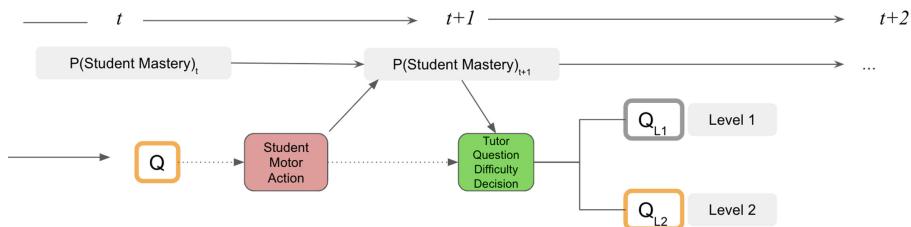
*Data Collection.* 49 students in an afterschool program participated. The experimental condition received hints with images of hand gestures that students should enact (Fig. 3); the control condition received textual hints that described ways to estimate. Pre- and Post- tests were conducted with both groups.

*Results.* Experimental condition students improved more ( $M=0.64$ ,  $SD=0.16$  on pre-test;  $M=0.68$ ,  $SD=0.17$  on post-test, n.s.) than those in the control condition ( $M=0.67$ ,  $SD=0.12$  on pre-test;  $M=0.68$ ,  $SD=0.11$  on post-test, n.s.).

*Findings.* Students in the experimental condition showed a trend of improvement, while those in the control condition did not improve. This trend suggests that incorporating aspects of embodiment, such as gestures, may be an effective way for tutoring systems to use embodiment to support learning. Students used gestures to denote relative and absolute size for measurement and estimation



**Fig. 3.** Examples of hints using hand gestures, Part-Part Whole, and size of an inch.



**Fig. 4.** An embodied ITS can work similarly to a non-embodied ITS, only that the questions (Q) are embodied quests that require motor action.

when discussing with their teammates. Using an embodied mode of communication for mathematical concepts may support students' analogy-building and gesturing, enriching their problem-solving and communicating strategies.

## 4 Conclusion

We presented three cases of how digital versus physical games in mathematics education can potentially impact student learning, and how a game-based tutor that uses mobile devices can leverage embodiment for learning. The first case showed that playing active games for different math concepts consistently had a positive impact on math learning, despite short exposure time. In the second and third case, the controlled studies revealed non-significant trends that favored embodied conditions, where the experimental condition was more "embodied". As the pre- and post-tests were given immediately before and after the games, their learning gains and trends of improvement suggest that the embodied-technology activities likely contributed to students' improvement.

**Limitations.** One limitation is that the case studies showed trends but non-significant effects. We suspect that the non-significant effects may be due to assessment instruments being paper & pencil (as opposed to having an embodied activity that reflects their learning through the game), failing to capture deeper modes of embodied thinking that are not necessarily reflected on multiple choice questions. For instance, if we assessed students' knowledge of triangles

through an embodied activity of cutting a right triangle with scissors and paper, this could potentially provide more meaningful insight into students' conceptual understanding. We aim to find ways to capture such differences in engagement in the future. Another limitation is that exposure time to the intervention was short, thereby constraining the impact that the activities could have brought about. Lastly, the experiments were deployed in after-school programs instead of math classes in formal classrooms. After-school environments are noisy due to looser student attendance regulations than in formal classrooms. Ideally, the intervention could have been deployed in schools led by math teachers within a math class, as this would be more suitable context than an after-school program.

**Future Implications: Embodied and Game-based ITS.** WearableLearning is not an ITS in a traditional sense yet, as it does not incorporate AI nor personalize instruction. Based on our work, we believe the following should be characteristics of embodied intelligent tutoring systems:

*Assessing Student Knowledge in Embodied ITS.* Students progress through WL games by finding, arranging, or manipulating physical objects in the environment and submitting answers into WL. Behind the scenes, each question has associated meta-data, such as the mathematics skill (standard) associated with a question. Success or failure at completing math tasks provides some evidence of students' math knowledge. Therefore, a traditional Bayesian Estimation of students' knowledge based on correctness, errors, and hint requests can be computed [5]; this is however dependent on students being exposed to enough questions regarding the same math skill (e.g., same Math Common Core standard).

*Changing Item Difficulty in Embodied Intelligent Games.* Lester et al. [9] described how, in Crystal Island, a Dynamic Decision Network can negotiate the narrative state of the game, the story world state, and the student user state. We envision a similar way of modeling students' knowledge and adjusting items in the games to satisfy the game narrative's requirements and constraints and the level of difficulty for each student player (Fig. 4). While the game continues to flow temporally in a horizontal direction, the intelligent game engine can update estimates of mastery, and make decisions in a vertical direction, making choices between questions or items of various difficulties (e.g., easy, medium, and hard) for the same "meta-level" math standard involved in the question.

*Embodied Hints and Support.* Teachers naturally incorporate gestures for illustrating and visualizing concepts as they communicate mathematical ideas [2]. Given the embodied nature of these learning tasks, gesture-based hints should be incorporated into embodied ITS. Underlying Case 3 described above is the work of Harrison et al. [8], who performed a deep cognitive task analysis of children's gestures and actions during measurement tasks. Using data collected from this study, Valente [13] created hints aimed at conveying motor strategies observed in college students and described the most common mistakes of elementary students during measurement tasks. Providing embodied hints introduces students to motor strategies to deploy during problem-solving (Fig. 3).

Future work will involve bringing WL to more math classrooms, creating more games aligned to mathematics standards, refining existing games, refining

our assessments for capturing math learning, and further analyzing the importance of motion, gesture, and action by collecting data on learners' movements.

## References

1. Abrahamson, D., Tancredi, S., Chen, R., Flood, V., Dutton, E.: Embodied design of digital resources for mathematics education: theory, methodology, and framework of a pedagogical research program. In: *Handbook of Digital (Curriculum) Resources in Mathematics Education*. Springer, Cham (2021)
2. Alibali, M.W., Nathan, M.J.: Embodiment in mathematics teaching and learning: evidence from learners' and teachers' gestures. *J. Learn. Sci.* **21**(2), 247–286 (2012)
3. Arroyo, I., Closser, A.H., Castro, F., Smith, H., Ottmar, E., Micciolo, M.: The WearableLearning platform: a computational thinking tool supporting game design and active play. *Technol. Knowl. Learn.*, 1–10 (2022). <https://doi.org/10.1007/s10758-022-09601-1>
4. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K., Rai, D., Tai, M.: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *Int. J. Artif. Intell. Educ.* **24**, 387–426 (2014)
5. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**, 253–278 (1994)
6. Eisenberg, M.: Embodiment as a strategy for mathematics education. In: *Proceedings of the International Conference on Interaction Design and Children* (2009)
7. Georgiou, Y., Ioannou, A.: Embodied learning in a digital world: a systematic review of empirical research in K-12 education. In: Díaz, P., Ioannou, A., Bhagat, K.K., Spector, J.M. (eds.) *Learning in a Digital World. SCI*, pp. 155–177. Springer, Singapore (2019). [https://doi.org/10.1007/978-981-13-8265-9\\_8](https://doi.org/10.1007/978-981-13-8265-9_8)
8. Harrison, A.: For good measure: identifying student measurement estimation strategies through actions, language, and gesture. In: *The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020*, vol. 2 (2020)
9. Lester, J.C., Ha, E.Y., Lee, S.Y., Mott, B.W., Rowe, J.P., Sabourin, J.L.: Serious games get smart: intelligent game-based learning environments. *AI Mag.* **34**(4), 31–45 (2013)
10. Link, T., Moeller, K., Huber, S., Fischer, U., Nuerk, H.C.: Walk the number line—an embodied training of numerical concepts. *Trends Neurosci. Educ.* **2**(2), 74–84 (2013)
11. McLaren, B.M., Nguyen, H.: Digital Learning Games in Artificial Intelligence in Education (AIED): A Review
12. Stoltz, S.A.: Embodied learning. *Educ. Philos. Theory* **47**(5), 474–487 (2015)
13. Valente, R.C.: Teaching students mathematical embodiment techniques using online learning game platform (WLCP). In: *Interactive Qualifying Project. Worcester Polytechnic Institute E-Projects Library* (2019)
14. Wai, J., Lubinski, D., Benbow, C.P.: Spatial ability for stem domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **101**(4), 817 (2009)
15. de Walle, J.V., Karp, K.S., Bay-Williams, J.M.: *Elementary and Middle School Mathematics: Teaching Developmentally* (2012)
16. Weisberg, S.M., Newcombe, N.S.: Embodied cognition and STEM learning: overview of a topical collection in CR: PI. *Cogn. Res. Principles Implications* **2**, 1–6 (2017)



# iQUIZ!: A Collaborative Online Learning System that Promotes Growth Mindset Using Persuasive Feedback

Mehnuma Tabassum Omar<sup>(✉)</sup> , Nafisul Kiron<sup>(✉)</sup> , and Julita Vassileva<sup>(✉)</sup>

University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada

{mehnuma.omar, ni.kiron}@usask.ca, jiv@cs.usask.ca

**Abstract.** Research has shown that learners learn best when they learn actively. In collaborative learning, students take charge of learning and work with others to achieve a common goal. It helps them experience seeking and synthesizing knowledge effectively by forming and planning their research and study strategies to reach their goals. Along with collaborative learning, a growth mindset helps them to believe that their effort and hard work can help them to succeed. To help students learn and adopt a growth mindset, we propose an online learning system called “iQUIZ!”, that uses persuasive strategies (*praise, simulation, self-monitoring, similarity, suggestion, and reminder*) to give students feedback to foster a growth mindset and motivate the students to practice active participation in a collaborative learning environment.

**Keywords:** Collaborative Learning · Growth Mindset · Persuasive Strategies

## 1 Introduction

Collaborative learning fosters higher levels of thought, information retention, and critical thinking than individual learning [1]. Collaborative learning means that students work together to achieve a common learning objective, taking the initiative for their learning and that of their fellow students [1]. A collaborative online learning system (COLS) is a platform that enables learners to interact with each other and with the learning materials synchronously or asynchronously. COLS can foster knowledge co-construction among learners, enhancing their learning outcomes and satisfaction. However, COLS also poses challenges, such as learners’ lack of motivation, engagement, and self-regulation. One way to address these challenges is to promote a growth mindset among learners, which is the belief that one’s abilities and intelligence can be improved through effort and learning. A growth mindset can help learners overcome difficulties, embrace challenges, and persist in their learning goals because they believe their effort will determine their achievement [2]. According to the theory of mindsets, learners with a growth mindset see intelligence as a trait that can be changed with effort and dedication, in contrast to those with a fixed mindset who see intellect as a fixed and unalterable trait [3].

Fostering a growth mindset has numerous benefits in education, including raising motivation levels among students and bridging social class divides while enhancing

academic achievement [4]. Furthermore, as a growth mindset inspires students to believe that their effort and hard work will ultimately pay off, they do not retreat when facing distractions and challenges. Therefore, practicing a growth mindset can help them adopt an effective learning behaviour that leads to success.

Mindsets serve as a mental framework that directs how people perceive, sense, and act when attempting to achieve a goal [5]. Unlike personality traits such as the big five, mindsets are domain-specific and can be developed with a single intervention [6]. For example, constructive feedback is vital in fostering a growth mindset. Students' feedback from their teachers or parents influences their attitudes and mindsets, impacting their performance. For example, although the conventional method of encouraging students to adopt a growth mindset is through praise, praising incorrectly can also have the opposite impact. The literature has demonstrated that praising talent promotes a fixed mindset among students, whereas praising meaningful effort promotes a growth mindset [7]. Therefore, providing students with appropriate feedback is essential to inspire them to make the effort needed to reach their objectives.

One possible approach to promote a growth mindset in a COLS is using persuasive messages/nudges. Designed using principles of social psychology and strategies for behaviour change from Persuasive Technology, these messages/nudges seek to alter users' attitudes, behaviours, and views without coercion [8]. Many studies show successful attitude and behaviour change examples applying persuasive messages in domains such as healthcare and education [8].

We propose an online collaborative learning system, called "iQUIZ!" that can be used in a blended learning environment, and promotes a growth mindset by providing performance-relevant feedback using persuasive messages. "iQUIZ!" is a wiki-based online quiz question dataset. The students can curate and edit the existing questions and post new questions related to the class material. By providing students with the option to create their questions and to modify the existing ones to improve their clarity, and possible answers, including the correct one and the distractors. The system allows them to rehearse the course material, critically examine the questions and construct new ones and thus allows them to utilize their conceptual structures on the learning task. To encourage a growth mindset, "iQUIZ!" uses feedback messages/nudges implementing six persuasive strategies [9]: *praise, simulation, self-monitoring, similarity, suggestion, and reminder*. We have chosen these six strategies in line with the results of previous research, indicating that they are widely used and successful in many domains and contexts [10]. We will carry out a field study in a University class and expect that by utilizing our approach, we will be able to increase student motivation, engagement, and the quality of their contributions to "iQUIZ!", as well as their academic achievement in the class. As a next step in our research, we will personalize the application of persuasive strategies by training a model based on the learner's demographics and personality features to amplify the effectiveness of the persuasive messages in encouraging a growth mindset, learner motivation, quality of contributions to "iQUIZ!", and academic success in the class.

## 2 Related Work

Several studies have shown that mindset is linked to educational outcomes. Students with a growth mindset accept challenges to develop their skills, and rather than giving

up; they persevere through those challenges. Despite the difficulties and challenges, they commit more time to their intended work, which enhances the likelihood that they will complete the academic task effectively [11]. According to L. B. Limeri et al., a positive loop exists between a growth mindset and academic success [12]. Previous research by E. Rhew, J. S. Piro, P. Goolkasian, and P. Cosentino has demonstrated that participants' motivation in the adolescent special education program changes significantly after receiving a growth mindset intervention [13]. Several educational tools employ the growth mindset concept as a form of intervention or application. One such program is Brainology, a computer program that Carol Dweck developed at Stanford University [14]. The program comprises tasks and challenges with the fundamental postulate of a growth mindset – that intelligence can be developed through individual effort and persistence [14]. Educational games with the growth mindset concept have been proposed as well. For example, to enhance student perseverance and growth mindset behaviours, the Brain Points game explicitly rewards effort, strategic use, and incremental advancement [7]. In addition, using Motion Math and teaching elementary math emphasizes growth mindset messages, especially for students vulnerable to failure [15].

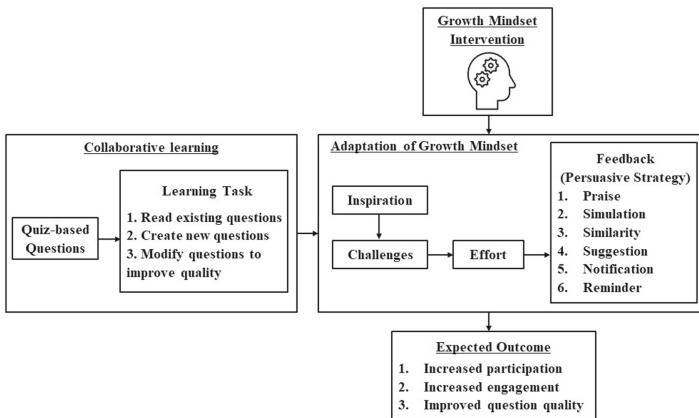
The environment in which students are immersed can impact their perspectives [16]. In collaborative learning, various learners with various mindsets (growth or fixed) collaborate and share their beliefs with the other group members, which can influence the task at hand [16]. Research [17] has shown how peer mindsets influence an individual's perception of a task's value. The study recommended including peers as one of the essential components in creating a growth mindset intervention.

Persuasive strategies can be used independently or with pedagogical strategies to engage students and encourage desired learning behaviour. For example, researchers [18] have used Web 2.0 technologies to create a persuasive system to enhance students' online learning behaviour. Others have integrated persuasive technology in a multi-disciplinary project named HANDS (Helping Autism-Diagnosed young people Navigate and Develop Socially) [19], which aids in the development of social and self-management skills of young children (aged 11–16) with high-functioning ASD (autism spectrum disorders). The work described in [20] integrates Gardner's Multiple Intelligence with persuasive strategies (*similarity, tailoring, and credibility*) to motivate students to develop their reading and writing skills.

To summarize, persuasive strategies can be applied to nurturing intended learning activities. Therefore, we propose to utilize feedback messages implementing persuasive strategies to promote a growth mindset in the target learning tasks – editing peer's questions and creating new questions to share with peer learners. Encouraging students to develop a growth mindset with persuasive techniques will improve their engagement, performance, and achievement in the learning activity.

### 3 Proposed Collaborative Online Learning System

We developed an application called “iQUIZ!”, a web-based collaborative learning system that supports students in editing quiz questions using a wiki. The system uses feedback messages using six persuasive strategies to foster a growth mindset among students, guiding them in active participation. The conceptual model of the “iQUIZ!” tool is shown in Fig. 1. The detailed model is described as follows:



**Fig. 1.** Conceptual Model of “iQUIZ!”

### 3.1 Collaborative Learning

**Quiz-based Questions.** The participants of the “iQUIZ!” tool will be students taking a face-to-face, blended, or entirely online class. “iQUIZ!” is a wiki-based platform that will provide an interface for the students to curate existing questions related to their course topic from the question database. The initial question database was collected from another game-based learning tool called the Tower of Questions (ToQ) [21]. The questions are quiz-type, true-false, multiple-choice, or short-answer questions.

**Learning Task.** The goal of the “iQUIZ!” tool is to assist students in learning course material with their peers in a collaborative setting. The students can review their peers’ work and help them as necessary. Students can perform three main tasks: *practice*, *edit*, and *add*. In the *practice* task, students can view and practice answering existing questions that are posted by previous students and newly added questions posted by their classmates using the question database. Students can modify existing questions by doing *edit* tasks to improve their quality. Finally, in the *add* task, students can post new and unique questions to the question database for their peers.

### 3.2 Growth Mindset Intervention

The purpose of the intervention is to inform the students that their goal might be achieved effectively with hard work and effort, making them stronger and more self-assured. “iQUIZ!” starts the intervention by introducing the basic definition of a growth mindset and then playing a video presenting an in-depth explanation of a growth mindset: how mindsets are formed, how a growth mindset can be achieved, the benefits, and why practicing a growth mindset is needed in “iQUIZ!”. The goal of this intervention is that students learn how to utilize “iQUIZ!” and the task activities and become aware of the concept of mindsets and the difference between growth and fixed mindset.

### 3.3 Adaptation of Growth Mindset

Learners can pursue self-development using four phases: inspiration, challenges, effort, and constructive feedback or criticism [2]. The “iQUIZ!” tool incorporates these four phases to perform its task. Each phase is described below:

**Inspiration.** The first phase for successfully performing a task involves the student thoroughly understanding the task they are being asked to complete. Providing students with examples of the required actions for the task is essential, so they can draw inspiration from the examples and develop original responses individually. The students can use the *practice* task in “iQUIZ!” to test their knowledge of the course material by answering the existing questions. They can also test their ability to ask questions about the material by viewing the questions created by other students. They can get an idea of the question structure from the available questions and attempt to generate questions with similar structures, forms, and topics for the database for the *add* task. As “iQUIZ!” is a collaborative learning environment, the students can also get inspiration for critical questions modifications from their peers’ modification style.

**Challenges.** Students should specify learning goals and responsibilities to complete a task successfully while focusing on self-development, which promotes learning ownership. Although examples or inspiration might assist students in comprehending how to do a task, being innovative and accepting challenges can help students develop their skills. For example, in the *edit* task, students can modify existing questions by fixing minor typos or grammatical errors or changing sentence structure. An *edit* that improves the quality of a question, referred to as a necessary modification, is more complex than a typical *edit*. This modification aims to enhance the question’s quality and make a question format acceptable for a wider audience. Another challenging task is to *add* new unique questions. A unique question should differ significantly from those already in the database (not just paraphrased). Initially, creating unique questions is relatively easy as the number of questions is low. Nevertheless, coming up with unique questions gets more challenging as time goes on. “iQUIZ!” encourages students to be creative. By accomplishing these tasks, students can attain learning goals and a step toward developing a growth mindset through their effort.

**Effort.** Students should independently utilize their effort to perform learning tasks while following the learning process. In “iQUIZ!”, students can add minor modifications and simple questions to the question database. However, their highest effort can lead them to perform advanced changes and new unique questions. They can use effort to achieve the learning objective, a fundamental requirement of a growth mindset, and then encourage their peers to do the same. This effort can also improve the quality of the “iQUIZ!” material for the upcoming participants.

**Feedback.** Constructive feedback or criticism can direct students’ thoughts and the growth of their learning process [2]. A growth mindset requires that students’ efforts be connected to results rather than just being praised for their work [4]. Students should be encouraged to concentrate on methods and materials that can be used to enhance learning when they fail to achieve a goal. In “iQUIZ!”, we have utilized six persuasive

strategies: *praise*, *simulation*, *self-monitoring*, *similarity*, *suggestion*, and *reminder*; to provide feedback messages to the students. The *praise* message can encourage the students through words or visuals in response to their behaviour. A *simulation* feedback message can present the impact of students' involvement or actions in the "iQUIZ!" in forms of visual effects. The *self-monitoring* strategy can convey feedback by including students' progress. *Suggestion* strategy in feedback can give students valuable options for carrying out the desired "iQUIZ!" activity. We can deliver feedback using a *similarity* strategy by reminding students of their connection with "iQUIZ!". The *reminder* strategy reiterates the "iQUIZ!" tool's goal and to adopt a growth mindset.

## 4 Conclusion and Future Work

This paper presents a conceptual model of a web-based collaborative learning tool named "iQUIZ!" that fosters a growth mindset among students with the help of persuasive strategies. This tool will allow students to help each other learn course materials through quiz-based questions. They can practice questions created by their friends and add new unique questions. In addition, students can create high-quality content in the underlying course material by critically modifying existing questions. The "iQUIZ!" supports active learning, requiring students to think of good unique questions and critical modifications of the existing ones. The students will get persuasive messages encouraging them to adopt a growth mindset or use their utmost effort to participate in the "iQUIZ!". We believe these messages will motivate the students to actively engage in the "iQUIZ!" which will enable them to learn the material and adequately prepare for academic assessments and tests.

At present, we are planning a pilot study with a first-year university-level programming course. Students will be divided into two groups: control (without persuasive feedback messages) and experiment (with persuasive feedback messages). We will measure the student's engagement and participation as well as the quality of the edits and new questions submitted by the students to see if there are any differences between efforts and performances between the two groups. We will use the log data to determine the students' engagement in "iQUIZ!" such as how much time they spent curating and creating content. The number of active participants in each group will determine the number of participations in the tool. We will review the quality of the contributions manually. We will also collect data about the students' performance in the mid-term and final exams. Our study concludes with a post-study survey to determine students' experience, satisfaction and motivation after using "iQUIZ!". We will analyze the data to find correlations between the different metrics and possibly use the data to train a model to predict the effectiveness of using specific persuasive strategies in "iQUIZ!" for promoting a growth mindset, motivation, and better performance in individual students. This will help personalize the system. The growing set of curated quiz questions will also be used in future deployments of "iQUIZ!" and a source of training and exam questions for the teachers.

## References

1. Laal, M., Laal, M.: Collaborative learning: what is it? Procedia – Soc. Behav. Sci. **31**, 491–495 (2012). <https://doi.org/10.1016/j.sbspro.2011.12.092>

2. Nuntasri, W., Chaichomchuen, S.: Development of positive thinking repository by improving growth mindset through project-based learning in a digital environment. In: ACM International Conference Proceeding Series, pp. 236–241. Association for Computing Machinery (2020). <https://doi.org/10.1145/3395245.3396444>
3. Dweck, C.S.: Self-Theories. Psychology Press (2013). <https://doi.org/10.4324/9781315783048>
4. Stohlmann, M.: Growth mindset in K-8 STEM education: a review of the literature since 2007. *J. Pedagogical Res.* **6**(2), 149–163 (2022). <https://doi.org/10.33902/JPR.202213029>
5. Keating, L.A., Heslin, P.A.: The potential role of mindsets in unleashing employee engagement. *Hum. Resour. Manag. Rev.* **25**(4), 329–341 (2015). <https://doi.org/10.1016/j.hrmr.2015.01.008>
6. Burnette, J.L., et al.: A growth mindset intervention: enhancing students' entrepreneurial self-efficacy and career development. *Entrepreneurship Theor. Pract.* **44**(5), 878–908 (2019). <https://doi.org/10.1177/1042258719864293>
7. O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C.S., Popović, Z.: Brain points: A growth mindset incentive structure boosts persistence in an educational game. In: Conference on Human Factors in Computing Systems – Proceedings, pp. 3339–3348. Association for Computing Machinery (2014). <https://doi.org/10.1145/2556288.2557157>
8. Orji, F., Vassileva, J.: Personalized Persuasion for Promoting Students' Engagement and Learning E-Commerce Trust View project Persuasive Technology View project. <https://www.researchgate.net/publication/330658206> (2018)
9. Oinas-Kukkonen, H., Harjumaa, M.: Persuasive systems design: key issues, process model, and system features. *Commun. Assoc. Inf. Syst.* **24**(1), 485–500 (2009). <https://doi.org/10.17705/1cais.02428>
10. Oyebode, O., Ndulue, C., Mulchandani, D.: Tailoring persuasive and behaviour change systems based on stages of change and motivation. In: Conference on Human Factors in Computing Systems – Proceedings, Association for Computing Machinery (2021). <https://doi.org/10.1145/3411764.3445619>
11. Kizilcec, R.F., Goldfarb, D.: Growth mindset predicts student achievement and behavior in mobile learning. In: Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019. Association for Computing Machinery, Inc. (2019). <https://doi.org/10.1145/3330430.3333632>
12. Limeri, L.B., et al.: Growing a growth mindset: characterizing how and why undergraduate students' mindsets change. *Int. J. STEM Educ.* **7**(1), 1–19 (2020). <https://doi.org/10.1186/s40594-020-00227-2>
13. Rhew, E., Piro, J.S., Goolkasian, P., Cosentino, P.: The effects of a growth mindset on self-efficacy and motivation. *Cogent Educ.* **5**(1), 1492337 (2018). <https://doi.org/10.1080/2331186X.2018.1492337>
14. Donohoe, C., Topping, K., Hannah, E.: The impact of an online intervention (Brainology) on the mindset and resiliency of secondary school pupils: a preliminary mixed methods study. *Educ. Psychol. (Lond.)* **32**(5), 641–655 (2012). <https://doi.org/10.1080/01443410.2012.675646>
15. Boaler, J., Bryant, C., Klein, J.: Promoting a Growth Mindset Through Personalized Mindset Coaching A Promising Experiment with Motion Math (2017). <https://www.youcubed.org/wp-content/uploads/2017/03/Motion-Math-final.pdf>. Accessed 18 Feb 2023
16. Sato, M.: Mindsets and language-related problem-solving behaviors during interaction in the classroom. *Innov. Lang. Learn. Teach.* **16**(3), 265–276 (2022). <https://doi.org/10.1080/17501229.2021.1895803>
17. Sheffler, P.C., Cheung, C.S.: The role of peer mindsets in students' learning: an experimental study. *Br. J. Educ. Psychol.* **90**(S1), 17–34 (2020). <https://doi.org/10.1111/bjep.12299>

18. Widyasari, Y.D.L., Nugroho, L.E., Permanasari, A.E.: Persuasive technology for enhanced learning behavior in higher education. *Int. J. Educ. Technol. High. Educ.* **16**(1), 1–16 (2019). <https://doi.org/10.1186/s41239-019-0142-5>
19. Mintz, J., Aagaard, M.: The application of persuasive technology to educational settings. *Educ. Tech. Res. Dev.* **60**(3), 483–499 (2012). <https://doi.org/10.1007/s11423-012-9232-y>
20. Lucero, A., Zuloaga, R., Mota, S., Muñoz, F.: Persuasive technologies in education: improving motivation to read and write for children. In: IJsselsteijn, W.A., de Kort, Y.A.W., Midden, C., Eggen, B., van den Hoven, E. (eds.) PERSUASIVE 2006. LNCS, vol. 3962, pp. 142–153. Springer, Heidelberg (2006). [https://doi.org/10.1007/11755494\\_20](https://doi.org/10.1007/11755494_20)
21. Kiron, N., Adaji, I., Long, J., Vassileva, J.: Tower of questions (TOQ): A serious game for peer learning. In: Liapis, A., Yannakakis, G. N., Gentile, M., Ninaus, M. (eds.) GALA 2019. LNCS, vol. 11899, pp. 276–286. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-34350-7\\_27/TABLES/4](https://doi.org/10.1007/978-3-030-34350-7_27/TABLES/4)



# Helping Teachers to Analyze Big Sets of Concept Maps

Michele La Barbera<sup>1</sup>, Filippo Sciarrone<sup>1</sup> , and Marco Temperini<sup>2</sup>

<sup>1</sup> Faculty of Economics, Universitas Mercatorum, Piazza Mattei, 10, Rome, Italy  
[filippo.sciarrone@unimercatorum.it](mailto:filippo.sciarrone@unimercatorum.it)

<sup>2</sup> DIAG, Sapienza, University of Rome, Via Ariosto, 25, Rome, Italy  
[marte@diag.uniroma1.it](mailto:marte@diag.uniroma1.it)

**Abstract.** We present a framework devised to help teachers analysing a collection of learners' concept maps, with the aim to give priority to maps that could reveal learners' cognitive problems, irrespective of the actual validity of the maps. The framework offers a 3-D visualization of maps' features, that can be the more useful, the greater is the number of students. We present the framework, its measures, and its prototypical implementation. We investigated the perceived usefulness of the framework, and the cognitive traits that it could point out, through a questionnaire answered by a sample of teachers. The results are overall positive, and may allow for further study and experimentation.

**Keywords:** Concept Maps · Learning · Cognitive Traits

## 1 Introduction

Many years ago, in [16], Ridley and Novak discussed assessment methods and posed that *there's a question about purposes [of assessment]. It's hard to find assessment invoked without a parallel call for improvement, particularly improved teaching and learning.* They explained the efforts of researchers in education aiming to add a constructivist viewpoint (based on Ausubel's Theory) to the more traditional objectivist/behaviourist approach, in order to eventually offer different assessment methods for different views of learning, namely those where *learners construct knowledge by applying what they already know in order to make sense of new phenomena* [16].

Along the above guidelines, *Concept Mapping* [14] developed, as the educational strategy based on the use of Concept Maps (CMs) by (teachers and) students. CMs allow to organize and express one's knowledge and understanding about a given knowledge domain. Teachers use CMs to organize the flow of a course [2] and to provide students with a structure helping deepen their understanding of the topic [5]. Learners build CMs to draw their mental knowledge domain organization: the CM can then be used to analyse learner's cognitive traits and for the more mundane grading operation [1].

CMs are convenient in Technology Enhanced Learning applications, as they can be represented as a Direct Acyclic Graph (DAG), where nodes are concepts, and labeled edges show semantic relationships. This allows integrating CMs into computer-supported systems, and manage them in an automated or semi-automated manner. According to literature, there are several learner's cognitive aspects that can be fostered by concept mapping, and assessed, based on the concept maps produced by the learner. Examples are critical thinking [3], creativity [18], motivation/engagement [11], comprehension [4].

In this paper we define a framework that should support the teacher through a part of the analysis task of a set (potentially large) of CMs. In particular, the teacher would be helped to select for further analysis, the maps that might point out learners' learning problems. Such CMs would then have precedence in the further teacher's analysis, to verify the actual presence of cognitive problems, and define appropriate interventions. The basic idea is that the framework would not perform a semantic analysis of the maps, to assess their correctness; rather it would guide the teacher to look for the occurrence of cognitive problems in learners who produced maps with certain characteristics. To such maps the teacher might give precedence in her/his further analysis.

We assume the above support could be useful in any context of class (and map) cardinality; however, it is likely that it would be more and more useful when the class is fairly large (with students in the hundreds) or even huge (as in the case of a Massive Open Online Course - MOOC's, with students in the thousands).

In the following sections, we define a measure of entropy for a CM; on the basis of such measure we define a second measure, *ME*, that we surmise can render the "effort" exerted by the learner while building her/his CM. We also surmise that a visualization of *ME*, over a large number of CMs, can help the teacher decide what maps should have precedence in her/his analysis, whereas a small effort could be indicative of some learning problems. By "learning problems" we mean factors that hinder the learning capability of the learner, such as any lack of knowledge, difficulty to get knowledge, lack of motivation/engagement, and other kinds of disturbances of a person's learning processes. In summary, based on *ME* (notice, with no indication about the correctness of the CM), by the data computation and visualization here proposed, the teacher can be guided toward those maps that might unveil learning problems occurring in students; then the teacher can decide whether such problems do actually occur, and whether interventions are in order to help the learner.

We can't, so far, validate theoretically the effectiveness of the proposed help to teachers, so we asked what a sample of teachers perceive about the usefulness of the framework and about the learners' cognitive aspects that could be captured through the visualization, and implicit classification of their CMs.

In particular, the Research Questions (RQs) of interest are as follows.

**Preliminary RQ** Does  $H$  actually capture the teacher's idea of quality of learning.

**RQ1** Irrespective of the individual maps' correctness, could our visual classification be useful to identify subsets of CMs that could bring the evidence of cognitive problematic aspects?

**RQ2** What cognitive traits (or lack thereof) are more likely to be pointed out in the visual classification we portrayed?

For this investigation we used the current implementation of our framework, as a system providing simulated data: it computes a set of 1000 CMs, according to a distribution of number of concepts and number of links. The visualization of such data is then proposed to the sample. The answers suggest that our framework is perceived as useful.

## 2 Related Work

As far as we could discern, there are two general approaches to CM analysis and assessment, where the topological and semantic structure of the CM is analysed and some conclusions are drawn (about grading and interventions).

The most practiced approach is about analysing each learner's CM, either by automated analysis, or by a rubric guiding the teacher. In these cases, the validity of the map is measured, and also information about the cognitive evolution of the learner can be appreciated. As an assessment tool, usually the CM is traversed: correct concepts, links, sequencing (or different distribution) of connected concepts, and other factors, give origin to points that are accumulated on a final "grade".

The second general approach, that we think is way less applied, tends to produce an evaluation of the map from other viewpoints than its propriety and validity. In these cases, personal traits (cognitive as well) of the CM's author can be modeled.

About the first general approach, [15] introduced the concept itself of semantic scoring of a CM.

In [13] a "semantic scoring" rubric is presented, that allows the teacher to give scores to various aspects of a CM, and provide an assessment. Points (or, rather, six categorical values, e.g. low, or high or NA) are awarded against presence/lack of a basic concept, correct/wrong propositions, dependency relations among concepts (dynamic propositions), cross-links, cycles of concepts.

In [9], an holistic scoring, based on [12], assesses the learner's state of knowledge. In this measure, propositions' correctness is scored on a scale [0, 3], a correct concept awards 2 points, a hierarchy gives 10 points, a true proposition gets 2 points (precision), further points are given for the number of levels of a hierarchy, and cross-links are considered as well. An example of map assessment not directly aiming to grade the map's elements, is in [10]. The research starts from the accepted usefulness of CMs as enhancer of student's learning gain, in various disciplinary fields and school settings. Basically the paper sets to analyse maps against the different types of knowledge (of which a preliminary classification is in "procedural" and "conceptual"). A semantic plane is defined, along the axis of *semantic density* and *semantic gravity*, the categories of knowledge

represented by different map layouts (novice, theoretical, professional and practical knowledge) are placed in the four quadrants. In such layouts the archetypal map morphologies of *Spoke*, *Chain*, and *Network* [8] are differently recognizable in the quadrants.

In [7] the authors introduce the Entropy applied to a CM, emphasizing, by some case studies, its property to be related to the topology of a CM.

In [6], the use of concept maps supports the learning process in three phases. In the initial phase the available maps depict the initial state of knowledge of each student, so that the next educational steps can be prepared accordingly. Then also a phase of exchange and peer evaluation among students takes place.

The framework presented in this paper is deemed to help the teacher according to the second of the above described approaches. We don't compute, the correctness, or the grade of a CM. Rather, by the  $H$  and  $ME$  measures, we provide a visualization of the effort applied by each learner in producing her/his map, surmising that such effort can help unveil learning problems of the learner. In particular, a CM with low  $ME$  might suggest that the learner is incurring in learning problems. The framework just points out the possibility of problems; then, the decision about such problems and possible interventions is left in the hands of the teacher. This help can be useful in general, and specially when the teacher deals with a great number of CMs, such as in the case of MOOC's classes. We think that, by offering the teacher the mentioned guidance through the learners' CMs, our proposal adds to the current area of research about Teaching and Learning with Concept Maps, in that it presents a point of view that (to our knowledge) is so far disregarded.

### 3 The Entropy $H$ and the Mapping Effort $ME$

The formula to compute the Entropy  $H$  was borrowed from Shannon's mathematical information theory [17]. This quantity is a topological measure of the *degree of networking* of a CM.

**Definition 1.** Given a CM, We define the Entropy  $H$  associated to a generic node  $c_i$ ,  $H(c_i)$  as follows: 
$$H(c_i) = \sum_{j=1}^n \frac{1}{|Edge(c_j)_{out}|} \log_2 \frac{1}{|Edge_{out}(c_j)|}$$

where  $|Edge(c_i)|$  is the number of outgoing links connecting  $c_i$  to other concepts. If a concept  $c$  has no outgoing links, we assume:  $H(c) = 0$ . Consequently, we have that  $H = 0$  for a *chain* configuration, while  $CH = 2$  for a *spoke* configuration having a root concept directly connected to four other concepts.

**Definition 2.** Given a CM, We define the Entropy  $H(CM)$  associated to the entire CM as follows: 
$$CH(CM) = \sum_{i=1}^n CH(c_i)$$

**Definition 3.** Given a CM, We define  $M_{max}(CM)$  the maximum of  $H$  for a given CM

The  $H_{max}(CM)$  value for a given CM is: 
$$H_{max}(CM) = \log_2[(n - 1)!]$$
.

**Definition 4.** Given a CM, We define the relative Entropy  $H_r(CM)$  associated to the entire CM as follows:  $H_r(CM) = \frac{H(CM)}{H_{max}(CM)}$ .

Based on a CM's number  $|C|$  of concepts, and number  $|L|$  of links, we then define the  $ME(CM)$  value:

**Definition 5.** Given a CM, We define the Mapping Effort  $ME(CM)$  associated to a CM as follows:  $ME(CM) = H_r|C||L|$

$ME$  measures the student's effort, exerted on building the CM. In fact, as two maps, having different numbers of concepts and links, might share the same value for  $H_r$ , if two students' maps have the same value of  $H_r$ , the  $ME$  value adds important information about the difference between the maps/students: in particular about the number of links and the number of concepts they used. For instance, given *student1*, whose map has  $H_r = 0.20$  with 10 concepts and 15 relations, and *student2*, whose map has  $H_r = 0.20$  as well, but with 5 concepts and 7 relations, the  $ME$  values would be 30 for *student1*, and 7 for *student2*, denoting a higher effort for *student1*.

## 4 The System

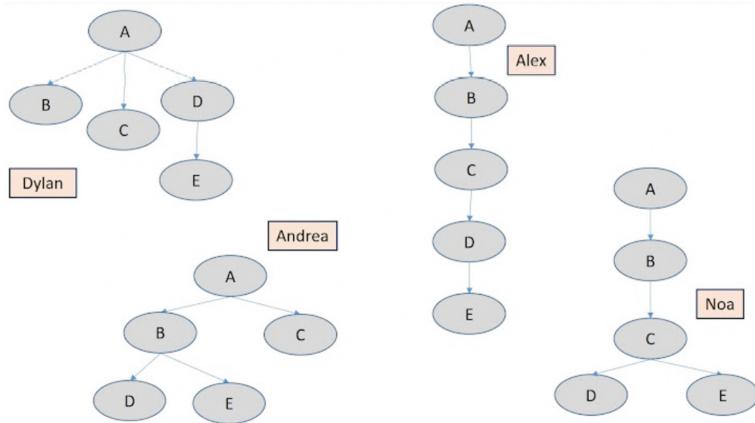
In this Section we briefly discuss the system that was used to evaluate our research questions. The system has a specific functionality, allowing to specify the features of a collection of simulated CMs (such the minimum and maximum number of concepts). The Python based back-end component generated a large number (1000) of randomly generated CMs. For each CM, three different measures are computed and shown: the *Entropy H*, the *Relative Entropy H<sub>r</sub>* and the *Mapping Effort ME*. The CMs are saved in a no-sql MongoDB database instance; they are displayed for the teacher, in three different 3-D scatter plots. In this way we obtained the images used in the questionnaire submitted to the teachers sample (cf. Sect.5).

## 5 Experimental Evaluation

In this section we explain the validation process of our Research Questions. It is comprised of two validation tasks: 1) Validation of the relationship between  $H$  and the learning process (the preliminary RQ), and 2) Validation of RQ1 and RQ2.

For the preliminary RQ, a questionnaire was administered to a sample of 25 high school teachers, randomly selected from a large set of teachers who, usually, use CMs as a didactic tool. The proposed question was: *Four students produced four concept maps about a given Knowledge Domain. Please give a mark to each map (min. 1, max 4) according to the level of student's learning witnessed by the map. To make the test less time expensive we did not ground the maps on a*

*particular Knowledge Domain, so we also avoided to give names to the concepts and relations: please do assume that the relations between concepts in the maps are valid.*



**Fig. 1.** The first questionnaire, about the perception of Entropy  $H$ .

Specifically, through this survey we wanted to verify whether the teacher would correlate the structure of a CM to the evidence of a student’s learning process (Fig. 1).

Basically the sample assessed the learning entailed by the maps in the same order of  $H$ . In particular, by using the t-test for two independent samples (as 25 people could be a sufficient number for using a parametric test), over the grades given to Andrea’s and Dylan’s CMs, we obtained a p-value  $p = 0.42$ , meaning that the two averages of the two distributions have no significant differences. However, the teachers’ evaluations of the four maps corroborates our preliminary hypothesis: the entropy  $H$  is perceived by the sample as related to the student’s learning process: The more networked is the map, the more the learning process could be better. Andrea’s and Dylan’s CMs show a small difference both in terms of  $H$  values (1.585 Vs. 2.0), and in the sample averages (3.48 Vs. 3.42).

For both RQ1 and RQ2, we used another sample, composed by a set of 20 units, randomly selected from a large set of teachers. The teachers’ disciplinary subjects were STEM (57.9%), Humanities (15.8%), Technical disciplines (10.5%), while 15.8% of the sample were teaching to students with disabilities.

A questionnaire was submitted to the sample through the Google Module platform<sup>1</sup> For the pictures in the questionnaire, we used the system briefly discussed in Sec. 4. By it, we first simulated a big class with 1000 CMs, built by different values for the number of concepts and relationships. The system computed for each CM the values  $H$ ,  $H_r$  and  $ME$ . Then these three variables were

<sup>1</sup> Available at: <https://forms.gle/85BMAVLN5ux3tk696>.

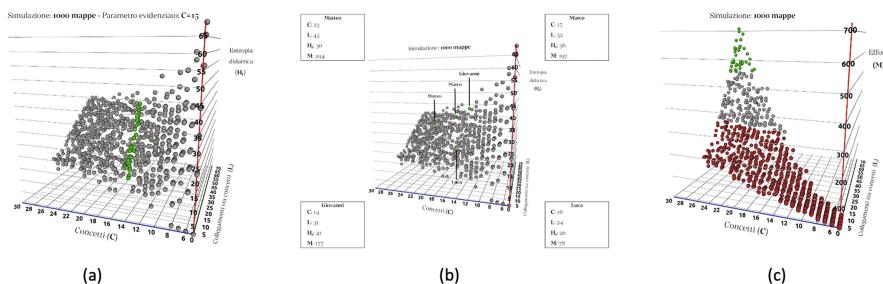
visualized in three different 3-D graphs (Fig. 2), and embedded into the questionnaire. An explanation of all the parameters was included in the instructions to clarify the purpose of the survey. Moreover, it was explicitly made clear in the instructions that no consideration was given to the semantic validity of CMs, whereas the CMs were classified only based on the above mentioned measures. (as explained in the Introduction, we are here focusing only on the perceived usefulness of our framework, as an help to point out possible learners' learning problems, with no information about the CMs correctness).

By the questionnaire, the sample was prompted to imagine a set of students who were assigned the following task: *Build a concept map to represent the K domain: use the concepts and connections between them that you deem necessary.*

Figure 2(a) shows the set of maps, classified according to the value of  $H_r$ : four of them are highlighted showing also the values  $H_r$ ,  $C$ ,  $L$ , and  $ME$ .

Figure 2(b) shows the set of the simulated CMs, highlighting those built with a given number of concepts (in this case  $C = 15$ ).

Figure 2 (c) shows the set of CMs, classified according to their value of  $ME$ . The points in the 3-D space are divided into three sets: the lowest set (in red), denotes maps with "low"  $ME$  values; the highest set (in green), corresponds to maps with "high"  $ME$  values. The other points denote CMs with intermediate  $ME$  values.



**Fig. 2.** The second questionnaire to check RQ1 and RQ2.

In order to validate the RQ1, the sample was asked to answer the question: *Do you think that the proposed visualization and classification, might be helpful to you in deciding which maps, or groups of maps, to prioritize in analysis, with the goal of undertaking appropriate instructional interventions on the students who produced them?* A 5-points Likert scale was used to classify the answers.

We obtained the following statistical parameters: mean  $\bar{x} = 3.65$  and standard deviation:  $s = 1.1$ . Assuming a normal distribution and applying an estimation based on the t-test, we have that the estimated mean  $\mu$  is in the range:  $3.22 \leq \mu \leq 4.07$ , with  $\alpha = 0.05$ .

About RQ2, the following question was used to check the correlation between the  $ME$  values and some significant cognitive aspects (still as perceived by the

sample): *In your opinion, the value of  $M$ , defined above, to which of the following cognitive aspects might be particularly related?* The cognitive aspects were selected from the CMs literature and were the following: *Understanding of the Knowledge domain, Motivation and involvement, Manual skills in CM building, Synthesis and critical thinking skills, and Creativity*.

Teachers had to express the perceived correlation between cognitive aspects and  $ME$ , using a 5-points Likert scale (from *None at all (Per niente* in the Questionnaire), through *Little (Poco)*, *Neutral (Neutro)*, *Fairly (Abbastanza)*, and *Much (Molto)*). Teachers related  $ME$  to all the proposed variables, but specially to Domain Understanding and Critical Thinking (highest score).

## 6 Discussion and Conclusions

We presented a framework, based on measures of Entropy and Effort, deemed to help the teacher in the endeavor of selecting CMs where it is more likely to unveil learning problems of the authors. In particular the framework does not provide information related to the actual correctness of the CMs: rather it points out, through visualization and classification of the maps, the above mentioned possibility of occurrence of learning problems, leaving to the teacher the further analysis and possible interventions.

A preliminary investigation strengthened the goodness of Entropy as a measure of quality of the student's learning: teachers' perception of the "CM's author learning" increases, as the CM's Entropy increases. This initial validation, albeit limited, has prompted us to investigate the use of this quantity more thoroughly, through a more extensive and in-depth experimentation.

However, in a setting where learners are requested to build a map freely, i.e. without starting from a fixed set of concepts/links,  $H$  could not discriminate well among different maps, so we introduced the  $ME$  measure, and conjectured that it could shed light on the presence of cognitive problems in a CM's author. This could then prompt the teacher to further analyse the CM and devise interventions, if that is the case.

Considering the answers given by the sample to the first question in the second questionnaire, the RQ1 gets a positive answer, as the sample's teachers deemed the proposed visualization useful to guide a preliminary phase of analysis of the maps, irrespective of the maps validity.

Then, considering the answers to the second question in the second questionnaire, also the RQ2 gets an encouraging answer, as we have seen that sample's teachers perceived  $ME$  as related to different cognitive aspects, and in particular to two of those we proposed.

As future developments we plan to release a web-based teacher support system for the analysis of CMs and an experimentation with real students.

## References

1. Anohina-Naumeca, A., Grundspenkis, J., Strautmane, M.: The concept map-based assessment system: functional capabilities, evolution, and experimental results. *Int. J. Continuing Eng. Educ. Life-Long Learn.* **21**(4), 308–327 (2011)
2. Asiksoy, G.: Computer-based concept mapping as a method for enhancing the effectiveness of concept learning in technology-enhanced learning. *Sustainability* **11**, 1005 (2019)
3. Barta, A., Fodor, L.A., Tamas, B., Szamoskozi, I.: The development of students critical thinking abilities and dispositions through the concept mapping learning method - a meta-analysis. *Educ. Res. Rev.* **37**, 100481 (2022)
4. Chang, K.E., Sung, Y.T., Chen, I.D.: The effect of concept mapping to enhance text comprehension and summarization. *J. Exp. Educ.* **71**(1), 5–23 (2002)
5. Cliburn, J.: Using concept maps to sequence instructional materials. *J. Coll. Sci. Teach.* **15**(4), 377–79 (1986)
6. Correia, P.: The use of concept maps for knowledge management: from classrooms to research labs. *Anal. Bioanal. Chem.* **402**(1), 1979–1986 (2012)
7. Fiume, A.F., Sciarrone, F., Temperini, M.: A new metric to help teachers unveil meaningful learning in concept maps. In: MIS4TEL 2022. Lecture Notes in Networks and Systems, vol. 580, pp. 65–74. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-20617-7\\_9](https://doi.org/10.1007/978-3-031-20617-7_9)
8. Hay, D., Kinchin, I., Lygo-Baker, S.: Making learning visible: the role of concept mapping in higher education. *Stud. High. Educ.* **33**(3), 295–311 (2008)
9. Ingeç, S.: Analysing concept maps as an assessment tool in teaching physics and comparison with the achievement tests. *Int. J. Sci. Educ.* **31**(1), 1897–1915 (2009)
10. Kinchin, I., Moellits, A., Reiska, P.: Uncovering types of knowledge in concept maps. *Educ. Sci.* **9**(2), 131–152 (2019)
11. Ma, A., Mm, A.: Using conceptual mapping for learning to affect students' motivation and academic achievement. *Sustainability* **13**(7), 4009–4029 (2021)
12. McClure, J., Sonak, B., Suen, H.: Concept map assessment of classroom learning: reliability, validity, and logistical practicality. *J. Res. Sci. Teach.* **4**(36), 475–492 (1999)
13. Miller, N., Cañas, A.: A semantic scoring rubric for concept maps: design and reliability. In: Concept Maps Connecting Educators, Proceedings of the Third International Conference Concept Mapping, pp. 60–67. Tallin University (2008)
14. Novak, J.: The importance of conceptual schemes for teaching science. *Sci. Teach.* **31**(6), 10–13 (1964)
15. Novak, J.D.: Learning How to Learn. Cambridge University Press, Cambridge (1984)
16. Ridley, D., Novak, J.: Assessing student learning in light of how students learn. In: The AAHE Assessment Forum (1988)
17. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948)
18. Tseng, K.H., Chang, C.C., Lou, S.J., Hsu, P.S.: Using creative problem solving to promote students' performance of concept mapping. *Int. J. Technol. Des. Educ.* **23**(1), 1093–1109 (2013)



# Learning Engagement and Peer Learning in MOOC: A Selective Systematic Review

Fatma Miladi<sup>(✉)</sup>, Daniel Lemire<sup>ID</sup>, and Valéry Psyché<sup>ID</sup>

TÉLUQ University, 5800 rue Saint-Denis, Montréal, QC H2S 3L5, Canada  
{fatma.miladi,daniel.lemire,valery.psyché}@teluq.ca

**Abstract.** Massive open online courses (MOOCs) improve learning; but their low completion rate remain problematic. Peer learning has been proposed as a method to increase learning engagement in MOOCs, thereby decreasing the dropout rate. However, the effectiveness of peer learning in promoting learning engagement in MOOCs remains underexplored. This systematic review aimed to examine the effect of peer learning on learning engagement in MOOCs. Eight articles met the inclusion and quality assurance criteria using the PRISMA method. We found that: (1) limited research has analyzed learner engagement within peer learning; (2) learner engagement can be measured through analysis of log, text, and survey data; and (3) peer learning can positively impact learning outcomes in MOOCs (completion rates, quiz completion rates, and quiz scores).

**Keywords:** MOOC · isolation · dropout · learner engagement · peer learning · peer interaction

## 1 Introduction

Massive open online courses (MOOCs) offer online learning opportunities to learners all over the world [1], giving them the flexibility to learn whenever and wherever they want [2]. Since the COVID-19 pandemic, MOOCs have become increasingly popular worldwide [3]. However, despite the high enrollment, the completion rate of MOOCs remains low [4]. On average, less than 10% of learners complete a MOOC [5, 6]. Many factors contribute to these high dropout rates, but one important factor that has a direct effect on learner dropout in MOOCs is the feeling of isolation [7, 8].

Feelings of isolation are common in fully online courses and MOOCs [9]. Geographical separation and lack of contact with fellow learners may lead to feelings of psychological isolation [10]. Feelings of isolation in completely online courses are not surprising. However, with an increased emphasis on blended learning in higher education, a sense of disconnection is frequent [11, 12].

One possible solution to the problem of isolation is to reinforce a sense of community [10, 13]. The notion of creating an online community for learning has been explored, suggesting that it can overcome the lack of physical contact between learners [14].

This systematic literature review aims to examine learning engagement using peer learning in MOOCs.

## 2 Related Works

The growing popularity of MOOCs has led to significant research on various aspects of MOOCs, including learner engagement and peer learning. This section provides an overview of the importance of peer learning in MOOCs and the fundamental description of learner engagement.

### 2.1 Peer Learning in MOOC

Peer learning is an educational method, where students interact with each other to acquire new knowledge and skills [15, 16]. Generally, MOOCs can provide discussion forums that enable learners to share their opinions and seek clarification on materials [17]. Furthermore, MOOCs often incorporate a peer assessment feature that allows learners to provide feedback and rate their peers' work as a way to learn from each other. Peer learning in MOOCs has demonstrated various benefits, such as increasing learner engagement [14, 18], which can potentially reduce the dropout rate of MOOCs.

### 2.2 Learning Engagement in MOOC

Learner engagement is defined as the behavioral, cognitive, emotional, and social connections that MOOC participants make with the course content, the instructor, and other learners [19]. Specifically, behavioral engagement refers to students' observable actions and their participation and involvement in educational activities [20].

Cognitive engagement refers to psychological investment in learning [17]. This is reflected in learners' efforts to acquire complex content or skills during the MOOC learning process [21].

Emotional engagement encompasses the positive and negative reactions demonstrated in learning [22]. It refers to students' emotional connections with institutions, instructors, peers, and the MOOC content itself [20].

Social engagement is centered on learner-learner and learner-instructor interactions [19]. Given that engagement can be considered a type of behavior, it is sometimes viewed as a subcategory of the behavioral engagement category. However, in many studies, social engagement is regarded as a fundamental component of students' perceptions and is measured separately from behavioral, cognitive, and emotional engagement [19].

This study investigated learning engagement using peer learning by reviewing eight studies on learner engagement in MOOCs published between 2016 and 2023. The research questions were as follows:

- RQ1: What research methods are used to quantify learner engagement using peer learning in MOOCs?
- RQ2: Does peer learning improve learning outcomes in MOOCs?

## 3 Methods

To answer the above questions, a systematic review was conducted based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020)

framework that guided this study [23]. PRISMA framework provides comprehensive tools for reporting transparent systematic review such as checklists and flow diagrams [23].

### 3.1 Search Strategy

Multiple databases were used to ensure a comprehensive search of the scientific literature, including Scopus, ERIC, and ProQuest. These databases were carefully selected based on their relevance to the research questions and their capacity to provide a diverse range of scholarly resources.

A search was conducted using a range of search keywords paired with the Boolean operators AND and OR. The keywords were selected to locate any studies that focused on the research questions. The broad search string used in this study was designed to be comprehensive and included keywords related to peer learning, MOOCs, and learner engagement [19]. Specifically, the search string used was as follows: ((“peer learning” OR “peer recommender” OR “social interaction”) AND MOOC AND (“learner engagement” OR “student engagement” OR “behavioral engagement” OR “cognitive engagement” OR “emotional engagement” OR “social engagement”)). In total, 20 papers containing the search string within the title and abstract were identified in the search.

### 3.2 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were used to select suitable studies for this review. The articles were selected based on the following inclusion criteria: (1) published between 2016 and 2023, (2) published in journals or conferences that have gone through a peer-review process, and (3) related to the impact of peer learning on learner engagement in MOOC.

Books, review articles, dissertations, and posters were excluded, as were studies that were published before 2016, were published in non-peer-reviewed journals or conferences, and were not related to behavioral engagement, or cognitive engagement or emotional engagement or social engagement. The summary of the inclusion and exclusion criteria is shown in Table 1.

### 3.3 Selection Process

The article selection process consisted of three main stages: identification, screening, and inclusion. The PRISMA flow diagram in Fig. 1 provides an overview of this process.

During the identification stage, a specific publication period from January 2016 to March 2023 was set to limit the search timeframe. To ensure a high-quality standard, the search was restricted to peer-reviewed journal or conference articles. Lastly, to eliminate any redundancy or duplication, all duplicate copies of the studies were removed, reducing the search to 14 papers.

The second stage was screening. The first screening was conducted based on the title, abstract, and keywords to ensure that the selected articles were relevant to the impact of peer learning on learner engagement in MOOCs. This reduced the number of studies

to 11. After conducting the initial screening of potential studies, a second screening process was conducted to assess each study's eligibility more deeply based on the full text of the article. The objective of this second screening was to verify that the study was related to behavioral engagement, or cognitive engagement or emotional engagement or social engagement. Finally, eight relevant articles were selected as studies included in our systematic literature review.

**Table 1.** Inclusion and exclusion criteria.

	Inclusion criteria	Exclusion criteria
Publication Year	Papers published from January 2016 to Mars 2023	Papers published before 2016
Source Types	Peer-reviewed journal or conference papers	Non-peer-reviewed journal or conference papers (e.g., books, editorials, dissertations, reviews)
Topic	The impact of peer learning on learner engagement in MOOC	Not related to behavioral engagement, or cognitive engagement, or emotional engagement or social engagement

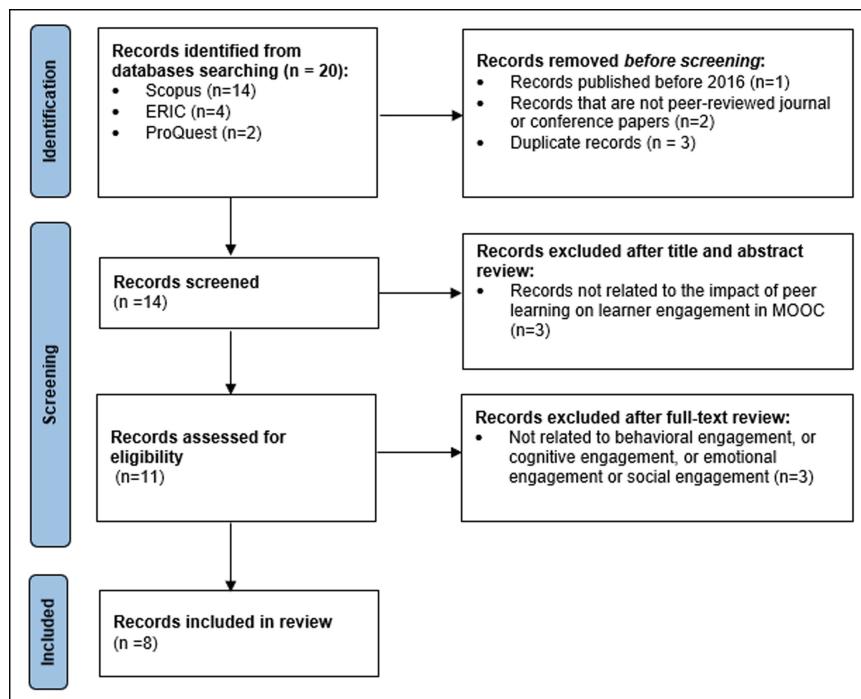
## 4 Results

Previous studies have identified four distinct dimensions of learner engagement: behavioral, cognitive, emotional, and social engagement [19]. Table 2 summarizes the articles discussed in this review.

### 4.1 RQ1: What Research Methods Are Used to Quantify Learner Engagement Using Peer Learning in MOOCs?

The three research methods commonly used in the literature are qualitative, quantitative, and mixed methods [32]. Qualitative methods involve non-numerical data such as interviews and observations to obtain descriptive data. Quantitative methods use tools, such as surveys, experiments, and statistical analysis, to gather numerical data. The mixed method combines both qualitative and quantitative methods for a more comprehensive understanding.

An analysis of eight relevant studies revealed that six used quantitative methods to quantify learner engagement in MOOCs [18, 25–27, 29, 30]. For example, [26] used the frequency of visits and posts made on a discussion board to measure learners' social engagement. The results indicate that learners who engaged in one-on-one discussions improved their social engagement. [27] analyzed learner engagement from both social and cognitive aspects based on forum text data, including discussion posts and comments. Specifically, they applied a social network analysis method to identify different social



**Fig. 1.** PRISMA Flow Diagram [23]

**Table 2.** Synthesis of the information extracted from the studies.

Authors	Research method	Data type	Learner engagement				Impact on learning outcomes
			Behavioral	Emotional	Cognitive	Social	
[24]	Mixed	Log, Text	-	-	-	X	Yes
[25]	Quantitative	Log	-	-	-	X	Yes
[26]	Quantitative	Log	-	-	-	X	Yes
[27]	Quantitative	Text	-	-	X	X	No
[28]	Qualitative	Text	-	X	-	-	No
[29]	Quantitative	Survey	X	X	-	-	No
[30]	Quantitative	Log	X	-	-	X	No
[18]	Quantitative	Log	X	-	-	X	Yes

participatory roles such as leaders, starters, influencers. Regarding the cognitive aspect, the researchers used the content analysis approach to determine low, medium, and high levels of contributions to individual knowledge inquiry (IKI) and group knowledge construction (GKC). [25] used posts, comments, replies, and follows to investigate social

engagement among learners in the forum. Their findings indicated that most learners who followed someone actively participated in discussions by writing comments or replying to someone else's comments. [29] used a survey to identify emotional and behavioral engagement of learners. The results showed that commitment was positively related to learner engagement. Finally, in [18] and [30], the authors used K-means clustering on log data to measure learners' behavioral and social engagement in MOOCs. The log data included metrics such as the number of pages the student accessed, the number of quizzes completed, the number of posts on discussions, and the average length of discussion posts. The researchers demonstrated that peer recommendations based on sociodemographic similarities increased learners' engagement in MOOC.

[28] used a qualitative method to measure the emotional engagement of learners. Specifically, they analyzed comments made by learners and found both positive and negative attitudes towards peer reviews.

[24] demonstrated that peer discussion on the forum can enhance social engagement among learners. They employed a mixed-methods research approach using both quantitative and qualitative analyses. The quantitative analysis involved examining the number of discussion forum participants and the number of forum postings per week. The findings revealed that a total of 29,355 posts were generated over a period of 8 weeks by 1691 participants. The qualitative analysis, on the other hand, showed that most of these discussions went beyond superficial comments, and none of them deviated from the focus question. Additionally, the study indicated that the learners were highly engaged in the discussions, even after the course ended, leading towards an evolving community of practice.

## 4.2 RQ2: Does Peer Learning Improve Learning Outcomes in MOOCs?

Of the eight studies analyzed, four highlighted the impact of peer learning on learning outcomes in MOOCs in terms of completion rates, quiz completion rates, and quiz scores.

[24] found that the average completion rate of the MOOCs studied was significantly higher (36.35%) than the average completion rates reported in the literature. In addition, their results showed that active learners had a completion rate of 59.38% on average.

[25] demonstrated that learners who fully participated in a forum were more likely to complete the course. Furthermore, they indicated that learners who followed someone in a MOOC were more likely to complete the course. Specifically, their findings revealed that over 30% of followers completed the course by completing at least half of the steps.

[26] found that students who engage in one-on-one discussions show improvement in their quiz completion rates by 7% to 10%, as well as a 2% to 10% increase in quiz scores in subsequent weeks. Similarly, [18] found that engaging in discussions improved learners' quiz completion rates and quiz scores.

## 5 Discussion and Conclusion

This systematic literature review investigates learner engagement in MOOCs through peer learning. Learner engagement is a complex phenomenon to understand; however,

it is considered a critical factor in fostering learning and success [31]. The concept of engagement is multifaceted and is based on how students behave, feel, think, and interact with one another. To obtain a comprehensive understanding of learner engagement, researchers should examine all four dimensions of learner engagement: cognitive, emotional, behavioral, and social engagement. Focusing solely on one dimension of engagement may restrict the understanding of learner engagement. Thus, measuring all dimensions is crucial for obtaining deeper insights into learner engagement.

Research should adopt both qualitative and quantitative methods to analyze learner engagement [32]. Using both methods can provide a more comprehensive understanding of learner engagement. Several studies have adopted a mixed methods approach to examine student engagement in MOOCs [24, 33].

Researchers should test the impact of the four dimensions of engagement (cognitive, emotional, behavioral, and social engagement) and multiple outcomes of interest. Specifically, they should explore the relationship between engagement and key learning-related factors such as retention [34] and course performance [26]. By doing so, researchers can gain a more holistic understanding of the learning process in MOOCs.

To conduct a comprehensive exploration of learning engagement in MOOC forum discussions, researchers should examine the interactive relationships between the various dimensions of learner engagement, such as emotional and cognitive engagement [35]. Through this examination, researchers can better comprehend how these dimensions are interrelated and how they jointly impact learning achievement [35].

## References

1. Gallego-Romero, J.M., Alario-Hoyos, C., Estévez-Ayres, I., Delgado Kloos, C.: Analyzing learners' engagement and behavior in MOOCs on programming with the Codeboard IDE. *Educ. Tech. Res. Dev.* **68**(5), 2505–2528 (2020). <https://doi.org/10.1007/s11423-020-09773-6>
2. Shen, J., Yang, H., Li, J., Cheng, Z.: Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimedia Syst.* **28**(2), 469–478 (2021). <https://doi.org/10.1007/s00530-021-00854-x>
3. Liu, C., Zou, D., Chen, X., Xie, H., Chan, W.H.: A bibliometric review on latent topics and trends of the empirical MOOC literature (2008–2019). *Asia Pac. Educ. Rev.* **22**(3), 515–534 (2021). <https://doi.org/10.1007/s12564-021-09692-y>
4. Kizilcec, R.F., et al.: Scaling up behavioral science interventions in online education. *Proc. Natl. Acad. Sci.* **117**(26), 14900–14905 (2020). <https://doi.org/10.1073/pnas.1921417117>
5. Jordan, K.: Initial trends in enrolment 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>
6. Yin, S., Shang, Q., Wang, H., Che, B.: The analysis and early warning of student loss in MOOC course. In: Proceedings of the ACM Turing Celebration Conference-China, pp. 1–6 (2019). <https://doi.org/10.1145/3321408.3322854>
7. Dalipi, F., Imran, A.S., Kastrati, Z.: MOOC dropout prediction using machine learning techniques: review and research challenges. In: 2018 IEEE Global Engineering Education Conference (EDUCON), pp. 1007–1014 (2018)
8. 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.comedu.2016.03.016>

9. Bulger, M., Bright, J., Cobo, C.: The real component of virtual learning: motivations for face-to-face MOOC meetings in developing and industrialised countries. *Inf. Commun. Soc.* **18**(10), 1200–1216 (2015). <https://doi.org/10.1080/1369118X.2015.1061571>
10. Croft, N., Dalton, A., Grant, M.: Overcoming isolation in distance learning: building a learning community through time and space. *J. Educ. Built Env.* **5**(1), 27–64 (2015). <https://doi.org/10.11120/jebc.2010.05010027>
11. Poon, J.: Blended learning: an institutional approach for enhancing students' learning experiences. *MERLOT J. Online Learn. Teach.* **9**(2), 271–289 (2013)
12. Smyth, S., Houghton, C., Cooney, A., Casey, D.: Students' experiences of blended learning across a range of postgraduate programmes. *Nurse Educ. Today* **32**(4), 464–468 (2012). <https://doi.org/10.1016/j.nedt.2011.05.014>
13. Rovai, A.P.: Building sense of community at a distance. *The Int. Rev. Res. Open and Distrib. Learn.* **3**(1), 1–16 (2002). <https://doi.org/10.19173/irrodl.v3i1.79>
14. Liu, S.: Student interaction experiences in distance learning courses: a phenomenological study. *Online J. Distance Learn. Adm.* **11**(1), 1–20 (2008)
15. Boud, D., Cohen, R., Sampson, J.: *Peer Learning in Higher Education: Learning From and with Each Other*. Routledge, London (2014). <https://doi.org/10.4324/9781315042565>
16. Topping, K.J.: Trends in peer learning. *Educ. Psychol.* **25**(6), 631–645 (2005). <https://doi.org/10.1080/01443410500345172>
17. Almatrafi, O., Johri, A., Rangwala, H.: Needle in a Haystack: identifying learner posts that require urgent response in MOOC discussion forums. *Comput. Educ.* **118**, 1–9 (2018). <https://doi.org/10.1016/j.compedu.2017.11.002>
18. Labarthe, H., Bouchet, F., Bachelet, R., Yacef, K.: Does a peer recommender foster students' engagement in MOOCs? In: 9th International Conference on Educational Data Mining, pp. 418–423 (2016)
19. Deng, R., Benckendorff, P., Gannaway, D.: Learner engagement in moocs: scale development and validation. *Br. J. Edu. Technol.* **51**(1), 245–262 (2020). <https://doi.org/10.1111/bjet.12810>
20. Jimerson, S.R., Campos, E., Greif, J.L.: Toward an understanding of definitions and measures of school engagement and related terms. *Calif. Sch. Psychol.* **8**, 7–27 (2003)
21. Jung, Y., Lee, J.: Learning engagement and persistence in massive open online courses (MOOCS). *Comput. Educ.* **122**, 9–22 (2018). <https://doi.org/10.1016/j.compedu.2018.02.013>
22. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**, 59–109 (2004)
23. Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* **88**, 105906 (2021). <https://doi.org/10.1016/j.ijsu.2021.105906>
24. Shah, V., Murthy, S., Warriem, J., Sahasrabudhe, S., Banerjee, G., Iyer, S.: Learner-centric MOOC model: a pedagogical design model towards active learner participation and higher completion rates. *Educ. Tech. Res. Dev.* **70**(1), 263–288 (2022). <https://doi.org/10.1007/s11423-022-10081-4>
25. Sunar, A.S., White, S., Abdullah, N.A., Davis, H.C.: How learners' interactions sustain engagement: a MOOC case study. *IEEE Trans. Learn. Technol.* **10**(4), 475–487 (2016)
26. Allon, G., Van Mieghem, J., Zhang, D.: Does social interaction improve learning outcomes? field evidence from massive open online education. *The Field Exp. Website* **9**, 347–367 (2016)
27. Ouyang, F., Li, X., Sun, D., Jiao, P., Yao, J.: Learners' discussion patterns, perceptions, and preferences in a Chinese massive open online course (MOOC). *Int. Rev. Res. Open Distrib. Learn.* **21**(3), 264–284 (2020). <https://doi.org/10.19173/irrodl.v21i3.4771>
28. 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>

29. Sun, Y., Guo, Y., Zhao, Y.: Understanding the determinants of learner engagement in MOOCs: an adaptive structuration perspective. *Comput. Educ.* **157**, 103963 (2020). <https://doi.org/10.1016/j.compedu.2020.103963>
30. Bouchet, F., Labarthe, H., Yacef, K., Bachelet, R.: Comparing peer recommendation strategies in a MOOC. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 129–134 (2017). <https://doi.org/10.1145/3099023.3099036>
31. Kahu, E.R.: Framing student engagement in higher education. *Stud. High. Educ.* **38**(5), 758–773 (2013). <https://doi.org/10.1080/03075079.2011.598505>
32. Creswell, J.W., Plano Clark, V.L., Gutmann, M.L., Hanson, W.E.: Advanced mixed methods research designs. *Handb. Mixed Methods in Soc. Behav. Res.* **209**(240), 209–240 (2003)
33. Lan, M., Hew, K.F.: Examining learning engagement in MOOCs: a self-determination theoretical perspective using mixed method. *Int. J. Educ. Technol. High. Educ.* **17**(1), 1–24 (2020). <https://doi.org/10.1186/s41239-020-0179-5>
34. Xiong, Y., Li, H., Kornhaber, M.L., Suen, H.K., Pursel, B., Goins, D.D.: Examining the relations among student motivation, engagement, and retention in a MOOC: a structural equation modeling approach. *Global Education Review* **2**(3), 23–33 (2015)
35. Liu, S., Liu, S., Liu, Z., Peng, X., Yang, Z.: Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Comput. Educ.* **181**, 104461 (2022). <https://doi.org/10.1016/j.compedu.2022.104461>



# Personalized Study Guide: A Moodle Plug-in Generating Personal Learning Path for Students

Ted Krahn<sup>1</sup>, Rita Kuo<sup>2</sup> , and Maiga Chang<sup>1</sup>

<sup>1</sup> Athabasca University, Edmonton, AB, Canada

[maiga.chang@gmail.com](mailto:maiga.chang@gmail.com)

<sup>2</sup> Utah Valley University, Orem, UT, USA

**Abstract.** Applying learning analytics to generate personalized learning paths is getting popular in recent research. This study designs a Moodle plug-in called Personalized Study Guide that can generate personalized learning paths according to students' learning styles. The plug-in supports two ways to determine students' learning styles: one is using Inventory of Learning Styles (ILS), the well-established learning style questionnaire developed by Felder and Silverman, and the other is through the analysis of students' past behaviour patterns on Moodle. The Personalized Study Guide analyzes the learning-style weightings for each learning resources/activities in the course. Using the calculated learning-style weightings, the Personalized Study Guide could determine which learning resources/activities are closer to students' learning styles to generate the recommended learning path. The research team is conducting the experiment to evaluate the perceived usefulness of the Personal Study Guide in 2023. The details of the evaluation plan are also described in this study.

**Keywords:** Recommendation System · Moodle Plug-in · Learning Style · Behaviour Pattern · Behaviour Analytics Block

## 1 Introduction

Learning analytics is used in analyzing learners' behaviour and the learning content to enhance learners' performance and the learning environment [5]. Learning Management System (LMS) is one of the most used platforms for the learning analytic research [7]. There are some existing learning analytics tools that are capable of extracting students' behaviour patterns on Moodle, e.g., Behaviour Analytics Moodle Plug-in [9] and Learning Object Relation Discovery (LORD) Moodle Plug-in [10].

The research team aims to use the existing behaviour analytics tool accompany with the Index of Learning Styles to generate a recommended learning path for students so instead of one-fit-all general study guide they can have their personalized study guide while learning according to the learning behaviour of the other students who share similar learning styles. Section 2 introduces the definition of the learning styles and its related studies. The system architecture of the Personal Study Guide is described in Sect. 3.

Section 4 demonstrates the system workflow of students using Personal Study Guide in learning. An evaluation plan examining the usability of Personal Study Guide and the effectiveness of using the existing behaviour analytics to identify student's learning style is illustrated in Sect. 5. Section 6 makes a brief conclusion.

## 2 Learning Styles

Students may have different preferences in receiving and processing information in the learning environment [11, 12]. Researchers start investigating how to categorize students' preferences in learning into different learning styles [2]. For example, Jackson's Learning Styles Profiler (LSP) is based on the personality to determine whether students are initiators, reasoners, analysts, or implementers. Each personality has the strengths and weaknesses based on the different preferences so the students can use the strengths to become more effective learners [6]. Kolb proposed students with different learning styles – which are diverging, assimilating, converging, and accommodating – are more comfortable with different stages in the learning circle; in this case, the students can identify which stages of the learning circle are their lesser preferred and strengthen these stages to enhance their learning outcome [8].

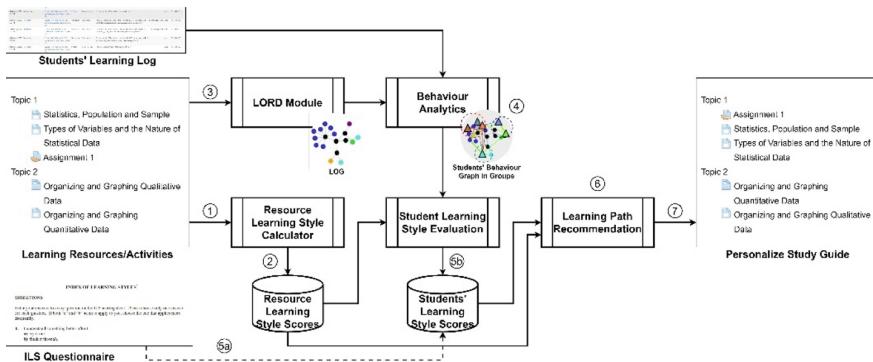
Felder and Silverman's learning- and teaching-style model (FSLSM) distinguish learners' learning style into each of the following four dimensions, which are information processing (active or reflective learners), information reception (visual or verbal learners), information perception (sensing or intuitive learners), and information understanding (sequential or global learners [3]). To access students' preference, the Index of Learning Styles (ILS) was developed as a web-based instrument [4]. ILS is a 44-item questionnaire – 11 questions in each dimension – where each item has two options sit in the two side of the learning style in the dimension. For example, in the *information processing* dimension, if the student selects the active option, the *active* preference is increased with a value of 1, and the *reflective* preference is decreased with a value of –1. At the end, the preference of each learning style is ranged between –11 to + 11.

Investigating the relationship between learning objects and learning styles decides how to deliver different content to students according to their learning styles [12]. In Assis and colleagues' research, 39 studies were included in the systematic literature review to discover the relationship between the learning object types and the learning styles in FSLSM. There are 46 types of learning objects identified from the past literature reviews and weighting of each learning styles corresponding to the learning objects types are calculated based on the number of occurrences that the learning objects has been connected to the learning styles [1].

## 3 Personalized Study Guide Generation

This research designs a Moodle plugin, Personalized Study Guide (PSG), that can adjust the study guide a student is referring for learning in Moodle based on either his/her learning style calculated from FSLSM or his/her learning behaviour patterns found from learning logs. Figure 1 shows the PSG's workflow. There are three sources feeding to the PSG:

1. The Learning Resources/Activities: are the structure and the content of the learning resources/activities on Moodle.
2. Students' Learning Log: contains the actions students were taking on Moodle for learning.
3. ILS Questionnaire: stores students' responses to the ILS questionnaire.



**Fig. 1.** The system flowchart.

At the first stage, the Resource Learning Style Calculator module examines the types of the learning resources/activities and assigns the learning style weights on the learning resources/activities according to the analysis of the relationship between learning object and learning styles [1]. The weights are stored in the Resource Learning Style Scores. The PSG can (optionally) includes the LORD (Learning Object Relation Discovery) Module [KWTC22] to analyze the relationships among learning resources/activities to build the Learning Object Graph (LOG) for teachers to effectively use the Behaviour Analytics plugin clustering students' learning behaviours (see Stage 3 in Fig. 1). Despite of using the optional LORD module or not, the PSG adopts the Behaviour Analytics module [9] to cluster students into different groups according to their learning behaviours stored in Students' Learning Log by as Stage 4 in Fig. 1 shows.

With the ILS Questionnaire, student responses are stored in the Students' Learning Style Scores database (refers to Stage 5a). In th case when teachers to opt out the ILS Questionnaire, the Student Learning Style Evaluation module will adopt the student clustering results according to their behaviour patterns (see Stage 5b). The Student Learning Style Evaluation module first identifies which group the student belongs to according to the analysis results (refers to Stage 4) from the Behaviour Analytics module and then finds what learning resources/activities this group of students frequently visited. The learning style scores in these common visited learning resources/activities are summed up to determine what would be the learning style this group of the students might be.

For example, the most frequently visited activities the students in Group A are listed in Table 1. When the Student Learning Style Evaluation module sums up the weights of each activity as: {active: 1.24, reflective: 0.55, sensitive: 0.66, intuitive: 0.41, visual:

0.73, verbal: 0.66, sequential: 0.38, global: 0.27}, the students in Group A are classified to the learning style: (active, sensitive, visual, sequential).

**Table 1.** An example of calculating the learning style of the student group with the frequently visited learning materials/activities.

Materials/Activities	Learning Style Weights							
	Act	Ref	Sen	Int	Vis	Verb	Seq	Glb
Turtle and For Loop	0.04	0.22	0.04	0.17	0.30	0.04	0.17	0.00
Examples of Turtle Graphics	0.26	0.13	0.19	0.00	0.26	0.06	0.06	0.03
Project 2 Description	0.38	0.05	0.19	0.05	0.05	0.19	0.05	0.05
Discussion – Turtle Graphics	0.21	0.00	0.14	0.14	0.07	0.29	0.00	0.14
Quiz – Turtle Graphics	0.35	0.15	0.10	0.05	0.05	0.08	0.10	0.05
Sum	1.24	0.55	0.66	0.41	0.73	0.66	0.38	0.27

In Stage 6, the Learning Path Recommendation module compares the similarities of the student's learning style and the learning resources/activities' learning style scores. The learning resources/activities with higher similarities (in either a unit or in the whole course, depending on the course teacher's decision) will be placed in the beginning of the recommended learning path for the students. At the end, the Learning Path Recommendation module presents the personal study guide according to the generated learning path to the students on Moodle (refer to Stage 7 in Fig. 1).

## 4 The Moodle Plugin

When a course teacher signs in the Moodle, he or she can go to the setting page of the Personalized Study Guide, where (as Fig. 2 shows) he or she can decide how the PSG generates the personal learning paths for students. The first option – **Personalise within sections** – tells the PSG to generate the personal study guide by following or not following the current designed course section structure. If “within sections” is *checked*, then the rearranged content will still follow the section or unit or week structure the course has. On the other hand, the PSG will generate personal study guide for students by rearranging everything presented in the course – therefore, the students might see some content from the later semester/term at the beginning of the course if the teacher *unchecks* the option.

Figure 3 is the system snapshot (with some extra information revealing for readers) when the option “Personalise within sections” is *checked*. The PSG follows the original section order: Course general section, Introduction du cours, Methodologie de la veille, Stocker, memoriser, traiter, Diffuser les resultats de la veille, and Topic. The activities under each section (e.g., Flux RSS, Introduction du cours, Activité 1, Act 1, etc. under the section **Introduction du cours**) are reorganized according to the signed in student's learning style calculated from either the ILS scores or the student clustering results (see the top first line on Fig. 3).

The screenshot shows the 'Personalised Study Guide format' configuration page. At the top, a green bar says 'Changes saved'. Below it, the title 'Personalised Study Guide format' is displayed. A sub-section titled 'Choose how to personalise the Learning Objects in this course' follows. It contains two options:

- Personalise within sections** (format\_psg | withinsection)
  - Default: Yes
  - This causes the Learning Objects within each section to be changed while keeping the section order.
- Personalise sections** (format\_psg | bysection)
  - Default: No
  - This causes the sections to have their order changed.

Below this, another section titled 'Use ILS questionnaire results or common links for learning style in personalisation?' is shown. It has one option:

- Use the Index of Learning Styles questionnaire results from Behaviour Analytics** (format\_psg | userils)
  - Default: Yes
  - Checking this will require students complete the ILS for the course where personalisation is used. Unchecked will see use of the common links from the clustering analysis checked for prediction in Behaviour Analytics.

A blue 'Save changes' button is located at the bottom of this section.

**Fig. 2.** Configuration of the Personalized Study Guide

The screenshot shows a generated study guide. At the top, there is some code: `array ( 'active' => 0, 'reflective' => 7, 'sensing' => 0, 'intuitive' => 3, 'visual' => 1, 'verbal' => 0, 'sequential' => 0, 'global' => 3, )` and 'Scored: 0.6666666666666667'. To the right, a progress bar is shown with the text 'Your progress 0/0'.

Below this, a list of activities with their scores and checkboxes is displayed:

Activity	Score	Checkboxes
Forum	1	
Announcements	1	
Chat	0	

Scored: 0.25

**Introduction du cours**

Activity	Score	Checkboxes
Flux RSS	1	
Introduction du cours	0	
Activité 1	0	
Act 1	0	

**Fig. 3.** Generated personal study guide following the original section order.

Take the section **Introduction du cours** for example, the Flux RSS activity is scored 1 and is the best option for the signed in student to read first comparing to other activities that were scored as 0s in the same section.

On the other hand, if the teacher *checked* the option “**Personalized sections**” in Fig. 2, the PSG will also consider the section scores and rearrange the sections accordingly. Take Fig. 4 for example, compared with Fig. 3 the rearranged section sequence became section **Topic 5** (scored 1) first followed by Course general section (scored 0.67), Diffuser les résultats de la veille (scored 0.5), Introduction du cours (scored 0.25), Methodologie de la veille (scored 0), and Stocker, memoriser, traiter (scored 0).



**Fig. 4.** Generated personal study guide with the rearranged section order.

## 5 Evaluation Plan

To understand whether the generated personal study guide can help students in learning, the research team plans to work with two university professors (who are the course instructor of Academic Writing in English and English Presentation for Social Sciences) in different countries in Asia. The evaluation process is separated into two stages are Table 2 shows.

At the very beginning of the experiment, all students will be asked to fill out a built-in 44-item learning style questionnaire (i.e., the Inventory of Learning Styles, ILS) that the Behaviour Analytics plugin has. During the first stage (e.g., before the mid-term exam), the Personalized Study Guide plugin will be set to re-arrange learning activities according to student's learning style based on the ILS results. At the end of the first stage (approximately before mid-term exam), the students will be asked to fill out another built-in 10-item System Usability Scale questionnaire so the research team could know how they perceive the usability of the PSG when the PSG uses ILS scores to generate the personal study guide for them.

During the midterm exam, the teacher needs to use the Behaviour Analytics plugin to group students according to their learning behaviour on Moodle. If the teacher believes the clustering results are inappropriate, he or she can manually drag-and-drop the selected students from a group to another. The teacher is also asked to put notes to each student group explaining his/her observations and thoughts on the characteristics of the students in the group.

At both first and the second stage, students can freely decide to switch back and forth to the original study guide or the personal study guide at any time. The PSG will have the choice changes recorded so the research team could also understand student preference and even getting know more about whether the changes are related to their perceived usability as well as whether the changes contribute to their academic performance (i.e., the mid-term and final exam marks or the final grades they received for the course at the end).

**Table 2.** The evaluation plan for the Personal Study Guide plug-in

Stage	PSG	Activity	Collected Data
1	Is set for using ILS responses to generate personal study guide Can be manually turned on/off by students	Students fill out the entry survey at beginning of the course	Students' ILS responses
		Students study in Moodle with the PSG plug-in	Students' learning behaviour log in Moodle as well as when they turn on/off PSG and how many times they turn on/off PSG
		Students fill out the first exit survey	Students' responses toward the perceived system usability of the PSG
Mid-term	N/A	Students take the midterm exam	Students' midterm marks
		The teacher adjusts the student groups clustered by the Behaviour Analytics plugin and takes notes on the potential observations and student characteristics that each group might have	Students' group and the observation notes from the teacher
2	Is set for using student clustering results to generate personal study guide Can be manually turned on/off by students	Students study in Moodle with PSG plug-in	Students' learning behaviour log in Moodle as well as when they turn on/off PSG and how many times they turn on/off PSG
		Students fill out the second exit survey	Students' responses toward the perceived usability of the PSG
Final	N/A	Students take the final exam	Students' final exam marks and final grades

At the end of the second stage, the students will be asked to fill out another built-in 10-item System Usability Scale questionnaire so the research team could know how they perceive the usability of the PSG when it uses the student clustering results for rearranging the learning activities/resources. If the second stage has similar or better perceived usability than the first stage, then it might indicate that in the future we don't really need students to fill out a learning style questionnaire, but the computer can cluster students automatically and help to provide them personal study guide (i.e., rearrange learning activities according to the clustering results).

## 6 Conclusion

This research designs the Personal Study Guide Moodle plugin which generates personal learning path according to students' learning styles. By using existing Moodle plugin (i.e., the Behaviour Analytics) and following the earlier found relationships between learning styles and objects/pedagogy, the PSG calculates the learning style scores for each learning resources/activities in Moodle and rearranges the best matching learning activities/resources for students accordingly.

The research team plans to conduct an experiment described in the previous section in 2023 in order to evaluate the system usability and effectiveness of the Personal Study Guide in learning. In the next stage, the research team will evaluate the collected data to determine whether the calculation of students learning styles by using their learning behaviour is accurate.

**Acknowledgements.** The authors acknowledge the support of the National Science and Engineering Research Council of Canada (NSERC) [RGPIN-2018-05545].

## References

1. Assis, L., Rodrigues, A.C., Vivas, A., Pitangui, C.G., Silva, C.M., Dorça, F.A.: Relationship between learning styles and learning objects: a systematic literature review. *Int. J. Dista. Edu. Technol. (IJDET)* **20**(1), 1–18 (2022)
2. Coffield, F., Moseley, D., Hall, E., Ecclestone, K.: Learning styles and pedagogy in post-16 learning: A systematic and critical review (2004)
3. Felder, R., Silverman, L.: Learning and teaching styles in engineering education. *Engr. Education* **78**(7), 674–681 (1988)
4. Felder, R.M., Solomon, B.A. Index of Learning Styles Questionnaire (1997). Retrieved 28 January 2023. from <https://www.webtools.ncsu.edu/learningstyles/>
5. Ferguson, R.: Learning analytics: drivers, developments and challenges. *Int. J. Technol. Enhan. Learn.* **4**(5–6), 304–317 (2012)
6. Jackson, C., Lawty-Jones, M.: Explaining the overlap between personality and learning style. *Personality Individ. Differ.* **20**(3), 293–300 (1996)
7. Kew, S.N., Tasir, Z.: Learning analytics in online learning environment: a systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning*, 1–23 (2021)
8. Kolb, D.A., Rubin, I.M., McIntyre, J.M.: Organizational psychology: readings on human behavior in organizations. Prentice-Hall, Englewood Cliffs, NJ (1984)
9. Kuo, R., Krahn, T., Chang, M.: Behaviour analytics - a moodle plug-in to visualize students' learning patterns. In: 17th International Conference on Intelligent Tutoring Systems, (ITS 2021). Hybrid, Athens, Greece, June 7–11, 2021, pp. 232–238 (2021)
10. Kuo, R., Wasowski, R., Krahn, T., Chang, M.: LORD: a moodle plug-in helps to find the relations among learning objects. In: 18th International Conference on Intelligent Tutoring Systems, (ITS 2022), June 29–July 1, 2022. Hybrid, Bucharest, Romania (2022)
11. Marosan, Z., Savic, N., Klasnja-Milicevic, A., Ivanovic, M., Vesin, B.: Students' perceptions of ILS as a learning-style-identification tool in e-learning environments. *Sustainability* **14**(8), 4426 (2022)

12. Sensuse, D.I., Hasani, L.M., Bagustari, B.: Personalization strategies based on Felder-Silverman learning styles and its impact on learning: A literature review. In: 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), pp. 293–298. IEEE (2020)



# Learning by Building Chatbot: A System Usability Study and Teachers' Views About the Educational Uses of Chatbots

Stavros A. Nikou<sup>1</sup> and Maiga Chang<sup>2</sup>

<sup>1</sup> School of Education, University of Strathclyde, Glasgow, Scotland, UK  
stavros.nikou@strath.ac.uk

<sup>2</sup> School of Computing and Information Systems, Athabasca University, Edmonton, Canada  
maiga.chang@gmail.com

**Abstract.** This article describes an experiment to investigate the usability of the “Learning by Building Chatbot” environment and the views of in-service teachers about the potential educational uses of chatbots. Chatbots are text or voice-based conversational interfaces that use natural language to simulate human conversations serving as virtual assistants to users. They have been used in a variety of areas such as e-commerce and healthcare while their use in education is relatively new. Chatbots have the potential to offer new dynamic forms of interactions, but there is often a burden to users to deploy and maintain their own chatbots. Our research group has proposed a web-based environment, “Learning by Building Chatbot”, that alleviates this burden by offering educators and students without any programming knowledge the opportunity to deploy and maintain their own chatbots for learning and training. The environment is a block-based, visual editing environment of creating RiveScript-powered chatbots for learning and training, without needing to know any RiveScript. Twelve in-service teachers registered in a postgraduate program in technology-enhanced learning used this environment to create their own chatbots. Afterwards, they completed the System Usability Scale (SUS) questionnaire. Results have shown that the visualized editing environment for building chatbot is marginally acceptable and fair to use with the SUS score 57.50. Participants found the system ease to use and felt quite confident to use it. Its functions are well integrated while there are a few inconsistencies in the current version. Moreover, participants self-reported their views about the educational uses of chatbots. They agreed that chatbots, if used appropriately, can be a valuable educational tool because they can automate administrative and teaching tasks and can be supportive and engaging for students. However, since chatbot development is challenging, teacher should be well supported to integrate chatbots in the educational practice if they intend to do so.

**Keywords:** Block · Chatbot · System Usability · RiveScript

## 1 Introduction

Conversational User Interfaces (CUI) or chatbots are software applications that use natural languages to conduct online conversations with users via text or speech, serving as

virtual assistants [1]. Chatbots offer the capacity to humans to use their natural language to communicate rather than syntax specific computer commands, opening up possibilities for more dynamic and accessible forms of interaction. Latest developments in Natural Language Processing (NLP), Natural Language Understanding (NLU) and Artificial Intelligence (AI), combined with the increased computing power available helped chatbots to become more mainstream.

There is an increased use of chatbots in a variety of areas such as e-commerce, retail services, customer support, healthcare and lately in education. However, studies related to chatbots in education are still in an early stage [2] and the potential of using chatbots in education have not yet fully explored [3]. A recent literature review by [4] categorized chatbots in education into three categories: learning chatbots, assisting chatbots, and mentoring chatbots. Chatbots can support student learning, e.g., they can assist in language learning [5] or programming [6]; they can assist students by providing administrative assistance, e.g., answer general questions about a course, timetable, assignment deadlines, etc. [7]; and, they can mentor students by providing scaffolding and recommendations, e.g., through adaptive formative quiz feedback [8].

To date many chatbot applications in education are offered as commercial off-the-shelf products. Users have quite a limited capacity to create, modify and maintain their own chatbots unless they have a programming background. The current project is aiming to develop a platform that teachers can freely use to create their own chatbots or have their students to create their chatbots for learning and training without prior programming knowledge and skills.

Based on [9] software attributes can be “hedonic” and “pragmatic”. “Pragmatic attributes emphasize the fulfilment of individuals’ behavioral goals, hedonic attributes emphasize individuals’ psychological well-being.” Usability is a pragmatic attribute that refers to the fulfilment of users’ functional goals and therefore it is important to be measured [10]. Usability is defined in ISO 9241–11:2018 (Sect. 3.1.1) [11] as ‘the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.’

The System Usability Scale (SUS) is a validated questionnaire [12] that can be used to assess the usability of any hardware, software system, device or service [13–16] and also it has wide acceptance and easy administration. The questionnaire consists of 10 items that are answered using a 5-step Likert scale ranging from “strongly disagree” to “strongly agree”, resulting in a single score between 0 and 100 (in 2.5 points increments) where higher scores indicate better usability. Scores<sup>1</sup> below ‘50’ is unacceptable, over ‘50’ is OK, over ‘70’ is acceptable, and a score of ‘85’ can be considered ‘Excellent’.

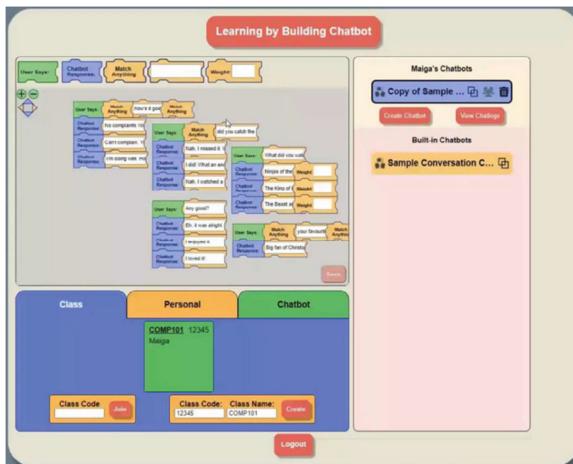
[17] has found the minimum number of test users for a usability study is 5 and [18] have the general rule  $10 \pm 2$  for optimal sample size and for major usability evaluations and meta-analysis results is 10. This study aims to investigate the perceived usability that in-service teachers have after they use the proposed “Learning by Building Chatbot” environment.

---

<sup>1</sup> <https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html>.

## 2 Learning by Building Chatbot

Our research team has proposed a block-based, visual editing environment based on RiveScript that enables users to create, view, test and manage their own chatbots [19]. Figure 1 shows a screenshot of the “Learning by Building Chatbot” environment.



**Fig. 1.** The Visualized Editing Environment for Building Chatbot (<https://vp.vipresearch.ca/>).

While RiveScript has a learning curve with distinct lexicons, formatting requirements and other nuances for the non-programmer, the “Learning by Building Chatbot” alleviates this burden for users. The editing environment handles the syntax, semantics, and other idiosyncrasies of the RiveScript language. New chatbots can be easily built by dragging and snapping together their building blocks and typing inside the blocks the conversations to be carried out. The platform offers a co-creating feature that users can share the chatbot they designed to others and multiple people can work on different parts of a chatbot together.

The platform allows the chatbots to be managed by their users and allows teachers to review student chatbot interactions from a single website without the need to self-host a chatbot or distribute a program.

## 3 Methodology

### 3.1 Participants and Procedures

The study took place in the context of the course “Technology Enhanced Teaching and Learning: Theory and Practice” course, part of the Postgraduate Certificate in Digital Education within the MEd in Education Studies program at the University of Strathclyde. The aim of the course is to provide students with a theoretical background on technology enhanced learning and introduce them to a variety of educational technologies. Conversational Agent is one of these technologies. Twelve in-service teachers who

were registered in the course participated in the study. The number of participants is within the usability study's general rule  $10 \pm 2$  for optimal sample size.

Once ethical approval was granted by the University of Strathclyde School of Education Ethics committee, researchers invited students to participate in the study. Participation was voluntarily and anonymous. The study took place during fall semester 2022. After a dedicated guest lecture on the use of chatbots in education, students were introduced to the "Learning by Building Chatbot" environment. Participants familiarised themselves with the environment, during an in-class session, by using it to build their own chatbots. They were asked to build simple chatbots that could be used in simple conversations with their pupils on frequently asked content or administration questions related to their classes. Finally, they have been asked to complete an online questionnaire made available to them through Qualtrics.

### 3.2 Instruments

The questionnaire had three parts. The first part included a series of socio-demographic and teaching related questions namely: gender, age, country of teaching, general digital skills level (basic, intermediate, advance), level and subject of teaching as well as previous experience with chatbots in education. The second part included the 10-item SUS questionnaire aiming to measure the usability of the "Learning by Building Chatbot" environment. The third part included the following open-ended questions aiming to increase our understanding on teachers' views on the potential applications, benefits, and challenges of chatbots in education:

- Q1: What do you think about the (potential) use of the chatbots for teaching and learning purposes? Any particular areas of possible applications?
- Q2: How would you perceive the value of the chatbots in your own teaching? What do you think are the associated benefits and/or challenges.

The following two sections present the results following with the main conclusions and related discussions.

## 4 Results

Based on the responses to the first part with the socio-demographic and teaching related questions, participants were 6 female, 5 male, and 1 preferred not to say. Their distribution in terms of their age was 4 participants 21–30 years old, 2 participants 31–40 years old, 5 participants 41–50 years old and 1 participant 51–60 years old. In terms of their general digital skills level, 3 participants had basic digital skills (use of a basic range of software such as office; and devices such as computer, tablet), 8 participants had intermediate digital skills (use of a big variety of software such as, audio and video processing software; and devices such as smart interactive whiteboards), and 1 participant had advanced digital skills (use of highly innovative and complex digital and communication technologies, advanced skills such as programming, software and web development).

Most of the participants were teaching in primary education (5 participants) with tertiary education (4 participants) and secondary education (2 participants) to follow, while 1 participant described the education level they were teaching as other. Their teaching experience in years was varied with 5 participants to have 1–5 years teaching experience, 4 participants to have 6–10 years teaching experience and 3 participants to have more than 16 years teaching experience.

The distribution of the participants in terms of the subject they were teaching was 3 participants were teaching Science (Math, Physics, Chemistry, Biology), 1 participant was teaching Informatics/Technology, 1 participant was teaching Social Sciences, 2 participants were teaching Language/Literature while 5 participants reported their teaching subject as other (as these participants were teaching in primary schools). Finally, regarding the previous use of chatbots in class, only 1 out of the 12 participants said that they had used chatbots in their classes before.

Regarding the results of the SUS questionnaire, a total of twelve responses were collected. While [12] suggested that SUS is a unidimensional instrument with its questionnaire items better not considered individually, to highlight each one questionnaire item, Table 1 shows the responses on the individual questionnaire items along with the median, the mean and standard deviation.

**Table 1.** SUS questionnaire and statistics for each item.

	Strongly Disagree 1	2	3	4	Strongly Agree 5	Median	Mean	SD
1. I think that I would like to use this system frequently	1	6	2	3	0	2	2.58	0.99
2. I found the system unnecessarily complex	2	5	5	0	0	2	2.25	0.75
3. I thought the system was easy to use	1	4	3	3	0	3	2.92	1.16
4. I think that I would need the support of a technical person to be able to use this system	3	6	3	0	0	2	2.00	0.74
5. I found the various functions in this system were well integrated	0	6	5	1	0	2.5	2.58	0.67

(continued)

**Table 1.** (*continued*)

	Strongly Disagree 1	2	3	4	Strongly Agree 5	Median	Mean	SD	
6. I thought there was too much inconsistency in this system	0		6	5	1	0	2.5	2.58	0.67
7. I would imagine that most people would learn to use this system very quickly	1		1	6	4	0	3	3.08	0.90
8. I found the system very cumbersome to use	1		6	4	1	0	2	2.42	0.79
9. I felt very confident using the system	1		2	4	5	0	3	3.08	0.99
10. I needed to learn a lot of things before I could get going with this system	4		3	3	2	0	2	2.08	1.16

Based on [12] the overall System Usability Score, representing the composite measure of the overall usability of the system was found 57.50, which indicates the environment and its functionality at this moment is OK for using. Regarding the third part, participants' responses to the open-ended questions about their views on the potential applications, benefits, and challenges of chatbots in education highlighted the following.

The majority said that chatbots can be used for administrative purposes: "Provide administrative information about marking procedure, assignments, school policies", "Provide feedback to common student questions mostly related to admin stuff, e.g., absences, marking policies etc.", "I think they could be good for answering basic admin questions such as due dates, class timetable."

Most participants said that chatbots can also be used for teaching and learning purposes: "It would be used to test factual knowledge", "It is good for digital literacy development" and "Good for comprehension in literacy", "I think it can be used in teaching languages and math" or "One use could be building a revision tool, whereby you essentially create a digital glossary on a series or unit of lessons."

One participant mentioned that chatbots are appealing for young generations because they are interactive: "Young people would enjoy the interactivity chatbot generate." One participant said that chatbots can be used to train preservice teachers: "Provide teaching scenarios simulating common classroom dialogues that student teachers can respond

and practice with" and another one said that they can be used for communication with the parents: "In primary it could be good for communication with parents."

Five participants agreed that chatbots could be used to automate tasks that are performed frequently to "answering a series of most asked questions" and integrate teachers' feedback "in areas that students need most clarification."

The majority of participants agreed that the use of chatbots can be beneficial for students as "they can help learners control their learning and be creative." One said that they can enhance student engagement as "Chatbots can be engaging for pupils" and another one said that they "Encourage independence during tasks and more technology awareness for pupils." Most participants agreed that chatbots have "Lots of benefits, save time, extra support" and seven participants said that chatbots be useful not only in teaching but in assessment as well: "It is a useful learning and assessment tool." Participants seemed to agree that "Any activity where knowledge must be stored for easy access would suit the use of a chatbot." However, they found that it is difficult for chatbots to "mimic human logic and empathy."

From the teachers' point of view, two participants mentioned that chatbots, once developed they can be very useful for teachers because they can "facilitate processes" and can "free up administration time". However, participants agreed that it is "Challenging for teachers to build their own chatbots." One teacher was skeptical about the generalized use of chatbots in education – "I do not believe these types of systems will ever be fully accepted by the teaching professionals ... Until such times as the subtle variations in human behavior (variations in spelling, addition of grammar symbols, etc.) can be automated without programming." A participant emphasized that "The time potentially required to construct a worthwhile system that could be used in place of an actual person" is a real challenge. Unless "...pre-made chatbots can be produced to specific coursework then their potential use would be much greater." A participant also said, "The time required to build a chatbot (particularly by those whose digital skills are limited) to a level where it can be deemed usable and interactive could prevent implementation by teaching professionals". Another one suggested a way that teachers could be motivated to use chatbots if they receive appropriate training – "Chatbot use can be demonstrated to teaching professionals in their area of expertise."

In conclusion, despite the aforementioned challenges, participants tended to agree that chatbots can be useful if used appropriately – "I can imagine a myriad of uses for a system such as this if used correctly."

## 5 Discussion

The result found for the SUS of the "Learning by Building chatbot" is below 68 which is considered as average. However, Lewis and Sauro (2009) did a sensitivity data analysis with 19 datasets, and they came out "the mean of our Overall SUS data was 62.1, with a 99.9% confidence interval ranging from 58.3 to 65.9" instead of earlier [13] distribution 70.1 (with a 99.9% confidence interval ranging from 68.7 to 71.5). Our result **57.50** is falling a little bit outside of 58.3 (i.e., the 99.9% interval range).

Moreover, as [20] reported, acceptable score corresponds to roughly above 70, **marginally acceptable to 50–70**, and unacceptable to below 50 [13]. The report also

categorized “promoters” are close to 81, “**passives**” are between **53 and 81**, and “detractors” are associated with 53 and below. In respect to the use of the adjectives including “Good,” “OK,” and “Poor”, “excellent” is associated with 85, “good” is above 71, “**OK**” (or “**Fair**”) is above **51**. Finally with respect to grades, A + is 84.1 to 100, A is 80.8 to 84, A- is 78.9 to 80.7, B + is 77.2 to 78.8, B is 74.1 to 77.1, B- is 72.6 to 74, C + is 71.1 to 72.5, C is 65 to 71, C- is 62.7 to 64.9, and D is 51.7 to 62.6. Therefore, we can safely say that the visualized editing environment for building chatbot is marginally acceptable and ok/fair to use with the SUS score **57.50**.

Considering the questionnaire items individually, the positive items #9, #7 and #3 had the higher mean values indicating that teachers felt confident using the system, most people would learn to use it very quickly and the system perceived as easy to use. We can marginally infer that teachers would like to use the system frequently (item #1); however, teachers have found that there was still some inconsistency in the system (item #6) despite the fact that various functions were well integrated (item #5).

The standard deviation of item #3 was relatively high indicating that not all teachers agreed with the level of easiness of the system, probably due to their own different digital skills level. The same holds for the standard deviation of item #10 indicating that teachers would need different levels of technical support before starting using the system.

Finally, considering that SUS actually has two factors [16], i.e., Usability (items #1, #2, #3, #5, #6, #7, #8) and Learnability (items #4 and #9), interpretation of the participants’ responses indicate that the system has high learnability since the median values for items 4 and 9 are quite low. Further studies can measure SUS in correlation to gender or digital skills.

In regard to the responses to the open questions related with teachers’ perceptions on the potential use of chatbots on education, their benefits and challenges, teachers agreed that chatbots, is a promising educational technology and if used appropriately, can be a valuable educational tool because they can automate administrative tasks (e.g., inform about assignments, class timetables, syllabus, due dates, school policies, etc.) and teaching tasks (e.g., answer subject related questions, provide automated feedback on areas where students need most clarification) or assessment tasks (e.g., personalized assessment, automate marking and assessment feedback). Moreover, due to their interactivity they can engage students [21]. Our findings agree with previous research. Chatbots can offer many opportunities to the teaching and learning process [2, 4, 22]. However, their deployment in education faces many challenges due to the technological limitations and appropriate pedagogical integration [2, 23]. Chatbot development is challenging, therefore teachers should have the support needed to integrate them in educational practice. The “Learning by Building” chatbot project is at its rather early stage of development and our future work aims to deliver a chatbot development environment for educational use to improve student learning through chatbot construction.

## References

1. Luo, X., Tong, S., Fang, Z., Qu, Z.: Frontiers: Machines Vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science* (2019). <https://doi.org/10.1287/mksc.2019.1192>

2. Hwang, G.-J., Ching-Yi Chang, C.-Y.: A review of opportunities and challenges of chatbots in education. *Interact. Learn. Environ.* (2021). <https://doi.org/10.1080/10494820.2021.1952615>
3. Winkler, R., Soellner, M.: Unleashing the potential of chatbots in education: a state-of-the-art analysis. *Academy of Management Annual Meeting Proceedings* **2018**(1), 15903 (2018)
4. Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., Drachsler, H.: Are we there yet? - a systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*. **4**, 654924 (2021). <https://doi.org/10.3389/frai.2021.654924>
5. Kim, N.-Y.: A study on the use of artificial intelligence chatbots for improving english grammar skills. *J. Digital Convergence* **17**, 37–46 (2019). <https://doi.org/10.14400/JDC.2019.17.8.037>
6. Daud, S.H.M., Teo, N.H.I., Zain, N.H.M.: E-java chatbot for learning programming language: a post-pandemic alternative virtual tutor. *Int. J. Emerging Trends Eng. Res.* **8**(7), 3290–3298 (2020). <https://doi.org/10.30534/ijeter/2020/67872020>
7. Sandoval, Z.V.: Design and implementation of a chatbot in online higher education settings. *Issues Inf. Syst.* **19**, 44–52 (2018). <https://doi.org/10.48009/4.iis.2018.44-52>
8. Draaijer, S., Joosten-ten Brinke, D., Ras, E. (eds.): CCIS, vol. 1014. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-25264-9>
9. Hassenzahl, M.: The thing and I: Understanding the relationship between user and product. In: Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (eds.) *Funology: from usability to enjoyment*, pp. 31–42. Springer Science + Business Media Inc (2005)
10. Hassenzahl, M.: User experience and experience design. *The Encyclopedia of Human-Computer Interaction*. Interaction Design Foundation (2013). <https://www.interaction-design.org/literature/book/the-encyclopedia-of-humancomputer-interaction-2nd-ed/user-experience-and-experience-design>
11. ISO: 9241–11:2018. Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and concepts. International Standardization Organization (ISO) (2018)
12. Brooke, J.: In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) *SUS: A ‘quick and dirty’ usability scale in Usability evaluation in industry*, pp. 189–194. Taylor and Francis, London/UK (1996)
13. Bangor, A., Kortum, P., Miller, J.: An empirical evaluation of the system usability scale. *Int. J. Human-Computer Interaction* **24**(6), 574–594 (2008). <https://doi.org/10.1080/10447310802205776>
14. Bangor, A., Kortum, P., Miller, J.: Determining what individual sus scores mean: adding an adjective rating scale. *J. Usability Stud.* **4**(3), 114–123 (2009)
15. Kortum, P., Bangor, A.: Usability ratings for everyday products measured with the system usability scale. *Int. J. Human-Computer Interaction* **29**(2), 67–76 (2013). <https://doi.org/10.1080/10447318.2012.681221>
16. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International*, pp. 94–103 (2009). [https://doi.org/10.1007/978-3-642-02806-9\\_12](https://doi.org/10.1007/978-3-642-02806-9_12)
17. Nielsen, J.: How many test users in a usability study? (2012) Retrieved from <https://www.nngroup.com/articles/how-many-test-users/>
18. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: The  $10 \pm 2$  rule. *Communications of the ACM* **53**(5), 130–133 (2010). Retrieved from <https://doi.org/10.1145/1735223.1735255>
19. Chang, M., McQuoid, S.: Visualized environment to build chatbot for learning and training. In: *1st International Workshop on Metaverse and Artificial Companions in Education and Society (MetaACES 2022)*, Virtual, June 24, 2022, pp. 37–38 (2022)
20. Sauro, J.: 5 Ways to Interpret a SUS Score. MeasuringU (2018). Access: <https://measuringu.com/interpret-sus-score/>

21. Cunningham-Nelson, S., Boles, W., Trouton, L., Margerison, E.: A review of chatbots in education: practical steps forward. 30th annual conference for the australasian association for engineering education (AAEE 2019): Educators becoming agents of change: Innovate, integrate. Motivate: Engineers Australia (2019)
22. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: a systematic review. *Comp. Edu. Artifi. Intelli.* **2**, 100033 (2021)
23. Huang, W., Hew, K.F., Fryer, L.K.: Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *J. Comput. Assist. Learn.* **38**(1), 237–257 (2022). <https://doi.org/10.1111/jcal.12610>



# Towards Integrating Learnersourcing, Microlearning and Gamification in Moodle

Mariana Madalina Nastase<sup>(✉)</sup> and Elvira Popescu

Computers and Information Technology Department, University of Craiova,  
Craiova, Romania

[nastase.mariana.e8b@student.ucv.ro](mailto:nastase.mariana.e8b@student.ucv.ro), [elvira.popescu@edu.ucv.ro](mailto:elvira.popescu@edu.ucv.ro)

**Abstract.** In this paper we aim to address the need for a switch in the educational system from a content-centred to a student-centred approach by peeking into the current trends and perspectives of integrating pedagogical approaches such as learnersourcing, microlearning and gamification in a widely-used LMS, such as Moodle. We present an overview of the recent literature treating the issues related to learnersourcing, microlearning and gamification and analyse existing shortcomings and challenges of the current approaches. We also aim to show the development capabilities that Moodle LMS encompasses and how it might serve our proposed solutions which deal with assisting the teachers in searching for micro activities that will suit the students' needs, on the one hand, and with the creation of a safe space in which students can express freely, collaborate and be assisted in their work, on the other hand.

**Keywords:** Moodle plugins · Learnersourcing · Collaborative learning · Microlearning · Gamification

## 1 Introduction

Sir Ken Robinson is widely acknowledged for advocating a more creative and innovative approach to education. He argues that the traditional educational system fails to meet the diverse needs and interests of the students since it is usually based on standardised approaches and he calls for a shift to a student-centred approach that emphasises creativity, critical thinking and innovation [1].

Integration of learnersourcing, microlearning and gamification in learning management systems is aligned with Sir Ken Robison's ideas in several ways. Learnersourcing involves empowering students to have an active role in their learning by contributing to content creation [2], which aligns with Robinson's call for students to be seen as co-creators rather than passive recipients. Microlearning and gamification, on the other hand, promote engagement and motivation by delivering small chunks of enjoyable and challenging information [3,4], reflecting on Robinson's belief that learning should be fun, engaging and relevant to students' interests.

A learning management system (LMS) is an online environment, usually a web-based platform, that facilitates learning by providing customised online instructional materials and interactive activities. More than 200 commercial LMSs offer features such as assignments, chat, resources, courses and forums [5]. Among these, Moodle is one of the most frequently used LMSs due to its easy access from mobile or laptop, the organisation of learning materials, the features for assignments and feedback, students' analytics, capabilities for collaborative learning and cost-time efficiency [6]. The platform can also be extended using plugins to create a personalised experience for the students and adapt to their particular needs [7]. Many studies attempted to address the issues related to the adaptability of learning using Moodle plugins due to its general technical stability [8–10]. Their results support the correlation between students' performance and the adaptation of learning materials according to the student's preferences, indicating the need for a student-centred approach to learning rather than a content-centred approach.

Despite the potential for Moodle to provide a personalized and adaptable learning environment, studies have shown that teachers often lack didactical knowledge of the platform and use it primarily as a resource repository, as shown by the survey conducted by Almenara et al. [11]. This highlights the need for teachers to invest effort in designing activities that use Moodle's capabilities for personalized and adaptive learning, which can be supported by the pedagogies related to learnersourcing, microlearning and gamification.

The rest of the paper is assembled to provide an overview of the current trends and approaches in learnersourcing, microlearning and gamification (Sect. 2), outline the existing Moodle plugins that facilitate these pedagogies (Sect. 3), summarise the shortcomings identified and propose several solutions and starting points for future work (Sect. 4).

## 2 Current Trends in Learnersourcing, Microlearning and Gamification

In this section, we provide an overview of the progress that has been made in the area of smart learning environments, highlighting the benefits and pitfalls of instructional approaches, such as learnersourcing, collaborative learning, microlearning and gamification.

**Learnersourcing** is a pedagogically supported crowdsourcing that empowers learners with an environment that allows them to contribute to teaching and learning while actively involved in learning activities themselves [2]. This type of learning was intensively studied in the latter years as it opened many opportunities for modern education. Singh et al. [12] propose a theoretical framework for learnersourcing by engaging the main stakeholders of the learning process and their contribution. Their results show that simple tasks are more effectively done when the system provides a high level of guidance, for example, by suggesting, editing or voting. The complex tasks are more suitable for proficient and expert learners requiring little guidance, but a group of learners could also complete

these complex tasks by working collaboratively. Thus, the systems should foster collaboration between the learners. The studies show that learners who choose to contribute and create new high-quality learning artefacts do so because they are interested in the novelty of the task and want to help future learners. The students who choose not to do it invoke motives like lack of time, confidence or interest.

Moore et al. [13] also have an opinion regarding the challenges faced by learnersourcing. They state that participation in learnersourcing tends to be low when the activity is optional. It is hard to evaluate how students act on the provided feedback, and it proved to be a challenge to integrate students' evaluation of the materials into the learning process; thus, the students require assistance in creating high-quality resources.

As Singh et al. [12] mentioned in their study, learnersourcing activities usually imply **collaborative learning**, especially since it has been proven to increase motivation, promote active work, and foster creativity [14]. However, multiple factors might influence the success of this approach, and a couple of aspects should be taken into account. Collaborative learning implies organisational work in choosing the appropriate tasks and resources, which usually becomes time-consuming for the instructors, leaving aside the technical knowledge required [14, 15]. There is also the worry of assessing the students' work as the current system still requires quantitative results such as grades and points [14].

The students' prior knowledge also influences the success of collaborative learning [15]. Research shows that this type of learning is proper when the participants have complementary knowledge they can share with their peers while acquiring new knowledge simultaneously. However, it is redundant for those with complete understanding and requires a high level of engagement for investing mental effort in transactive activities [15–17]. Bause et al. [18] attempt to balance the discussion bias caused by prior knowledge and raise the discussion intensity and duration by using a multi-touch table and controlled conditions for collaborative learning. Their work was successful as the participants were more engaged and involved in the task, and the percentage of correct answers was high.

Whatever the challenges, studies show that learners are more likely to recall and remember information if they construct it themselves rather than having it given by somebody else [12]. Further development in this field will make education more accessible and beneficial to students. Endless types of activities could be created and evaluated using learnersourcing techniques. Spreading these methods worldwide in educational environments will lead to millions of resources that could be used to create more and more personalised learning experiences [13].

Significant research was also conducted in the direction of **microlearning** and **gamification** as a solution to the educational system's cognitive challenges nowadays, such as students' cognitive overload, short attention span and the need for instant gratification. The human brain has evolved to seek instant gratification, and microlearning activities provide immediate satisfaction, which increases

the probability that the action will be repeated. Tracking progress is another way to motivate learners to engage in the activities, as it provides instant rewards [19].

Recent literature reviews analyse the progress that has been made in this area. They summarise the benefits and threats of these pedagogical approaches and draw guidelines for developing successful systems that could facilitate these theories. De Gagne et al. [3] analysed 17 studies from 2011 to 2018 in health professions education. The review showed positive results in students' reactions to microlearning, knowledge and skill acquisition and overall behaviour. These results are also supported by the study conducted by Garshasbi et al. [20]. They review a significant body of literature on STEM education, starting from a couple of pedagogical theories, such as Sweller's Cognitive Load Theory [21,22], Mayer's Cognitive Theory of Multimedia Learning [23], Ryan's Self-Determination Theory of Motivation [24], Bouillion and Gomez (learning in the context of daily habits) [25], Collaborative Cognitive Load Theory [26], and draw essential guidelines in designing a system that facilitates microlearning.

Their investigation shows that microlearning should be flexible enough to facilitate learning in multiple ways: it should support various learning needs and encourage lifelong learning while connecting diverse pedagogical theories and approaches. It should also not be dependent on the environment and context. It should support both online and face-to-face learning, as well as academic, corporate or individual education, allowing the learner to take control over their learning process and thus enhance their motivation and preference for lifelong learning. It could be a complement to overcome certain shortcomings of the curriculum. It should be robust and scalable to be easy to integrate into existing or upcoming learning platforms. This will allow the designers to track learners' progress and feedback.

However, it is not suitable for in-depth training or complex concepts. While the benefits of using microlearning are noticeable, the results show that these activities should be complementary. Taylor et al. [27] discovered that the systems that use only microlearning produced lower satisfaction levels than those that use microlearning as a part of an extensive learning ecosystem.

Gamification is another pedagogy that can act combined with learnersourcing and microlearning to increase their effectiveness. Most studies attempt to prove a direct correlation between the gamification of an educational tool and an increase in the motivation and engagement of the students, leading to a rise in their academic performance. The study by Denny et al. [4] supports Landers' theory of gamified learning [28]. It proposes a model that explains how gamification triggers a behaviour that might generate a specific outcome. In the study context, the gamification of the learning process will increase self-testing and question authoring behaviour, improving exam performance. The study addresses one criticism of gamification, which argues that using external rewards, such as points and badges, might harm the learners' intrinsic motivation. They conclude that while giving external rewards will trigger a sense of competence, it might also reduce that students' autonomy. On the contrary, the

study by Van Roy and Zaman [29], based on the self-determination theory [30], concludes that game elements do not necessarily increase motivation. Rather, individual personality traits influence this. One of the main insights that Gordon et al. [31] demonstrated in the last decade refers to some of the primary responses of the brain. Their research shows that the brain is built to “minimize the danger and maximize the reward” [31], which might be reliable evidence supporting gamification.

However, the systematic reviews conducted by Majuri et al. [32] and Manzano et al. [33] concluded that most researches show a positive effect of using gamification elements to increase students’ motivation and engagement, leading to an improvement in their academic performance.

### 3 Relevant Moodle Plugins

As Moodle representatives state, their mission is to “empower educators to improve our world” and “give the world the most effective platform for learning”<sup>1</sup>. Because of that, Moodle already facilitates various learning activities for learnersourcing, microlearning and gamification through numerous plugins developed by their employees or by their worldwide community.

**Learnersourcing** is an activity highly dependent on students’ behaviour, and it usually goes hand in hand with **collaborative activities**. Moodle provides plugins that empower students with tools for creating and sharing learning resources and providing feedback. “*Student Blog*”<sup>2</sup> and “*Student Wiki*”<sup>3</sup> allow the creation of public posts that foster self-expression and sharing knowledge and ideas. “*Student Journal*”<sup>4</sup> encourage self-reflection and self-assessment. “*Student Quiz*”<sup>5</sup> allows students to create quiz questions that could be used in class. “*Poll*” plugin<sup>6</sup> promotes students’ autonomy by allowing them to vote on topics of interest or further learning paths. “*Student Feedback*”<sup>7</sup> and “*Point of View*”<sup>8</sup> gives the students the freedom to express their satisfaction with a topic or learning activity. All of these plugins encourage active participation and thoughts sharing.

Moodle also provides a set of plugins that aim to facilitate group work and collaboration. The “*Group Choice*” plugin<sup>9</sup> gives the students autonomy to choose the group they want to work with based on their interests. On the other hand, The “*Group Self-Selection*” plugin<sup>10</sup> provides the students with complete control over the group formation. “*Team Assignment*”<sup>11</sup> allows the teachers to

---

<sup>1</sup> <https://moodledev.io/general/community/mission>.

<sup>2</sup> [https://docs.moodle.org/35/en/Blog\\_settings](https://docs.moodle.org/35/en/Blog_settings).

<sup>3</sup> [https://docs.moodle.org/35/en/Wiki\\_activity](https://docs.moodle.org/35/en/Wiki_activity).

<sup>4</sup> [https://docs.moodle.org/35/en/Journal\\_module](https://docs.moodle.org/35/en/Journal_module).

<sup>5</sup> [https://moodle.org/plugins/mod\\_studentquiz](https://moodle.org/plugins/mod_studentquiz).

<sup>6</sup> [https://moodle.org/plugins/block\\_poll](https://moodle.org/plugins/block_poll).

<sup>7</sup> [https://moodle.org/plugins/qbehaveavour\\_studentfeedbackdeferred](https://moodle.org/plugins/qbehaveavour_studentfeedbackdeferred).

<sup>8</sup> [https://moodle.org/plugins/block\\_point\\_view](https://moodle.org/plugins/block_point_view).

<sup>9</sup> [https://moodle.org/plugins/mod\\_choicegroup](https://moodle.org/plugins/mod_choicegroup).

<sup>10</sup> [https://moodle.org/plugins/mod\\_groupselect](https://moodle.org/plugins/mod_groupselect).

<sup>11</sup> [https://moodle.org/plugins/assignment\\_team](https://moodle.org/plugins/assignment_team).

create and manage group assignments. The “*Collabora*” plugin<sup>12</sup> enables students to collaborate on a shared document in real-time. The “*Big Blue Button*” plugin<sup>13</sup> provides web conferencing and collaboration tools, such as video, audio and screen sharing.

Karampa and Paraskeva [34] successfully integrate multiple Moodle plugins (theme, blocks, grids, badges) to incorporate the ARCS (Attention - Relevance - Confidence - Satisfaction) model in a collaborative learning environment. They take advantage of the flexibility of the Moodle platform as it promotes active learning through the generation of content in a collaborative manner. Badea et al. [35, 36] propose a solution to overcome the grading issues during the peer assessment process by extending the “*Workshop*” plugin<sup>14</sup> from Moodle. They use an API based on the Bayesian Network model, which computes the students’ Competence and Assessment capability during a collaborative task.

Verdu et al. [37] developed a Moodle plugin through which students can access social media without losing their LMS presence, and teachers can analyse the students’ social interactions to improve the learning process. Nalli et al. [38] propose a plugin for creating heterogenous groups to enhance students’ performance in collaborative learning activities. Constapel et al. [39] support collaborative work through a plugin that supervises and provides intelligent feedback to students with the purpose of improving teamwork and interactions. Hasan et al. [40] explore Moodle’s game mechanics and develop a gamified discussion environment suitable for collaborative learning.

**Microlearning activities** are the most versatile since they take various forms, such as games, short videos, quizzes, or flashcards. Multiple plugins offer the possibility of creating quizzes, questionnaires or flashcards. Out of these, “*H5P*”<sup>15</sup>, and “*Lesson*”<sup>16</sup> are the most used, offering interactive and multimedia-rich lessons that make it easier for learners to focus on small chunks of information at once. Polasek and Javorcik [41], and Hudson [42] test the effectiveness of microlearning using the H5P plugin. Their results showed that this approach creates a student-centred environment that motivates students to engage and be more proactive.

Some Moodle plugins offer support for **gamification**, either by creating activities presented in a game format, such as crosswords, hangman and quizzes<sup>17</sup> or by gamifying the entire course through leaderboards, points or badges. The “*Quizventure*” plugin<sup>18</sup> provides the tools to create quizzes with game mechanics such as time limits, lives and power-ups. “*Block Game*”<sup>19</sup> adds gamification elements to a course through avatars, rankings and points.

---

<sup>12</sup> [https://moodle.org/plugins/mod\\_collabora](https://moodle.org/plugins/mod_collabora).

<sup>13</sup> [https://moodle.org/plugins/mod\\_bigbluebuttonbn](https://moodle.org/plugins/mod_bigbluebuttonbn).

<sup>14</sup> [https://docs.moodle.org/35/en/Using\\_Workshop](https://docs.moodle.org/35/en/Using_Workshop).

<sup>15</sup> [https://moodle.org/plugins/mod\\_hvp](https://moodle.org/plugins/mod_hvp).

<sup>16</sup> [https://docs.moodle.org/35/en/Lesson\\_activity](https://docs.moodle.org/35/en/Lesson_activity).

<sup>17</sup> [https://moodle.org/plugins/mod\\_game](https://moodle.org/plugins/mod_game).

<sup>18</sup> [https://moodle.org/plugins/mod\\_quizgame](https://moodle.org/plugins/mod_quizgame).

<sup>19</sup> [https://moodle.org/plugins/block\\_game](https://moodle.org/plugins/block_game).

The plugin proposed by Kotama et al. [43] enhances Moodle “Quiz” plugin to create a multiplayer matching quiz. Songkram et al. [44] present an integration of Virtual 3D Classrooms into Moodle environment. Zaric et al. [45] developed a gamified recommendation system to assist students during the learning process and foster self-regulated strategies.

In their review, Sinnott and Xia [46] evaluate the effectiveness of the Moodle plugin “Level Up”<sup>20</sup> in increasing the engagement and motivation of the students. The plugin creates a gamified learning space for the students by adding game elements, such as badges, points and leaderboards. The authors identified the most significant weakness as the difficulty of creating an actual educational game in Moodle environment, which is more complex than the simple tasks available.

Overall, the growing body of research conducted in this area focuses on the extension of Moodle capabilities through plugins to overcome shortcomings or implement novel strategies aimed to enhance the learning process.

## 4 Conclusions and Perspectives

Integrating learnersourcing, microlearning, and gamification in LMSs like Moodle holds great promise for promoting student-centred learning and aligns with Sir Ken Robinson’s ideas for education. This paper sheds light on Moodle’s existing capabilities and limitations in facilitating these pedagogical approaches. The results of our literature review indicate that there is still much work to be done to fully realize the potential of Moodle as a student-centred LMS. However, the findings of this paper suggest that with suitable investments in technology and pedagogy, it may be possible to achieve a more creative, innovative, and engaging learning experience for students.

The analysis of the state-of-the-art also underlines several challenges. Rodriguez et al. [14], and Retnowati et al. [15] identified in their research that instructors face difficulties in choosing appropriate activities that will suit the need of the students and will help them achieve the goals of the course at the same time, since it can be time-consuming and not suitable for every student. Based on the studies conducted by Singh et al. [12] and Moore et al. [13], there is also a challenge to motivate the students to contribute to the learning process and create high-quality resources since the students sometimes require assistance for complex tasks. From the studies conducted by Rodriguez et al. [14], and Moore et al. [13], we can mention the issues related to assessment, especially in the context of learnersourcing and collaborative learning, since the current educational system still requires quantitative data, such as grades.

To answer these challenges, future steps in our research will be the development of several Moodle plugins, which will help us provide solutions to the issues identified during our review.

We stated earlier in this paper that one of the most significant issues that teachers encounter is that creating learning activities is time-consuming and

---

<sup>20</sup> [https://moodle.org/plugins/block\\_xp](https://moodle.org/plugins/block_xp).

requires careful planning. Our solution is a plugin that will assist teachers in the retrieval of activities suitable for their courses. This system will work with multiple possible settings to suit the user's needs. The first approach might be a simple parameter-based search on existing Learning Object Repository, as De Medio et al. [47] propose in their paper, but targeted on micro activities and gamified learning. The direct extension of this solution is an ML algorithm that will scan the existing resources on the course (lessons, labs, books) and recommend activities suitable for the existing content. This will address the concern of using microlearning activities complementary to other resources and activities as highlighted by Taylor et al. [27]. OpenAI's APIs, such as ChatGPT [48], could also serve as a great virtual assistant for the teacher in choosing suitable activities.

The second issue identified is related to students' engagement during a collaborative task and the assessment of their work, as this activity is prone to various threats, such as knowledge bias and self-regulation abilities. Bause et al. [18] propose a solution to balance the knowledge bias of the students using a multi-touch table to prove the efficiency of collaborative learning in a controlled setting where students are provided with different pieces of information that they have to put together to solve the given task.

Inspired by their idea, we propose an activity plugin that creates a virtual private space for each student where they can take notes, work on assignments and develop ideas, and a common public space, where the students can share the notes from their private space and potentially discuss it with their peers. The private space will take the form of a list of notes that could be created or edited using an essential toolbox for text editing, media insertion, attachments, and other interactive features. These notes could be kept in the private space or moved to the public space as topics on a discussion forum. Some potential use cases for this plugin are knowledge sharing, collaborative work and activity tracking that could assess students' critical thinking and creativity in such settings. Since we found out from Singh et al.'s [12] research that some students might require guidance, we also propose the integration of ChatGPT [48] in the students' private space to fill in the knowledge gaps the students might have.

Another future direction for our research could be related to the assessment of skills such as creativity and critical thinking which is a complex topic of discussion that was intensively studied [13, 14, 35, 36]. Some of the challenges that we identified are the difficulty of quantifying and comparing students performance and the subjectivity of assessment that can not be standardised. There is also the need for time and resources, such as specialised training for the teachers, suitable tools and materials, which often requires financial support that many institutions do not have, but technology can play a significant role in making these solutions more accessible.

## References

1. Robinson, K., Robinson, K.: *Imagine If . . . : Creating a Future for Us All*. Penguin Publishing Group (2022). <https://books.google.ro/books?id=O4c3EAAAQBAJ>

2. Khosravi, H., Demartini, G., Sadiq, S., Gasevic, D.: Charting the design and analytics agenda of learnersourcing systems. In: 11th International Learning Analytics and Knowledge Conference, pp. 32–42 (2021). <https://doi.org/10.1145/3448139.3448143>
3. De Gagne, J., Park, K., Hall, K., Woodward, A., Yamane, S., Kim, S.: Microlearning in health professions education: scoping review. *JMIR Med. Educ.* **5**(2), e13997 (2019). <https://doi.org/10.2196/13997>
4. Denny, P., McDonald, F., Empson, R., Kelly, P., Petersen, A.: Empirical support for a causal relationship between gamification and learning outcomes. In: The 2018 CHI Conference, pp. 1–13 (2018). <https://doi.org/10.1145/3173574.3173885>
5. Ka Yuk Chan, C.: Laboratory learning. In: Seel, N.M. (eds) Encyclopedia of the Sciences of Learning, vol. 4, pp. 1705–1708. Springer, Boston (2012). [https://doi.org/10.1007/978-1-4419-1428-6\\_966](https://doi.org/10.1007/978-1-4419-1428-6_966)
6. Aikina, T., Bolsunovskaya, L.: Moodle-based learning: motivating and demotivating factors. *Int. J. Emerg. Technol. Learn. (iJET)* **15**, 239–248 (2020). <https://doi.org/10.3991/ijet.v15i02.11297>
7. Shchedrina, E., Valiev, I., Sabirova, F., Babaskin, D.: Providing adaptivity in Moodle LMS courses. *Int. J. Emerg. Technol. Learn. (iJET)* **16**, 95–107 (2021). <https://doi.org/10.3991/ijet.v16i02.18813>
8. Leka, L., Kika, A.: Enhancing Moodle to adapt to students different learning styles. In: International Conference on Recent Trends and Applications in Computer Science and Information Technology, pp. 186–189 (2018)
9. Kika, A., Leka, L., Maxhelaku, S., Ktona, A.: Using data mining techniques on Moodle data for classification of student's learning styles. In: 47th International Academic Conference, Prague, pp. 26–33 (2019). <https://doi.org/10.20472/IAC.2019.047.010>
10. Kolekar, S.V., Pai, R.M., MM, M.P.: Adaptive user interface for Moodle based e-learning system using learning styles. *Procedia Comput. Sci.* **135**, 606–615 (2018). <https://doi.org/10.1016/j.procs.2018.08.226>
11. Almenara, J., Arancibia Muñoz, M., Del Prete, A.: Technical and didactic knowledge of the Moodle LMS in higher education. Beyond functional use. *J. New Approaches Educ. Res.* **8**, 25–33 (2019). <https://doi.org/10.7821/naer.2019.1.327>
12. Singh, A., Brooks, C., Doroudi, S.: Learnersourcing in theory and practice: synthesizing the literature and charting the future. In: Proceedings of the Ninth ACM Conference On Learning @ Scale, pp. 234–245 (2022). <https://doi.org/10.1145/3491140.3528277>
13. Moore, S., Stamper, J., Brooks, C., Denny, P., Khosravi, H.: Learnersourcing: student-generated content @ scale. In: Proceedings of the Ninth ACM Conference On Learning @ Scale, pp. 259–262 (2022). <https://doi.org/10.1145/3491140.3528286>
14. Rodríguez, A.I., Riaza, B., Gómez, M.S.: Collaborative learning and mobile devices: an educational experience in Primary Education. *Comput. Hum. Behav.* **72**, 664–677 (2016). <https://doi.org/10.1016/j.chb.2016.07.019>
15. Retnowati, E., Ayres, P., Sweller, J.: Collaborative learning effects when students have complete or incomplete knowledge. *Appl. Cogn. Psychol.* **32**(6), 681–692 (2018). <https://doi.org/10.1002/acp.3444>
16. Janssen, J., Kirschner, P.A.: Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Educ. Technol. Res. Dev.* **68**(2), 783–805 (2020). <https://doi.org/10.1007/s11423-019-09729-5>

17. Slof, B., Leeuwen, A.V., Janssen, J., Kirschner, P.: Mine ours and yours: whose engagement and prior knowledge affects individual achievement from online collaborative learning. *J. Comput. Assist. Learn.* **37**(1), 39–50 (2020). <https://doi.org/10.1111/jcal.12466>
18. Bause, I.M., Brich, I.R., Wesslein, A.-K., Hesse, F.W.: Using technological functions on a multi-touch table and their affordances to counteract biases and foster collaborative problem solving. *Int. J. Comput.-Support. Collaborative Learn.* **13**(1), 7–33 (2018). <https://doi.org/10.1007/s11412-018-9271-4>
19. Schultz, W.: Multiple reward signals in the brain. *Nat. Rev. Neurosci.* **1**(3), 199–207 (2000). <https://doi.org/10.1038/35044563>
20. Garshasbi, S., Yecies, B., Shen, J.: Microlearning and computer-supported collaborative learning: an agenda towards a comprehensive online learning system. *STEM Educ.* **1**, 225–255 (2021). <https://doi.org/10.3934/steme.2021016>
21. Sweller, J., Ayres, P., Kalyuga, S.: Cognitive Load Theory. Springer, New York (2011). <https://doi.org/10.1007/978-1-4419-8126-4>
22. Sweller, J.: Cognitive load theory and educational technology. *Educ. Technol. Res. Dev.* **68**(1), 1–16 (2019). <https://doi.org/10.1007/s11423-019-09701-3>
23. Mayer, R.E.: Cognitive Theory of Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, pp. 43–71. Cambridge University Press, Cambridge (2014) <https://doi.org/10.1017/CBO9780511816819>
24. Ryan, R.M., Deci, E.L.: Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. The Guilford Press, New York (2017). <https://doi.org/10.1521/978.14625/28806>
25. Bouillion, L.M., Gomez, L.M.: Connecting school and community with science learning: real world problems and school-community partnerships as contextual scaffolds. *J. Res. Sci. Teach.* **38**(8), 878–898 (2001). <https://doi.org/10.1002/tea.1037>
26. Kirschner, P.A., Sweller, J., Kirschner, F., Zambrano R., J.: From cognitive load theory to collaborative cognitive load theory. *Int. J. Comput.-Support. Collaborative Learn.* **13**(2), 213–233 (2018). <https://doi.org/10.1007/s11412-018-9277-y>
27. Taylor, A., Hung, W.: The effects of microlearning: a scoping review. *Educ. Technol. Res. Dev.* **70**(1), 1–33 (2022). <https://doi.org/10.1007/s11423-022-10084-1>
28. Landers, R.: Developing a theory of gamified learning: linking serious games and gamification of learning. *Simul. Gaming* **45**(6), 1–17 (2015). <https://doi.org/10.1177/1046878114563660>
29. Van Roy, R., Zaman, B.: Need-supporting gamification in education: an assessment of motivational effects over time. *Comput. Educ.* **127**(2), 283–297 (2018). <https://doi.org/10.1016/j.compedu.2018.08.018>
30. Deci, E., Ryan, R.: Self-determination theory. International encyclopedia of the social & behavioral sciences (Second Edition), pp. 486–491 (2015) <https://doi.org/10.1016/B978-0-08-097086-8.26036-4>
31. Gordon, E., Barnett, K., Cooper, N., Tran, N., Williams, L.: An “integrative neuroscience” platform: application to profiles of negativity and positivity bias. *J. Integr. Neurosci.* **7**, 345–66 (2008). <https://doi.org/10.1142/S0219635208001927>
32. Majuri, J., Koivisto, J., Hamari, J.: Gamification of education and learning: a review of empirical literature. In: Proceedings of the 2nd International GamiFIN Conference, pp. 11–19 (2018)
33. Manzano León, A., et al.: Between level up and game over: a systematic literature review of gamification in education. *Sustainability* **13**(4), 1–14 (2021). <https://doi.org/10.3390/su13042247>

34. Karampa, V., Paraskeva, F.: A motivational design of a flipped classroom on collaborative programming and STEAM. *Commun. Comput. Inf. Sci.* **870**, 226–238 (2018). [https://doi.org/10.1007/978-3-319-95522-3\\_19](https://doi.org/10.1007/978-3-319-95522-3_19)
35. Badea, G., Popescu, E., Sterbini, A., Temperini, M.: A service-oriented architecture for student modeling in peer assessment environments. In: Hao, T., Chen, W., Xie, H., Nadee, W., Lau, R. (eds.) SETE 2018. LNCS, vol. 11284, pp. 32–37. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-03580-8\\_4](https://doi.org/10.1007/978-3-030-03580-8_4)
36. Badea, G., Popescu, E., Sterbini, A., Temperini, M.: Integrating enhanced peer assessment features in Moodle learning management system. In: Foundations and Trends in Smart Learning. LNET, pp. 135–144. Springer, Singapore (2019). [https://doi.org/10.1007/978-981-13-6908-7\\_19](https://doi.org/10.1007/978-981-13-6908-7_19)
37. Verdú, M., De Castro, J., Regueras, L., Corell, A.: MSocial: practical integration of social learning analytics into Moodle. *IEEE Access* **9**, 23705–23716 (2021). <https://doi.org/10.1109/ACCESS.2021.3056914>
38. Nalli, G., Amendola, D., Perali, A., Mostarda, L.: Comparative analysis of clustering algorithms and Moodle plugin for creation of student heterogeneous groups in online university courses. *Appl. Sci.* **11**(13), 5800 (2021). <https://doi.org/10.3390/app11135800>
39. Constapel, M., Doberstein, D., Hoppe, H., Hellbruck, H.: IKARion: enhancing a learning platform with intelligent feedback to improve team collaboration and interaction in small groups. In: 18th International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 1–10 (2019). <https://doi.org/10.1109/ITHET46829.2019.8937348>
40. Al-Delawi, H., Cemal Nat, M., Vanduhe, V.: Gamified collaborative environment in Moodle. *IEEE Access* **7**, 89833–89844 (2019). <https://doi.org/10.1109/ACCESS.2019.2926622>
41. Polasek, R., Javorcik, T.: Results of pilot study into the application of microlearning in teaching the subject computer architecture and operating system basics. In: International Symposium on Educational Technology, Czech Republic, pp. 196–201 (2019). <https://doi.org/10.1109/ISET.2019.00048>
42. Hudson, J.: Flipping lessons with Moodle: using the H5P Moodle plugin to deliver online sessions. In: 12th Annual University of Glasgow Learning and Teaching Conference, Glasgow, UK (2019). <https://eprints.lancs.ac.uk/id/eprint/168168>
43. Kotama, I., Saputra, K., Linawati, L.: Proposed model of multiplayer matching game plugins using websocket in Moodle. *Int. Jo. Emerg. Technol. Learn. (iJET)* **14**(11), 194–201 (2019). <https://doi.org/10.3991/ijet.v14i11.10190>
44. Songkram, N., Songkram, N., Chootongchai, S., Samanakupt, T.: Developing students' learning and innovation skills using the virtual smart classroom. *Int. J. Emerg. Technol. Learn. (iJET)* **16**(4), 34–51 (2021). <https://doi.org/10.3991/ijet.v16i04.15221>
45. Zaric, N., Deb, A., Schroeder, U.: Learning strategizer - LESTRAT: a gamified moodle plugin to support self-regulated learners in the forethought phase. In: 13th Annual International Conference of Education, Research and Innovation, pp. 1637–1643 (2020). <https://doi.org/10.21125/iceri.2020.0415>
46. Sinnott, M., Xia, L.: A review of the Moodle gamification plugin “level up”: using a Moodle plugin to gamify learning of academic vocabulary. *Int. J. Comput.-Assist. Lang. Learn. Teach.* **10**(3), 89–95 (2020). <https://doi.org/10.4018/IJCALLT.2020070107>

47. Medio, C., Limongelli, C., Sciarrone, F., Temperini, M.: MoodleREC: a recommendation system for creating courses using the Moodle e-learning platform. *Comput. Hum. Behav.* **104**, 106168 (2020). <https://doi.org/10.1016/j.chb.2019.106168>
48. OpenAI: GPT-3: Generative Pretrained Transformer, vol. 3 (2021). <https://beta.openai.com/docs/models/gpt-3>



# Towards a Social Learning Analysis Using Learning Management System and Learning Experience to Predict Learners' Success

Hiba Gharbaoui<sup>1</sup>(✉), Khalifa Mansouri<sup>1</sup>, and Franck Poirier<sup>2</sup>

<sup>1</sup> Modeling and Simulation of Intelligent Industrial Systems Laboratory – M2S2I ENSET  
Mohammedia, University Hassan II of Casablanca, Casablanca, Morocco

hiba.gharbaoui-etu@etu.univh2c.ma

<sup>2</sup> Lab-STICC, University Bretagne Sud, Vannes, France  
franck.poirier@univ-ubs.fr

**Abstract.** Online learning is an alternative to the traditional (face-to-face) educational system. Contrary to traditional teaching (classroom), distance learning is characterized by the lack of physical contact between instructors and students and also between students within a classroom. Unfortunately, some learners may fail or even stop learning online quite quickly. In addition, learners find it more difficult to learn, not because there is not enough content, but because there is too much and they cannot find what is useful and up-to-date. This study aims to prove the importance of using social and adaptive learning to overcome the problems that learners face in online learning platforms, we propose a model to determine the effects of using the two types of platforms in online learning, namely Learning Management System (LMS) and Learning Experience Platform (LXP) for two groups of learners in Moroccan higher education applying social learning analysis (SLA). Specifically, in this proposal we attempt to analyze data from the LMS and LXP platforms to build knowledge models about students and determine their learning styles based on their interactions on these platforms, improve their learning experience and predict the conditions that favor their progress and success in order to recreate the conditions for quality social learning and help learners find possible paths to success in the online learning platforms in Morocco.

**Keywords:** E-learning · social learning · learning analysis · Learning Management System · Learning Experience Platform

## 1 Introduction

The social learning hypothesis, proposed by Albert Bandura, accentuates the significance of noticing, displaying and imitating the ways of behaving, perspectives, and profound reactions of others. Social learning theory examines the interaction of environmental and cognitive factors in influencing human learning and behavior [1].

In order to better understand the role of learners' social presence, this study tries to find out how these interactions work this will give us a lot of information to adapt our

teaching to the way our students learn. The approach aims to enhance the effectiveness of student participation, for this we will use Learning Management System (LMS) and Learning Experience System (LXP) which are two types of platforms in online learning [2], and specific social learning analytics methods to analyze the data from both platforms such as the connections and interactions between students in order to predict the conditions that enable their progress and find the most appropriate type of learning for each. In this paper, Sect. 2 presents a theoretical background to understand the concepts of social presence in online learning and also presents the methods of social learning analysis as well as an explanation of LMS and LXP online learning platforms. Section 3 to discuss the research that was being done in this area. Our proposed system is to analyze data from the LMS and LXP platforms using social learning analytics in order to predict the conditions favoring learner success in e-learning platforms in Morocco this will be discussed in Sect. 4. The last section presents a conclusion and identifies our perspectives and the challenges that need to be addressed in the future.

## 2 Theoretical Background

### 2.1 Social Presence in Online Learning (Social Networks / Social Networks Analytics)

Social networks allow people to quickly exchange information and knowledge and have been promoted with various applications for interpersonal interaction, communication and sharing [3, 4] Furthermore, learning is a social and collective outcome that is achieved through conversations, knowledge dissemination and social networks [5]. According to Christine Greenhow, “Now that we know what skills students are learning on the Internet, we need to encourage and help them to develop them further” [6]. Teenagers are not always aware of developing such abilities, which can be useful from both an academic and professional point of view. Going further, we also learn that the reduction of the digital divide between social backgrounds would take place with the use of these community media [7].

In recent years, Social Network Analysis (SNA) has attracted increasing attention as a methodology in the field of education [8]. SNA is often applied to analyze interconnected forum user networks to study the structural patterns and underlying relational organization of learning communities in online courses. The participation gap between active and inactive learners has been studied by Poquet and Dawson [9, 10]. Thoughtfully, Social Network Analysis (SNA) considers each actor as a node of the whole network, Whereas the various connections between them are conceptualized as lines interfacing the hubs and called edges or connections [11].

### 2.2 Learning Analytics / Social Learning Analytics

Currently, it is possible to retrieve social data from various social sources, such as Twitter, Facebook, Instagram, YouTube and learning platforms like LMSs and LXPs. This data can be the main source of information to do the Learning analytics (LA) and Social Learning analytics (SLA).

Learning analytics (LA) is a discipline that aims to exploit the hidden potential of interaction data generated by the use of online learning environments [12]. According to Sclater [13], The goal is to analyze the digital traces left by learners in order to better understand them and optimize learning. In fact, learning analytics focuses on the detection, collection, analysis and exploitation of digital traces left by learners in their daily activities, in order to understand and optimize learning and the different environments in which it occurs [14, 15]. On the other hand, Learning Analytics (LA) is a very important discipline of artificial intelligence, due to its ability to exploit learning data, generate information and knowledge to improve learning processes.

One of the subclasses of Learning Analytics (LA) is Social Learning Analytics (SLA), The SLA focuses on relationship building and provides an opportunity to identify interventions that can enhance the potential of the network to support stakeholder learning [2]. Social Learning Analytics (SLA) seeks to gain hidden insights from large amounts of data, usually educational in nature. SLA focuses primarily on Social Network Analysis (SNA) and web analysis to uncover patterns of interaction and behavior of educational social actors. Social Learning Analytic (SLA) is a subfield of LA, which focuses on how students acquire knowledge in a social context, often without being controlled, in collaboration with other [12, 16, 17].

### 2.3 Learning Management System / Learning Experience Platform

Supporting learners via web-based platforms is not new. For decades, learning management systems (LMS) such as Blackboard and Moodle have served as the foundation for online learner support. LMSs provide a comprehensive digital learning environment where teaching and learning materials can be delivered, stored, and assessed. Often these systems are also linked to administrative functions such as student registration and student performance tracking [18].

However in recent years, third generation platforms called LXP have been developed that claim to provide a more interactive, current and social learning experience, the so-called learning experience platforms recently begun to replace the use of LMSs in higher education. LXPs combine aspects of a social media platform with a virtual learning environment. The term was coined by Josh Bersin, who also compared these new platforms to "Netflix for learning" and the expression is still used today [19]. Uploading content is quick and easy for all users, with the integration of a limited number of applications - an LXP functions as a platform rather than a system with different building blocks like an LMS.

## 3 Related Works

Social networks can enhance peer-to-peer and educator-student interaction, especially when used to search for data, publish articles, engage in discussions and communicate [20]. Other than social networks, there are user-generated content (UGC) platforms in e-learning such as LMSs and LXPs. Content creation is no longer limited to the course instructor, but in an LXP, everyone creates content [21]. In the LMS, the learning journey is largely linear, with a learner completing one task and then moving on to

the next. An LXP system aims to increase engagement and experience sharing among users, rather than functioning as a repository of information. The recent push in higher education toward online and distance learning, already a trend prior to the pandemic, has accelerated and allowed for a more rapid introduction of LXPs into higher education [18]. To analyze student interactions in online learning platforms, researchers have introduced five categories of analysis. The first two categories are intrinsically social (Social network analysis, Discourse analysis), while the other three are potentially "socialized", i.e., usefully applied in social settings (Content analysis, Alignment analysis and Contextual analysis) [16].

In [17], They suggest analyzing external data from the web and social networks to build knowledge models about students and to improve the learning processes that occur in smart classrooms. For this there are several factors and dimensions to consider that could affect the learning experience (LX) to decrease the failure or dropout rate in Online Learning Environments (OLEs). In [22], the proposed research model consists of 38 factors categorized along 5 dimensions: learner characteristics, instructor characteristics, system characteristics, course characteristics, and social aspects. This study [23] examined learner satisfaction with e-learning systems in higher education. Results indicate that diversity of assessments, course flexibility, social interactions, system quality, and perceived usefulness have positive effects on learner satisfaction. The results of this research will be useful to both institutions and practitioners of e-learning systems.

## 4 Proposed System

Our model (Fig. 1) focuses on social learning as a solution to our problem in order to predict the success or failure of learners in Moroccan universities (higher education), improve their learning experience and suggest the type of platform suitable for each student. For this we proposed to use two e-learning platforms LMS and LXP, LMS to focus more on the administration of learning than on the experience, whereas LXP to focus on the ability to facilitate a personalized, learner-oriented learning.

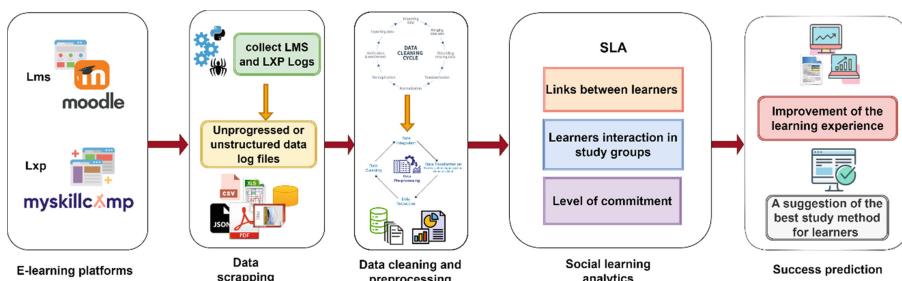
The chosen LMS is Moodle which is a learning platform designed to provide teachers, administrators and learners with a single robust, secure and integrated system for creating personalized learning environments [24], and the chosen LXP is Myskillcamp [25] which is an all-in-one learning experience platform. One can create, edit, import and purchase courses directly in Myskillcamp, build and launch learners on personalized learning paths and even check progress and completion rates with their analytics.

We chose these two platforms because they are designed for teaching and learning, easy to use, free, with no licensing costs, their multilingual capabilities ensure that no language barrier hinders online learning, they are flexible and customizable platforms that can serve small classrooms as well as large organizations, robust, secure and private with a large number of resources available.

So we will collect the data from these platforms using the data retrieval technique, after this generated data will be cleaned and preprocessed. There are several tools to facilitate data mining, namely MMT (Moodle Mining Tool) and ADE (Automatic Data Extraction) [26]. Generally speaking, there are three methods of data extraction, the first method of extraction is xAPI. For example, Watershed which has a way to extract data

from virtually any LRS "Learning Record Store" which is a database used to store all the information of a learning ecosystem in xAPI format, even if it is not fully compliant. And many LRSs have xAPI functionality that can pass xAPI data to Watershed as an activity provider. An API connection is used for the second approach. Most platforms include APIs that enable access to their data or enable the creation and download of reports. API access involves technical know-how, comprehension of the applicable documentation, as well as the necessary tools and credentials. The final approach uses an Excel or CSV file. You may get a report from most platforms and explore with the data in Excel or another data tool of your choosing [27].

The data that will be extracted is the content of the links between students, which means the interests that connect them, their interactions in the learning groups and the level of engagement in each group. After cleaning the data we will have interpretable data that will be analyzed using Social Learning Analytics (SLA) where we seek to obtain information hidden in large amounts of educational data. This concept focuses mainly on Social Network Analysis (SNA) and web analysis, to discover interaction and behavior patterns of educational social actors [17].



**Fig. 1.** Social learning as a solution to predict the success of learners

The analysis will therefore focus on three main axes “the content of the links between students”, “the interaction of the learners in the work groups” and the “level of commitment (increase or decrease)” in the two LMS and LXP platforms, we will therefore look for the relationship between these axes and the success rate of the learners.

This will allow us to conclude the role of social learning in the success of students in these platforms specifically their interactions and participations in the online discussion forums of the courses and thus track the progress of learners and give them targeted and better feedback, in order to provide them with an improved learning experience that will help them succeed and self-regulate.

## 5 Conclusion

The social environment in which an e-learning action takes place can play a major role in a learner's persistence and success. Peer and instructor opinions can influence learners' choices in using a learning system. This study aims to test the hypothesis of the role and importance of the use of social and adaptive learning to overcome the problems

that learners face in online learning platforms and its role in their success, we propose a model to determine the effects of the use of the two main online learning platforms "LMS and LXP" in Moroccan higher education, We seek to analyze the interactions and social presence in learning groups as well as the level of engagement of learners, how their activity is evaluated on the basis of social learning analysis (SLA) in order to predict the success and the best types of learning for each learner, thus improving their learning experience.

Our future goal is to conduct an experiment where we will apply SLA concepts to analyze data extracted from the LMS and LXP platforms.

From this analysis, we will dedicate the importance of social learning in e-learning environments and suggest the best learning method for each learner in order to predict and guarantee the success of learners in e-learning environments in Moroccan educational systems.

## References

1. « Albert Bandura's Social Learning Theory | Simply Psychology ». <https://www.simplpsychology.org/bandura.html>
2. Díaz-Lázaro, J.J., Fernández, I.M.S., Sánchez-Vera, M. del M.: Social learning analytics in higher education. an experience at the primary education stage.J. New Approaches Educ. Res. **6**(2) (juill. 2017). <https://doi.org/10.7821/naer.2017.7.232>
3. Chen, C.-M., Hsieh, Y.-L., Hsu, S.-H.: Mining learner profile utilizing association rule for web-based learning diagnosis.Expert Syst. Appl. **33**(1), 6-22 (juill. 2007). <https://doi.org/10.1016/j.eswa.2006.04.025>
4. Wu, T.-T.: Using smart mobile devices in social-network-based health education practice: a learning behavior analysis.Nurse Educ. Today **34**(6), 958-963 (juin 2014). <https://doi.org/10.1016/j.nedt.2014.01.013>
5. Brown, J.S., Duguid, P.: Organizational Learning and communities-of-practice: toward a unified view of working, learning, and innovation.Organ. Sci. **2**(1), 40-57 (févr. 1991). <https://doi.org/10.1287/orsc.2.1.40>
6. Article 4 - Les réseaux sociaux: outils d'apprentissage en devenir? - Article 1 – Apprenance. <https://b.21-bal.com/doc/2311/index.html?>
7. Pinte, J.-P.: Vers des réseaux sociaux d'apprentissage en éducation. Cah. Dyn. **47**(2), 82–86 (2010). <https://doi.org/10.3917/lcd.047.0082>
8. de Laat, M., Lally, V., Lippinen, L., Simons, R.-J.: Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis.Int. J. Comput.-Support. Collab. Learn. **2**(1), 87-103 (mars 2007). <https://doi.org/10.1007/s11412-007-9006-4>
9. Oleksandra, P., Shane, D.: Untangling MOOC learner networks. In:Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 208–212. New York, NY, USA (avr. 2016). <https://doi.org/10.1145/2883851.2883919>
10. Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., Liu, M.: Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis.Comput. Hum. Behav. **115**, 106582 (févr. 2021). <https://doi.org/10.1016/j.chb.2020.106582>
11. Hernández-García, Á., González-González, I., Jiménez-Zarco, A.I., Chaparro-Peláez, J.: Visualizations of online course interactions for social network learning analytics.Int. J. Emerg. Technol. Learn. IJET **11**(07) (juill. 2016). <https://doi.org/10.3991/ijet.v11i07.5889>

12. Ferguson, R.: Learning analytics: drivers, developments and challenges. *Int. J. Technol. Enhanc. Learn.* **4**(5-6), 304-317 (janv. 2012). <https://doi.org/10.1504/IJTEL.2012.051816>
13. 13 « Learning Analytics Explained | Niall Sclater | Taylor & Francis Group ». <https://doi.org/10.4324/9781315679563>. Consulté le 18 juin 2022
14. Safsouf, Y., Mansouri, K., Poirier, F.: TABAT: design and experimentation of a learning analysis dashboard for teachers and learners. *J. Inf. Technol. Educ. Res.* **20**, 331-350 (juill. 2021)
15. Best Coursera Courses & Certificates in 25 categories [2022] (7 mai 2022). <https://www.codespaces.com/coursera.html>. Consulté le 18 juin 2022
16. 16 Ferguson, R., Buckingham Shum, S.: Social learning analytics: five approaches. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 23-33, New York, NY, USA (avr. 2012). <https://doi.org/10.1145/2330601.2330616>
17. Aguilar, J., Buendia, O., Pinto, A., Gutiérrez, J.: Social learning analytics for determining learning styles in a smart classroom. *Interact. Learn. Environ.* **30**(2), 245-261 (févr. 2022). <https://doi.org/10.1080/10494820.2019.1651745>
18. Cockrill, A.: From Learning Management Systems to Learning Experience Platforms: Do they keep what they promise? Reflections on a rapidly changing learning environment. (mai 2021). <https://doi.org/10.25401/cardiffmet.14611932.v1>
19. A New World of Corporate Learning Arrives: And It Looks Like TV – JOSH BERSIN. [https://joshbersin.com/2017/06/a-new-world-of-corporate-learning-arrives-and-it-looks-like-tv/?zd\\_source=hrt&zd\\_campaign=3171&zd\\_term=chiradeepbasumallick](https://joshbersin.com/2017/06/a-new-world-of-corporate-learning-arrives-and-it-looks-like-tv/?zd_source=hrt&zd_campaign=3171&zd_term=chiradeepbasumallick)
20. Wu, T.-T.: Using smart mobile devices in social-network-based health education practice: a learning behavior analysis. *Nurse Educ. Today* **34**(6), Art. n° 6 (juin 2014). <https://doi.org/10.1016/j.nedt.2014.01.013>
21. LMS vs. LXP: How and Why They're Different. <https://unboxedtechnology.com/blog/learning-management-system-lms/>
22. Safsouf, Y., Mansouri, K., Poirier, F.: A New Model of Learner Experience in Online Learning Environments. In: Rocha, Á., Serrhini, M. (eds.) *EMENA-ISTL 2018. SIST*, vol. 111, pp. 29–38. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-03577-8\\_4](https://doi.org/10.1007/978-3-030-03577-8_4)
23. Safsouf, Y., Mansouri, K., Poirier, F.: Smart learning environment, measure online student satisfaction: a case study in the context of higher education in Morocco. In: *2020 International Conference on Electrical and Information Technologies (ICEIT)*, pp. 1–5 (mars 2020). <https://doi.org/10.1109/ICEIT48248.2020.9113189>
24. Moodle - Open-source learning platform | Moodle.org. <https://moodle.org/>
25. About us | myskillcamp. <https://myskillcamp.com/about-us>
26. Ali, F., Ng, S.: Moodle data retrieval for educational data mining. *Int. J. Sci. Eng. Technol.* **4**, 523–525 (nov. 2015) <https://doi.org/10.17950/ijset/v4s11/1105>
27. How Do I Extract Learning Data from My LMS or LXP? <https://www.watershedlrs.com/blog/learning-data-ecosystems/how-to-extract-learning-data-from-lrs/>



# Explanatory Didactic Dialogue in the Intelligent Tutoring Systems Based on the Cross-Cutting Approach

Viktor Uglev<sup>(✉)</sup>

Siberian Federal University, Zheleznogorsk, Russia  
[vaugelev@sfu-kras.ru](mailto:vaugelev@sfu-kras.ru)

**Abstract.** The paper deals with the problem of developing a dialogue in the interaction of a learner and Intelligent Tutoring Systems. We demonstrate the approach to synthesizing the text of a didactic dialogue on the example of using the cross-cutting approach to decision-making support by a tutoring system planner. The parametric mapping method is used for this purpose. The maps are used not only for synthesizing the explanatory text of a didactic dialogue, but also for visual representation of a learning situation in the Cognitive Maps of Knowledge Diagnosis notation. We demonstrate the process of entering a didactic dialogue and transition between dialogue forms on the example of a graduate student (course “Simulation Modeling”). The study has shown the interest of the students in receiving explanations of the ITS decisions, including both textual and visual presentation of arguments.

**Keywords:** Intelligent Tutoring Systems · cognitive visualization · explainable AI · mapping · explanation of decisions · dialogue · Cognitive Maps of Knowledge Diagnosis

## 1 Introduction

The dialogue, as a sequence of acts of exchange of verbal and non-verbal information portions between two entities, involves an element of uncertainty. First of all, it concerns the trajectory of discussion. When two people communicate they can effectively exchange information, even if they have different levels of expertise in the topic of discussion. When a person interacts not with another person, but with a machine possessing the elements of artificial intelligence, the dialogue loses its effectiveness. The reasons for this can be the template-based generation of phrases and the structure of the whole explanatory text; the weak connection between the verbal (textual/sound) and non-verbal (visual) components of the dialogue; the difficulties of explaining the situation (not only stating the facts); the difficulties of explaining in detail the decisions made; the difficulties of developing the dialogue in the direction the person is interested in; the difficulties of expanding the subjects of discussion (shifting the focus of discussion, dynamic expansion of the list of the subjects as the interaction progresses);

the problem of including many potentially useful entities in the dialogue (e.g., teacher and tutor models [9]); low degree of personalization (both in role and terminological aspects). The consequence is a psychological fixation to machine responses, therefore its decisions shall be treated with some distrust (the problem of XAI [3]). If the dialogue is synthesized by convolutional neural networks, then the problem of explainability of decisions (the answer to questions “Why?” and “What for?”) becomes fundamental [11]. If an explanatory text is formed based on a certain structure of knowledge (for example, an ontology of a subject area), then it shall be supported by a certain logic of the dialogue’s development. As a rule, this logic is described in the form of a tree of dialogue trajectories with a predetermined structure [13]. It is this structure that makes the dialogue “unnatural”. We consider this problem in more detail on the example of conducting a didactic dialogue between a learner and an Intelligent Tutoring System (ITS).

## 2 Related Works

As a rule, the ITS learning subsystems include an intellectual solver (planner), a set of models (learner, teacher, tutor and e-course), a knowledge base, a database with learning material and sets of practical assignments and tests, a subsystem of translation of the learning material and tests, and a subsystem of conducting a dialogue [11]. A learner does not see the whole model of the course and works with a sequence of learning units as well as with tests and assignments. As a result, ITS protocols record the time-frequency data concerning dealing with the e-course elements, answers to assignments and tests, and questionnaire data, thus forming a digital educational footprint. The object of the dialogue is the current learning situation as a whole or its separate aspects (including the didactic aspect).

The dialogue in the ITS is of interest because the intelligent system transmits knowledge to the person, which he/she does not possess at the proper level *a priori*. This involves the presence of a person’s interest in both the subject of discussion (first of all, the learning material), and the didactic aspect of its organization and presentation. If the learning material itself can be represented as an ontology of the subject area, the didactic aspect (the answer to the questions “Why was this decision made?” and “Why do I need this?”) shall not only satisfy cognitive needs, but also perform a control function (through motivation). Tutoring systems, as a rule, provide a dialogue with a detailed tree of its development trajectories (e.g., [15]). As a consequence, most of the disadvantages mentioned above are also typical for the ITS dialogue subsystems.

The dialogue occurs at the moment of the occurrence of any event in the ITS (for example, the calculation of the grade for a test assignment) or at the initiative of a learner. First, the text explaining the current situation is displayed. It is a kind of entry point into the dialogue, which can be interrupted or developed at the initiative of a learner. According to the nature of the dialogue, it is possible to distinguish between a subject-based dialogue and a didactic dialogue. The former is based on the ontology underlying the learning material of the discipline being studied. The advantage of this approach is the flexibility of the model of knowledge and the ability to answer the questions “Why...?” and “How...?”,

relating directly to the material being studied. The disadvantage of this approach is that the ontology is not transferable to any other course and requires enormous labor for qualitative formalization.

The didactic dialogue is focused on the technology of organizing the learning material and methods of analyzing the learning situation when developing the ITS decisions. The models and methods of organizing such a dialogue can be transferred to a wide range of e-courses if similar models of knowledge representation form their basis. The goal of this type of a dialogue is not only to answer the questions “Why was this decision made?” and “Why do I need this?”, but also to provide an opportunity to develop discussion in the direction of learner’s interest (being limited to the didactic aspect of the learning process). An element of the learning process (a curriculum, a particular e-course, a learning topic, a learning unit, a practical assignment, and a particular term) as well as other learning entities (competencies, standards, personal goals, methods of grading and decision-making, results of learner’s actions, etc.) can be the topic of a didactic dialogue.

The main characteristics of organizing a didactic dialogue are the necessity of taking into account different types of entities (A), consideration of objects in different scales of the learning process (B), the possibility to take into account different points of view on the learning situation (C), presentation of results in different aspects of analysis (D), the possibility of non-verbal supplement of an explanatory text by means of cognitive visualization (E), the possibility to explain decision-making methods by the ITS itself (F). A comparison of various technologies within ITS relative to the didactic dialogue is given in Table 1. Not all of the software solutions given in the table (see column “Examples”) can be attributed to only one type (for example, solutions citeBrus2019 and citeKelsey2015 are hybrid solutions), and some ITS which belong to this or that type and were not included in the review (for example, [10] and many others) have indicators other than the ones given here, which are less important for organization of a didactic dialogue (including technologies of genetic algorithms, artificial neural networks, etc.). In spite of this, it follows from the table that the organization of a didactic dialogue cannot be provided by typical ITS solutions and it is necessary to conduct research of hybrid approaches.

**Table 1.** The implementation of the opportunity to conduct a didactic dialogue within ITS using various technologies

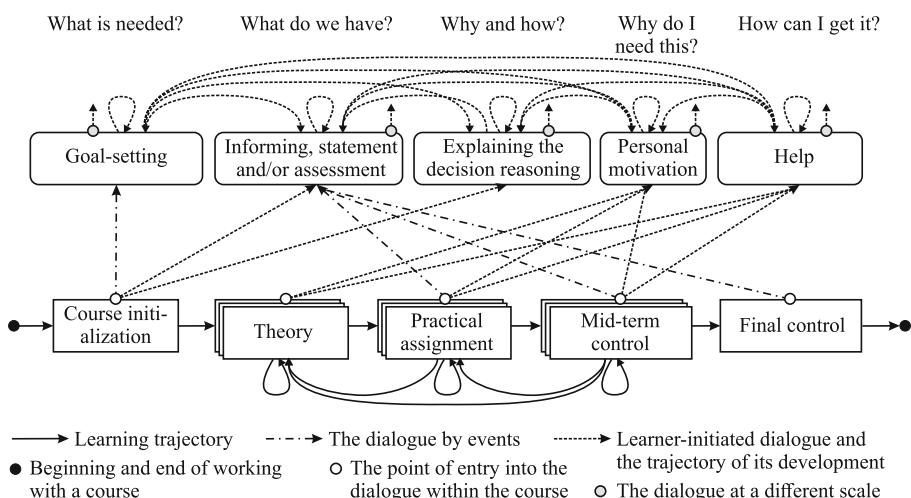
Basic technologies	Examples	A	B	C	D	E	F
Decision making theory	[2]	No	No	No	Yes-No	No	No
Finite-state machines	[14]	No	No	No	No	No	No
Ontologies	[6]	Yes-No	Yes-No	No	Yes	No	Yes
Expert systems	[1]	No	No	No	Yes	No	Yes
Knowledge graphs	[16]	No	No	No	Yes-No	No	Yes-No
Multi-agent systems	[7]	No	No	Yes	No	No	Yes
Cognitive maps [5]	–	Yes	Yes	Yes-No	Yes	Yes	No
Applied semiotics [12]	–	No	No	Yes	Yes	No	Yes

The purpose of this study is to demonstrate the possibility of implementing a didactic dialogue on ITS basis, which is accompanied with cognitive visualization. We will describe the dialogue as an element of the cross-cutting approach (including the visual component), propose an approach to organizing (developing) a didactic dialogue, give an example of applying the proposed approach and formulate recommendations for its application.

### 3 Method

#### 3.1 A Didactic Dialogue in the Context of the Cross-Cutting Approach

The result of a decision made by the ITS planner will be informing a learner. This informing in the form of an explanatory text is the “initialization point” of the potential dialogue. We represented the main stages of the learner’s interaction with the e-course as a sequence of blocks (see the bottom sequence in the diagram shown in Fig. 1). The white circles in the diagram indicate the invocation of a dialogue form both by an event in the ITS (dot-dotted arrow) and by the learner’s initiative (dotted arrow). Both cases focus on explaining the current learning situation in order to increase the learner’s knowledge of it. The learner, in turn, can either interrupt the dialogue or request further clarification.



**Fig. 1.** The diagram of the transition from an e-course (bottom row) to the elements of the didactic dialogue (top row)

According to [18], the cross-cutting approach allows combining the basic independent parameters from the set of parameters considered by the ITS ( $P$ ) when describing a learning situation: the scope of the learning situation  $P^M$  is defined

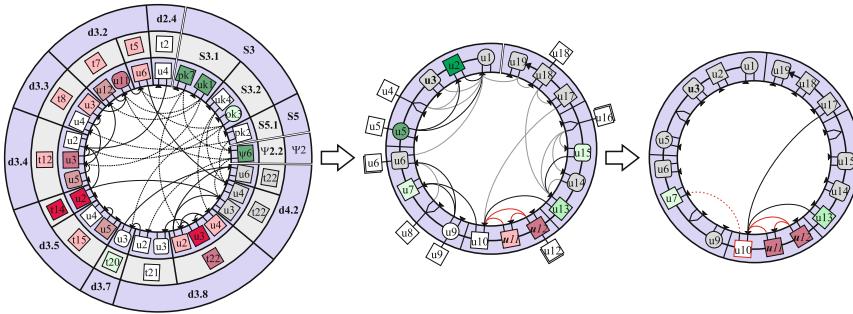
by the following set of states, i.e. 1-nano, 2-micro, 3-meso, 4-macro, 5-meta; the decision-making levels ( $P^L$ ) are defined by the following set of states, i.e. 1-real-time, 2-tactical, 3-strategic (evaluated relative to the goal-setting); aspects of the learning situation consideration ( $P^A$ ) are defined by the following set of states, i.e. 1-subject, 2-competency, 3-normative, 4-target; temporal dynamics ( $P^B$ ) is defined by the following set of states, i.e. 1-current indicators, 2-retrospective data, 3-planned/estimated indicators. Following the generalization according to [17] the set of parameters stored in the digital educational footprint form a parametric map. The analysis of the learning situation is performed relative to each of the models  $Q$  implemented as part of the ITS (opposing “points of view” of the learner’s model and teacher’s model, as well as the trade-off model implemented by a tutor) to produce such decisions  $y$ , which are supported by the system of arguments  $\gamma$ .

The explanatory text  $\gamma$  includes four blocks: a block assessing the current learning situation (statement concerning the object of consideration), a block with information about the most significant difficulties (supplemented, if possible, by non-verbal visualization), a block with suggested actions for a learner (recommendations), and a block with an invitation to develop a dialogue. In the cross-cutting approach it is possible to operate with the parameters  $P^M$ ,  $P^L$ ,  $P^A$  and  $P^B$  independently. This allows changing the scale of the analysis (vertical navigation in the factor space relative to  $P^M$ ), as well as switching between the aspects of the analysis (horizontal navigation on  $P^B$ ).

The parametric map synthesized for the current learning situation without taking into account the combination of values from  $P$  is called a combined map. In order to minimize the load on the ITS intelligent solver when accompanying the dialogue (after formation of the primary explanatory text), the parameter subsets that correspond to the object of discussion (specifying the values of the vector  $P = < m, l, a, b >$ ) are extracted from the combined map. But since it is not necessary to display all the components of the particular map for explanation to a learner, the most significant components are highlighted and displayed during the argumentation  $\gamma$ . Graphically this process can be represented using a Cognitive Map of Knowledge Diagnosis (CMKD). An example of a combined map with the problem points of the learning process for all entities (including external ones) of the course “Simulation Modeling” is shown in Fig. 2 on the left. The central part of the figure contains a particular map for the competency aspect, a simplified representation of which to be used for integration into the dialogue form is shown on the right. Therefore, the verbal and non-verbal representation of the explanatory information is combined in any configuration of didactically significant parameters  $P$  for the learning situation.

### 3.2 Developing a Didactic Dialogue

The cross-cutting approach to decision-making and dialogue support in the ITS implies three ways of developing a dialogue: interactive interaction with the cognitive map of knowledge diagnosis and formulation of a question in a verbal form. In addition, there is a possibility to traverse hyperlinks in the verbal part



**Fig. 2.** Representation of the learning situation for a graduate student in the frame as a combined CMKD, particular CMKD and its simplified representation

of the explanatory text (provided for all the objects of the learning material and entities mentioned on the form) and interrupt the dialogue.

Interactive interaction implies not only the display of extended information about the objects on the map (when hovering over it) but also the movement in the factor space of the learning situation characteristics. This happens due to the possibility to select different display notations (dotted arrows in the upper sequence of blocks shown in Fig. 1; for example, switching from the structural representation in the CMKD notation to the functional one in Unified Graphic Visualization of Activity notation, UGVA); changing the coverage of the displayed entities in the current scale (for example, adding the elements having links of the current course with other ones within the curriculum); switching the display aspect (for example, switching from the subject aspect to the target or competency one); and also changing the scale of the map (for example, transition from the course level to the curriculum level; such transition is indicated by gray circles in Fig. 1). Then the explanatory text corresponding to the request is synthesized and the CMKD is redrawn (see the diagrams in [20] for details).

The dialogue based on the verbal form is developed through the construction of a query (a direct close-ended question of the clarifying type [4]). This process is accompanied with the corresponding components in the drop-down block of the dialogue form (a query builder). The sequence of the element-by-element initialization of the query is shown below:

1. selection of the question word type: the item corresponding to the type of the question which a learner wants to address to the ITS (for example, “Why?”) is selected from the list;
2. selection of the subject of the dialogue (the subject): when selecting the type of the entity from the list, the corresponding dictionaries with meanings are loaded (for example, “Fragment of the learning material”) and a learner is invited to select one of them or all of them at once;
3. selection of the nature of the question relevant to the type and subject (the main verb): the process that requires clarification, justification or explanation is selected from the list (e.g. “Assessed”);

4. selection of one or more (if possible) additional parameters clarifying the query (optional): the type of the parameter and the value of interest are selected from the list (e.g. “By parameter” is selected from the list and the value of “Competency mastery level” is specified for dictionary item).

The minimum configuration of the query (if it is acceptable) is selection of the first two items. When switching to a new dialogue form, the map that is most relevant to the subject of discussion is also updated. Obviously, it is more convenient to formulate the question in an arbitrary form (without the query builder), but the entry method is not in the scope of interest of this study.

The structure of the explanatory text synthesized during the development of a dialogue includes various blocks and an element of cognitive visualization, which are combined into a template. For example, the explanatory text concerning the method of calculating the grades includes the following: a phrase of qualitative assessment of the selected parameter accompanied with the obtained quantitative values and sources of information  $\theta$  (the significance of factors is evaluated using an integrated expert system based on the FLM\_Builder module); a phrase reflecting the method used to assess the measure of confidence of the ITS intelligent solver  $\vartheta$ ; the sentence that invites the learner for further explanatory dialogue  $\delta$ .

The block structure described above varies for different configurations of queries based on pre-designed templates. All of them are supplemented with elements of cognitive visualization in the CMKD or UGVA notation [18].

## 4 Case Study, Results and Conclusion

Suppose that there is a result of interaction of the learner (the graduate student) with the e-course “Simulation Modeling” (*d3.2*) (the course is taught in the frame of the master’s degree program “Informatics and Computer Science” in the Siberian Federal University in the third semester, enrollment in 2020). The student is working with an experimental ITS AESU<sup>1</sup>. The student has previously completed a questionnaire indicating his preferences for goals  $\psi_6$ , learning units ( $u_1, u_6, u_7, u_9, u_{10}, u_{11}, u_{12}, u_{18}$  and  $u_{19}$ ), competencies to be developed (*PK-7, OK-3* and *UK-1*), etc. As a result of passing the regular (mid-term) test a dialog is displayed that contains an explanatory text  $\gamma$  with an overall grade, a description of the problem points and recommendations for repeating the particular learning material (see Fig. 3 in article [19] from this collection of conference materials). This will be the dialogue entry point.

For the current learning situation the parametric model can be visualized in the form of a CMKD shown in Fig. 2a, demonstrating the problem entities and semantic links between them both within the course and beyond the scope of the course (within the entire curriculum). The most significant emphasis in the recommendations is made on the learning topic  $t_7$  (learning units  $u_{11}$  and  $u_{12}$ ),

---

<sup>1</sup> <https://aesfu.ru/>.

which can be highlighted in the structural parametric map reflecting the competency aspect (see Fig. 2b and its simplified version in Fig. 2c for textitUK-1 displayed as an explanatory element in Fig. 3 in article [19]). For these, the parametric model of the map will be defined by the configuration  $P = < 2, 2, 2(uk1), 1 >$ . The map shows that the expert system synthesizing the explanatory text of recommendations suggests repeating the  $u_7$  and  $u_{10}$ .

Having received the ITS recommendation, the student requests an explanation of the decisions using the query builder (the query block is collapsed on the bottom of Fig. 3 in article [19]). To do this, he enters data into the query template as shown in Fig. 4 in article [19]. It corresponds to the text:

Why is the learning unit “Monte Carlo Method” ( $u_{10}$ ) assessed as insufficiently mastered by the parameter “Subject Assessment”?

Next, the user initiates the execution of the query and receives a new dialogue form, thus developing the dialogue in the direction of the block “Explanation of the decision” shown in Fig. 1. An explanatory text will be synthesized, relying on both the combined parametric map and data from the ITS knowledge base. Its structure will include a block of qualitative assessment of the student’s actions, clarification and an invitation to further dialogue (only  $\theta$  and  $\vartheta$ ):

The level of mastery of the “Monte Carlo Method” ( $u_{10}$ ) in the subject context is assessed through three assignments (assignment #2 and tests #15 and #17). The grade is 47% per 100%. This grade level is interpreted qualitatively by the ITS as insufficiently mastered (the fuzzy logic method is applied, the confidence factor in the answer was equal to 0.86).

The explanatory didactic dialogue as a means of increasing the learner’s trust in ITS decisions can be assessed by both the fact of following the system’s recommendations accompanied with their reasoning, and by the degree to which the learner is engaged in communication with the intelligent planner. The student whose data were used to illustrate this approach developed the dialogue in 38% of cases (114 entries in total) while studying the discipline “Simulation Modeling” (in 68% of cases when the dialogue was developed the student interacted with the CMKD). The student followed the ITS recommendations in 78.2% of cases when the dialogue was developed (when the dialogue did not continue, the student followed the recommendations only in 44.5% of the cases). A summary of the results of the experiment with student groups in the Siberian Federal University is given in [20]. Despite the results obtained, we consider them preliminary and there is a need of a larger pedagogical experiment.

The explanatory didactic dialogue between a learner and the ITS gives the possibility to change the objects of discussion (cross-cutting approach [18]) and aims to increase the level of trust in the system’s decisions. Our research has shown that such an effect is in general observed and it is consistent with the trends that exist in the field of the explanation synthesis [8]. The possibility of combining verbal and non-verbal types of information within a single dialogue form (using the CMKD) plays a particularly positive role. The current areas of our work are collection and generalization of more experimental data.

## References

1. Agarwal, A., Mishra, D.S., Kolekar, S.V.: Knowledge-based recommendation system using semantic web rules based on learning styles for MOOCs. *Cogent Eng.* **9**(1), 2022568 (2022). <https://doi.org/10.1080/23311916.2021.2022568>
2. Alepis, E., Virvou, M., Kabassi, K.: Development process of an affective bi-modal intelligent tutoring system. *Intell. Decis. Technol.* **1**(3), 117–126 (2007)
3. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020)
4. Belnap, N.D., Steel, T.B.: *The Logic of Questions and Answers*. Yale University Press, London (1976)
5. Du, H., Li, N., Ma, F., Palaoag, T.: Personalization exercise recommendation based on cognitive diagnosis. In: Proceedings of the 6th International Conference on Computer Science and Application Engineering. CSAE 2022, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3565387.3565416>
6. Gavrilova, T.A., Kokoulina, L.: Using ontology engineering to design an artificial intelligence course. In: Uskov, V.L., Howlett, R.J., Jain, L.C. (eds.) *Smart Education and e-Learning 2019*. SIST, vol. 144, pp. 201–207. Springer, Singapore (2019). [https://doi.org/10.1007/978-981-13-8260-4\\_19](https://doi.org/10.1007/978-981-13-8260-4_19)
7. Gribova, V.V., Ostrovskii, G.E.: An intelligent cloud service for training diagnostics skills. *Biomed. Eng.* **51**(6), 416–421 (2018). <https://doi.org/10.1007/s10527-018-9761-8>
8. Ilves, K., Leinonen, J., Hellas, A.: Supporting self-regulated learning with visualizations in online learning environments. In: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, pp. 257–262. Association for Computing Machinery, New York, USA (2018). <https://doi.org/10.1145/3159450.3159509>
9. Karpenko, A., Dobryakov, A.: Model for automated training systems. overview, science and education. *Sci. Educ.* **7**, 1–63 (2011). <http://technomag.edu.ru/file/out/504730>, (in Russian)
10. Kumar, A.N.: The effectiveness of visualization for learning expression evaluation: a reproducibility study. In: Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education, pp. 192–197. ITiCSE 2016, Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2899415.2899427>
11. Mashbitz, E., Andrievskays, V., Komissarova, E.: Dialog in a Tutoring System. Higher school, Kiev (1989). (in Russian)
12. Pospelov, D., Osipov, G.: Applied semiotics. *News Artif. Intell.* **1**, 9–35 (1999). (in Russian)
13. Song, Y.Y., Ying, L.: Decision tree methods: applications for classification and prediction. *Shanghai Arch. psychiatry* **27**(2), 130 (2015). <https://doi.org/10.11919/j.issn.1002-0829.215044>
14. Stefanuk, V.L.: Learning levels in intelligent tutoring systems. *IEICE Trans. Inf. Syst.* **78**(9), 1103–1107 (1995)
15. Sychev, O., Penskoy, N., Anikin, A., Denisov, M., Prokudin, A.: Improving comprehension: intelligent tutoring system explaining the domain rules when students break them. *Educ. Sci.* **11**(11), 719 (2021). <https://doi.org/10.3390/educsci11110719>

16. Troussas, C., Krouská, A.: Path-based recommender system for learning activities using knowledge graphs. *Information* **14**(1), 9 (2023)
17. Ugly, V., Zakharin, K., Baryshev, R.: Cognitive maps of knowledge diagnosis as an element of a digital educational footprint and a copyright object. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds.) CoMeSySo 2020. AISC, vol. 1295, pp. 349–357. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-63319-6\\_31](https://doi.org/10.1007/978-3-030-63319-6_31)
18. Uglev, V., Gavrilova, T.: Cross-cutting visual support of decision making for forming personalized learning spaces. In: Krouská, A., Troussas, C., Caro, J. (eds.) Novel and Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022). NiDS 2022. Lecture Notes in Networks and Systems, vol. 556, pp. 3–12. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-17601-2\\_1](https://doi.org/10.1007/978-3-031-17601-2_1)
19. Uglev, V., Sychev, O.: Synthesizing didactic explanatory texts in intelligent tutoring system based on the information in cognitive maps. In: Frasson, C., et al. (eds.) ITS 2023. LNCS, vol. 13891, pp. 233–246. Springer, Cham (2023)
20. Uglev, V., Sychev, O., Gavrilova, T.: Cross-cutting support of making and explaining decisions in intelligent tutoring systems using cognitive maps of knowledge diagnosis. In: Crossley, S., Popescu, E. (eds.) ITS 2022. Lecture Notes in Computer Science, vol. 13284, pp. 51–64. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_5](https://doi.org/10.1007/978-3-031-09680-8_5)



# Estimation of Piloting Attention Level Based on the Correlation of Pupil Dilation and EEG

Maryam Ghaderi, Hamdi Ben Abdessalem<sup>(✉)</sup>, Maxime Antoine, and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal, H3C 3J7  
Montréal, Canada

{maryam.ghaderi,hamdi.ben.abdessalem,  
maxime.antoine}@umontreal.ca, frasson@iro.umontreal.ca

**Abstract.** An aircrew’s behaviour greatly affects aviation safety and efficiency, especially in emergency situations. Considering vision is the primary method by which pilots acquire information, the way in which they distribute their attention greatly influences the amount of information they acquire. The purpose of this study was to examine pilots’ attention levels during a simulated takeoff procedure. In our study, we collected time series data on 136 takeoffs from 13 pilots over nine hours. In this study, we aimed to identify if pupil dilation measured by eye tracking correlates with engagement measured by EEG, especially during critical periods such as engine failures. According to the results, pupil dilation is correlated with EEG activity, suggesting that changes in pupil size indicate changes in engagement. Based on these findings, pupil dilation and EEG can be used together to examine pilot behaviour more comprehensively, since both measures are valid measures of engagement and cognitive workload.

**Keywords:** Attention · Engagement · Pupil dilation · EEG · Workload

## 1 Introduction

Aircrew member behaviour is crucial for aviation safety and efficient operation, especially during emergencies. It is a critical element for the stable development of the civil aviation industry [1, 2]. In the takeoff phase, pilots are required to perform a series of complex operation procedures and monitor the external environment and specific flight instruments (e.g., altimeter, attitude indicator, engine parameters and speed) and organize it into a cohesive picture for flight management [3]. As a result of this extremely complex task, pilots would be subjected to a great deal of load on the cognitive system [2, 4]. A pilot’s visual-motor skills and capability are impacted by cognitive load, which affects aircraft control accuracy [5]. It is evident that high cognitive load can lead to illusions and misjudgments, so correcting parameter deviations in a timely manner ensures an optimal level of safety [6].

Human activities are affected by attention, a complex cognitive process that limits the ability to focus on and understand everything at the same time [7]. Basically, attention is the ability to focus on and avoid distractions while maintaining an interest in a particular

task or idea [8]. In visual information processing activities, the study of human vision is the source of the visual attention mechanism [9]. Pilots' attention distribution mode greatly impacts the extent to which they acquire information because vision is the most significant channel for information acquisition [2, 10]. An aircraft's position, motion, and attitude can be accurately and reliably determined by a pilot's reasonable visual scanning strategy. This information can be used to respond to and prevent spatial disorientation [10]. Conversely, improper allocation of attention resources can result in overlooking or forgetting certain parameters, leading to serious risks to aircraft safety [11, 12].

A reduction in attention to special tasks can have terrible consequences, which makes this especially important. It is advantageous to direct the brain's attention toward the relevant stimulus, location, or action when it is engaged in attention-demanding tasks. Thus, a hypothesis that this research attempts to answer is whether pupil dilation follows the same pattern as EEG engagement or not. The main goal of this study is to identify if pupil dilation measured by eye tracking correlates with engagement measured by EEG, especially during critical periods such as engine failures and how these findings can be used to train novice pilots.

The rest of this article is organized as follows: in Sect. 2 several past articles that studied attention in aviation are summarized. In Sect. 3, we explain the concepts of engagement and attention. In Sect. 4, we describe our methodology, including participants, materials, and procedures. Section 5 presents the data analysis, discussion, and results of this study.

## 2 Related Work

Recently, researchers are widely using eye movement technologies in their studies about pilots' attention. They investigated whether the regularity of attention distribution can be objectively reflected by eye movement indicators. Results of [13] show that observing human gaze behaviour can reveal a lot about attention distribution. Another scholar [14], monitored pilots' attention during different flight phases using eye movement indexes. The authors found that high workload resulted in smaller sweep amplitude, longer fixation duration, and more pupil dilation. Moreover, Authors in [15] studied pilots' eye fixation during turning maneuvers near intersections by analyzing fixation rate, duration, and dwell time percentage in different AOI [15]. As noted in [16, 17], pilots face challenges in managing their attention during high and low workload situations, which can negatively affect performance and safety.

## 3 Attention and Engagement

Attention and engagement are related concepts in human behavior and cognition, where attention refers to an individual's ability to focus on a specific task, information, or object while avoiding distractions. It involves selectively filtering out distractions and directing focus towards a target [18]. Aviation researchers have studied attention and engagement for decades to prevent repetitive aircraft accidents. Engagement refers to being fully absorbed in a task, with characteristics such as motivation, focus, and enjoyment. It is

more than just enjoyment and can involve withdrawing attention as a way of engaging [19].

Engagement and attention are often interdependent. When people are highly engaged in an activity, they tend to pay close attention, which helps them maintain engagement. Similarly, paying close attention to something can increase engagement and result in a positive experience. These two concepts are interconnected and influence our behaviors and experiences [18, 19].

Various sensors can be used to track participants' electro-dermal activity, heart rate, eye movements, brain activity, facial expressions, or posture to infer states such as attention, engagement, boredom, flow, frustration, etc. [15].

Electroencephalogram (EEG) signals can be used to extract two mental indices: mental engagement and cognitive load. Several fields, including aviation, robotics, and the military, use these states to dynamically assess changes in the user's state, since they are closely related to the user's performance and experience [20, 21]. The engagement index can be calculated directly from EEG signal frequency bands [22]. This is based on three EEG frequency bands, which are: (4–8 Hz), (8–13 Hz) and 13–22 Hz:

$$\text{Engagement} = \beta / (\theta + \alpha) \quad (1)$$

Emotiv headsets measure 14 regions simultaneously, so you can sum the values of all the frequency bands over all the measured regions to get a combined value of the  $\theta$ ,  $\alpha$  and  $\beta$  frequency bands. From the EEG signal, an engagement index is calculated every second [22, 23].

## 4 Experiments

An experiment was conducted aiming to gather pilot CW and PD during a takeoff procedure in an Airbus A320 in real time. It consisted of six different scenarios with varying weather, time, and conditions as well as the possibility of a failure. The first to third scenarios relate to standard takeoff sessions, while scenarios four to six relate to failure sessions. Different scenarios were given to participants to perform the takeoff procedure. In a session, no limits were placed on the number of takeoffs that a participant could take. A pilot monitor and a pilot are required to operate the Airbus A320. In this experiment, the participants served as pilots and the experimenter as pilot monitors. The experiment was conducted at the University of Montreal and with participants from CAE and Bombardier who all signed informed written consent forms.

### 4.1 Participants

Participants from CAE and Bombardier participated in the experiment at the University of Montreal. The study included 13 male subjects (all between 24 and 49 years of age) with an average age of 36 years. It was required that participants work in an industry related to aviation. The 13 participants included 7 pilots with piloting experience, piloting license and A320 piloting experience. The other 6 participants were engineers at Bombardier and CAE who were familiar with most aircraft procedures but without holding a piloting license. For these participants, the minimum and maximum flight hours were 1 h and 3000 h, respectively.

## 4.2 Materials

**Eye-Tracking Module.** Gaze point is the most affordable, research-grade eye tracker on the market, used in the experiment to measure pupil dilation.

Eye-tracking is a technique to investigate visual attention, where eye movements such as fixations (stationary phase) and saccades (rapid phase) are measured along with gaze point location to study visual attention [24, 25].

**Pupil Dilation.** Pupil dilation is a physiological response that can provide information about a person's level of attention and engagement.

When a person is paying attention to a visual stimulus, the dilator muscle contracts, causing the pupil to enlarge. This increase in pupil size is known as pupil dilation. Studies have shown that pupil dilation can be used as a measure of attention and engagement in a variety of contexts [26].

**Electroencephalography (EEG).** The electroencephalography technique measures and records electrical activity in the brain. An EEG measures electrical signals generated by the brain by placing electrodes on the scalp, which are amplified and recorded. In the traditional EEG analysis method, EEG signals (the brain activity) are divided into the different five wavebands Based on the frequency range, including:

- **Delta (1–4 Hz),** or  $\delta$  activity is a periodic electromagnetic wave that has a frequency ranging from 1 to 3 Hz and an amplitude between 100 and 200 V.
- **Theta (4–8 Hz),** or  $\theta$  activity is a periodic electromagnetic wave that has a frequency ranging from 4 to 7 Hz and an amplitude of under 30  $\mu$ V. When people are under stress or during a period of profound relaxation, these waves are generated.
- **Alpha (8–13 Hz),** or  $\alpha$  activity is a type of periodic electromagnetic wave that has a frequency ranging from 8 to 13 Hz and an amplitude between 30 and 50 V. Whenever a person is conscious, relaxed, or at rest, the brain emits waves in the parietal region and the occipital region.
- **Beta (13–30 Hz),**  $\beta$  activity wave is a periodic electromagnetic wave with a frequency between 14 and 30 Hz and an amplitude ranging from 5 to 20 V. During times of thought or sensory stimulation, these waves are especially noticeable.
- **Gamma (30–60 Hz),** or  $\gamma$  activity is a type of periodic electromagnetic wave with a frequency of 31–60 Hz and an amplitude of 5–10 V.  $\gamma$  activity is associated with selective attention, cognition, and perceptual activity.

EEG can be used to calculate attention and engagement levels by measuring specific frequency bands such as alpha and beta. Higher levels of beta and lower levels of alpha indicate increased attention, while higher levels of alpha and lower levels of beta indicate decreased attention [27]. It should be noted that the workload is directly gained from the EEG Module. By connecting to the EEG headset, the module is able to measure the cognitive workload of the pilot in real-time. This module is actually part of the NCO software, which is owned by the BMU lab and is specifically designed for this purpose [28].

### 4.3 Procedure

The experiment environment was comprised of a participant (for each experiment) as the pilot and the experimenter as the pilot monitor. Ethics committee and partners' approval were obtained for the experiment's procedure. Participants received a detailed description of the A320 takeoff procedure a week before their experiment to familiarize themselves with the handling characteristics of the joystick, side stick, and pedals used to control the aircraft during takeoff. The experiment included various scenarios with and without failure, such as a rejected takeoff, an engine failure after V1 procedure, and a standard takeoff procedure. The experiment also featured several visual displays, including the navigation display, primary flight display, engine/warning display, system display, and engine display. In order to generate a more cognitive workload for every scenario, participants were not aware of the scenarios beforehand. The experiment was found to be easier for participants with piloting experience compared to those without.

## 5 Results and Discussion

One of the hypotheses that this research tries to answer is: "Is it possible to measure pilot attention during a takeoff flight?".

Here we show the data output examples of two participants. The first one (Fig. 1) shows the data output of an official A320 pilot with eight years of experience and more than 250 official flight hours during a rejected takeoff. The other figure (Fig. 4) is related to output of a non-licensed pilot for the same scenario. This pilot is an A320 engineer who knows the ins and outs of the A320 and has flown on the simulator with it for testing purposes but is not a licensed pilot. The figures include:

1. The Engagement, Pupil Dilation, and Cognitive Workload of the pilot at a time t
2. Important events, which resulted in a Cognitive Workload increase.

The three plots are aligned vertically based on time. Time 05<sup>th</sup> s is the start of this scenario when the parking brake is released (first vertical line) and 50<sup>th</sup> s is almost the end of scenario, ECAM action and when the pilot confirms the engine must be turned off (last two vertical lines).

As pupil dilation directly relates to someone's attention and is often used as a physiological measure of cognitive load, attention, and engagement, our main focus will be on the dilation directly and its relation with cognitive workload. As we can see in the Fig. 1, from time zero to 05<sup>th</sup> s, which is the start of this scenario when the parking brake is released (first vertical line), the workload is low due to the fact that the flight is a regular operation and does not require excessive resources. The pupil dilation is high from the beginning until 15<sup>th</sup> s, because the plane is about to take off and the participant wants to maintain it in the middle of the runway. There is a medium to high level of engagement during this period of time. This is the first reaction of pilots and shows that attention is triggered before the workload. The workload increased noticeable after a short delay because the brain energy was mobilized. What we see before reaching 15<sup>th</sup> s, is the peak of workload and we see that the engagement and pupil dilation has been reduced a little and then maintained at a medium to high level.

Although there is some peak in workload between 15<sup>th</sup> and 33<sup>th</sup> s, in general the workload is low. At the time of peaks in workload (which are mainly a search), we see a decrease in engagement and pupil dilation (20<sup>th</sup> s after the plane leaves the ground), but shortly thereafter, both engagement and pupil dilation increase as the pilot realizes its normal occurrence. Which means that the pilot has identified the problem, knows what to do and has applied the right task (The pilot is in full control of the problem).

Again, at 30<sup>th</sup>, there is a search and the workload increase again in parallel until 34<sup>th</sup> s. There is an engine failure at 34<sup>th</sup> s. Thus, after the pilot discovered the engine failure at 33<sup>th</sup> and 34<sup>th</sup> s (when it was announced by the copilot), there was a considerable workload at first, but it decreased rapidly at 35<sup>th</sup>, which indicates that the pilot was not surprised, knows the problem and has applied the right task. The pupil dilation changed quickly because of the urgency of the situation, and he knew he was going to need to be very careful. At this time, engagement also increased somewhat, but not too much because as we mentioned, he was a master pilot and had flight under control even during a failure. Again, at 50<sup>th</sup> s we see a considerable increase of workload which is almost the end of scenario(takeoff), ECAM action and when the pilot confirms the engine must be turned off (last vertical line). The workload increased because it is the end of takeoff and pilot had to think about the next action and to see what to do. Then, as we progressed from 50<sup>th</sup> to the end of the flight, after making the correct action, when arriving at 52<sup>th</sup> s, the workload decreased and was low in general since he has done the correct action.

In this period, we see that the engagement was medium which means that in pilots mind, he has completed the takeoff and now he is entering to navigate. However, the pupil dilation was high due to his concern for maintaining the plane's course.

We should take into account the fact that participants know that a failure may occur during takeoff. The PD and CW curves show that despite the fact that the workload decreased and increased at different points (each CW increase was due to a specific action the pilot needed to perform), the pilot continued to be engaged even when a failure occurred, which proves that the pilot has identified the problem and knows what to do about it.

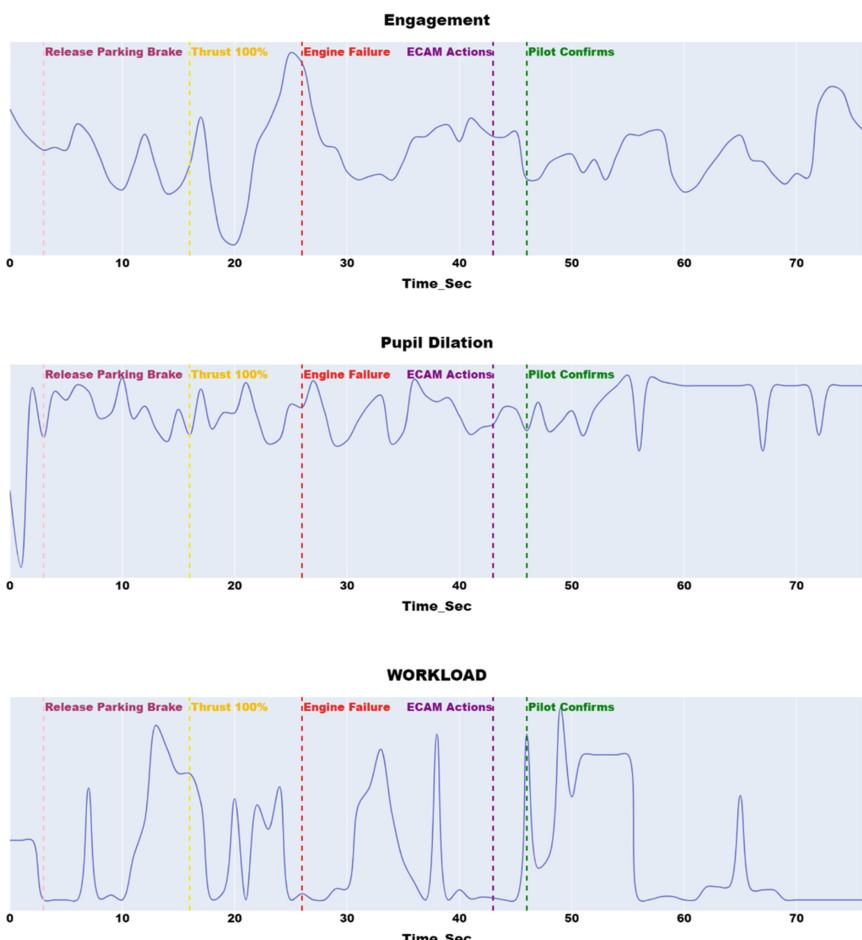
On the other hand, for the other participant (pilot 5) who was a non-licensed pilot (Fig. 2), everything was completely different. At the beginning of the flight, he has a high workload since it is not a normal operation and will require a lot of energy consumption. He is unfamiliar with the procedure, and this is why the engagement and pupil dilation are low (decreased). In general, even though there are two peaks in workload, we have seen a decrease in workload until the 10<sup>th</sup> s. This is while both engagement and pupil dilation decreased and maintained a medium level after a sharp increase at around 4<sup>th</sup> s. Which is not normal because the pilot has decreased the workload and not fully mobilized on the task and the workload is decreased while the attention remains medium to high.

From the 10<sup>th</sup> s to the 26<sup>th</sup>-s pilot realized that he has to face over steps of tasks that is why workload increased. At this point, the workload is maximum and in parallel, the pilot's Engagement and pupil dilation increased slowly. This means that the capability of pilot is fully mobilized. The engine failure was announced in Sec 26. Pilots, however, did not realize this before announcing it as we see that the workload increased following the announcement of engine failure. Engagement and pupil dilation increased at first,

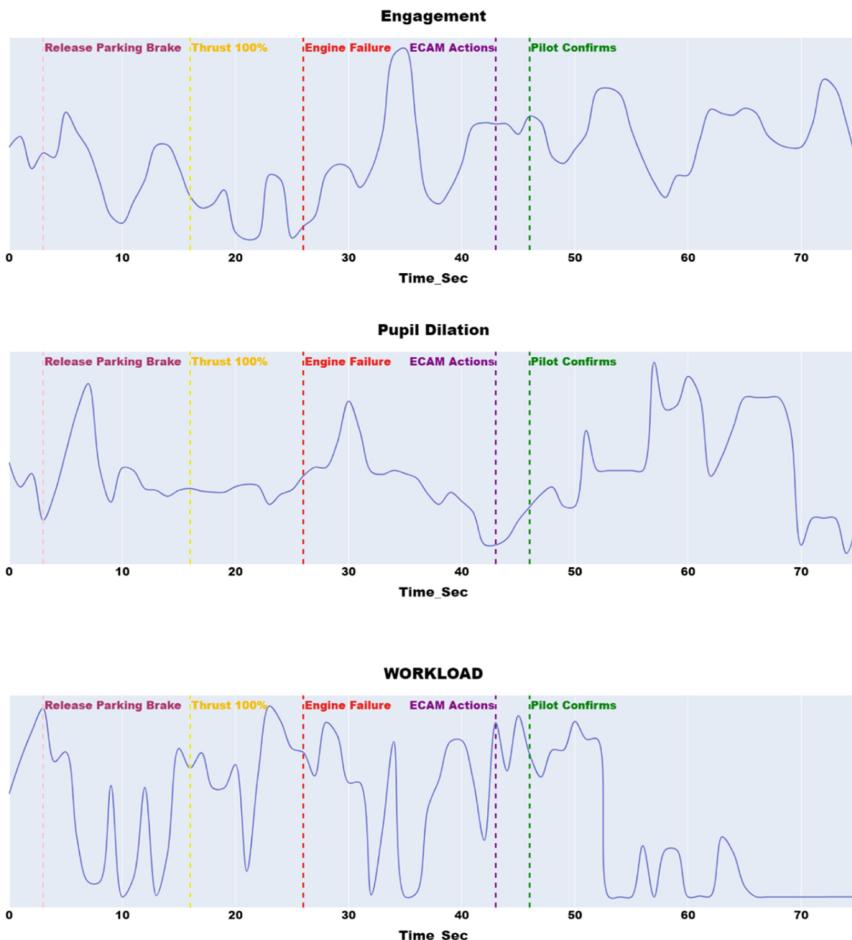
but not much (just a small amount) and then decreased, showing he wasn't engaged and unsure of what to do.

From 35<sup>th</sup> s to 43<sup>th</sup> s, the workload increased while both engagement and pupil dilation decreased which is not good and he should be fully attended, because he did not know what to do and looking for what to do. After that, we see a sharp increase in workload ((very high workload)) during the ECAM action period, which is between 43<sup>th</sup> and 50<sup>th</sup> s, whereas engagement and pupil dilation are low or medium. Which is again not good and normal.

The graph shows that most of the time when the workload is high, engagement is low or average. When the workload is low, engagement is high, but that's no sign of flight control. When the engagement is increase while the workload is low, means that the pilot has not enough resources and look around to see what to do.



**Fig. 1.** Data output of an official A320 pilot with eight years of experience and more than 250 official flight hours.



**Fig. 2.** The data output of a non-licensed pilot (an A320 engineer).

## 6 Conclusion

In conclusion, this study investigated the relationship between pupil dilation, engagement measured by EEG, and workload in pilots. Pupil dilation is a commonly used physiological measure of cognitive workload and attention, and EEG measures the electrical activity of the brain. The results showed a strong correlation between pupil dilation and engagement measured by EEG, indicating that these two physiological measures are related and follow the same pattern. A positive correlation was also found between pupil dilation and engagement with the workload as measured by EEG, further emphasizing pupil dilation's usefulness as a cognitive process indicator. Combining both of these physiological measures has been shown to be useful for understanding pilots' engagement levels and cognitive workload.

The study's results have significant implications for aviation safety and pilot training programs. By identifying areas of weakness in pilots' performance, targeted training programs can be developed to enhance their skills and improve their overall performance. This approach can help reduce the number of accidents caused by human error and ensure the safety of passengers and crew. Additionally, the use of physiological measures such as pupil dilation and EEG to understand engagement and workload in pilots can provide valuable insights into how pilots react to various flight phases and situations. This information can be used to optimize pilot training programs and improve aviation safety and efficiency. Overall, the study's findings have the potential to make a significant contribution to the field of aviation research and improve air travel safety.

**Acknowledgments.** We acknowledge the support from NSERC-Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

## References

1. Harris, D.: The influence of human factors on operational efficiency. *Aircr. Eng. Aerosp. Technol.* **78**(1), 20–25 (2006). <https://doi.org/10.1108/17488840610639645>
2. Jin, H., et al.: Study on how expert and novice pilots can distribute their visual attention to improve flight performance. *IEEE Access* **9**, 44757–44769 (2021). <https://doi.org/10.1109/ACCESS.2021.3066880>
3. Peißl, S., Wickens, C.D., Baruah, R.: Eye-tracking measures in aviation: a selective literature review. *Int. J. Aerosp. Psychol.* **28**(3–4), 98–112 (2018). <https://doi.org/10.1080/24721840.2018.1514978>
4. Lounis, C., Peysakhovich, V., Causse, M.: Visual scanning strategies in the cockpit are modulated by pilots' expertise: a flight simulator study. *PLOS ONE* **16**(2), e0247061 (2021). <https://doi.org/10.1371/journal.pone.0247061>
5. Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., Wickens, C.D.: Pilot flying and pilot Monitoring's aircraft state awareness during go-around execution in aviation: a behavioral and eye tracking study. *Int. J. Aerosp. Psychol.* **27**(1–2), 15–28 (2017). <https://doi.org/10.1080/10508414.2017.1366269>
6. Wei, H., Zhuang, D., Wanyan, X., Wang, Q.: An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chin. J. Aeronaut.* **26**(4), 884–889 (2013). <https://doi.org/10.1016/j.cja.2013.04.053>
7. [www.skybrary.aero](http://www.skybrary.aero), Attention and vigilance
8. Wickens, C.D., McCarley, J.S.: *Attention-Situation Awareness (A-SA) Model of Pilot Error*. CRC Press (2007)
9. Yang, B., Li, H.: A visual attention model based on eye tracking in 3D scene maps. *ISPRS Int. J. Geo-Inf.* **10**, 664 (2021). <https://doi.org/10.3390/ijgi10100664>
10. Balaj, B., et al.: Spatial disorientation cue effects on gaze behaviour in pilots and non-pilots. *Cogn. Technol. Work* **21**(3), 473–486 (2018). <https://doi.org/10.1007/s10111-018-0534-7>
11. Chen, H.: 'Error analysis of student pilots in the approach phase. *Sci. Technol. Inf.* **15**, 218 and 220, 201. <https://doi.org/10.16661/j.cnki.1672-3791.2012.15.175> (in Chinese)
12. Li, G., Lai, W., Sui, X., Li, X., Qu, X., Zhang, T., Li, Y.: Influence of traffic congestion on driver behavior in post-congestion driving. *Acc. Anal. Prev.* **141**, 105508 (2020). <https://doi.org/10.1016/j.aap.2020.105508>
13. Wanyan, X., Zhuang, D., Wei, H., Song, J.: Pilot attention allocation model based on fuzzy theory. *Comput. Math. Appl.* **62**(7), 2727–2735 (2011)

14. Liu, Z.Q., Yuan, X.G., Liu, W., Kang, W.Y., Han, Y.D., Ma, R.: Analysis on eye movement indices based on simulated flight task. *China Saf. Sci. J.* **16**(2), 51–55 and 149 (2006)
15. Zhang, X., Li, G., Xue, H., Zhao, H.: Pilots' scanning behavior between different airport intersection maneuvers in a simulated taxiing task. *IEEE Access* **7**, 150395–150402 (2019). <https://doi.org/10.1109/ACCESS.2019.2947530>
16. Ghaderi, M., Ben Abdessalem, H., Frasson, C.: An analysis of mental workload involved in piloting tasks. In: Krouská, A., Troussas, C., Caro, J. (eds.) *Novel & Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022)*. NiDS 2022. Lecture Notes in Networks and Systems, vol. 556. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-17601-2\\_21](https://doi.org/10.1007/978-3-031-17601-2_21)
17. Ghaderi, M., Courtemanche, M.A., Ben Abdessalem, H., Nkambou, R., Frasson, C.: Attentional tasks model: a focus group approach. In: Krouská, A., Troussas, C., Caro, J. (eds.) *Novel & Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022)*. NiDS 2022. Lecture Notes in Networks and Systems, vol. 556. Springer, Cham (2023) [https://doi.org/10.1007/978-3-031-17601-2\\_29](https://doi.org/10.1007/978-3-031-17601-2_29)
18. Charles, H.M., Leung, Y.K.: Pilot mental workload: how well do pilots really perform? *Ergonomics* **49**(15), 1581–1596 (2006)
19. Battiste, V., Bortolussi, M.: Transport pilot workload: a comparison of two subjective techniques. *Proc. Hum. Factors Soc. Annu. Meet.* **32**(2), 150–154 (1988). <https://doi.org/10.1177/154193128803200232>
20. Li, Y., Li, X., Ratcliffe, M., Liu, L., Qi, Y., Liu, Q.: A real-time EEG-based BCI system for attention recognition in ubiquitous environment. In: *Proceedings of 2011 International Workshop on Ubiquitous Affective Awareness and Intelligent Interaction (UAAII '11)*. ACM, New York, NY, USA, pp. 33–40 (2011). <https://doi.org/10.1145/2030092.2030099>
21. Liu, N.-H., Chiang, C.-Y., Chu, H.-C.: Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors* **13**(8), 10273–10286 (2013). <https://doi.org/10.3390/s130810273>
22. Kosmyna, N., Morris, C., Sarawgi, U., Maes, P.: AttentivU: a biofeedback system for real-time monitoring and improvement of engagement. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*, Article VS07, 2 p. ACM, New York, NY, USA (2019). <https://doi.org/10.1145/3290607.3311768>
23. Carreiras, C., Lourenço, A., Aidós, H., da Silva, H.P., Fred, A.L.N.: Unsupervised analysis of morphological ECG features for attention detection. In: Madani, K., Dourado, A., Rosa, A., Filipe, J., Kacprzyk, J. (eds.) *Computational Intelligence. SCI*, vol. 613, pp. 437–453. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-23392-5\\_24](https://doi.org/10.1007/978-3-319-23392-5_24)
24. Abdelrahman, Y., et al.: Classifying attention types with thermal imaging and eye tracking. *Proc. ACM on Interact., Mob., Wearable Ubiquitous Technol.* **3**(3), 1–27 (2019). <https://doi.org/10.1145/3351227>
25. Mancas, M., Ferrera, V.P.P., Riche, N., Taylor, J.G.G. (eds.): *SSCNS*, vol. 10. Springer, New York (2016). <https://doi.org/10.1007/978-1-4939-3435-5>
26. Jacob, R.J., Nessler, D.: Pupil dilation reveals sympathetic activation during emotional picture viewing. *Sci. Rep.* **7**(1), 7986 (2017)
27. Liu, Y., Perlmuter, S.I.: The functional significance of alpha and beta oscillations in the human brain. *Neuropharmacology* **43**(1), 4–17 (2018)
28. Benlamine, M. S., Chaouachi, M., Frasson, C., Dufresne, A.: Physiology-based recognition of facial micro-expressions using EEG and identification of the relevant sensors by emotion. In: *Proceedings of the 3rd International Conference on Physiological Computing Systems*, pp. 130–137 (2016). <https://doi.org/10.5220/0006002701300137>



# Functional Programming of Intelligent Systems

Vladymyr Meitus<sup>(✉)</sup> and Clara Simon de Blas<sup>(✉)</sup>

Rey Juan Carlos University, 28933 Mostoles, Madrid, Spain

vmeitus@gmail.com, clara.simon@urjc.es

**Abstract.** The paper considers the problem associated with the possibility of functional programming of intelligent systems, which are based on the definition of intelligence as the ability to model the environment around the system in order to use this model to form the specified behavior of the system in this environment. Such behavior is considered as the result of a consistent solution of intermediate tasks, into which the general task is divided, determined by the goal set for the system. In the variant under consideration, the environment model is built on the basis of knowledge collected by the system or obtained from its knowledge base. Separate knowledge has a multi-element representation, making available for the user several tools for solving problems. The options proposed in his paper are: sets of properties, logical and ontological representations of individual components of the environment surrounding the system, and related associations of these components. It should be noted that various variants of logics can be incorporated into the system, including non-classical ones, on which the system builds its conclusions. In addition, the system can use various variants of mathematical structures that are stored in its knowledge base when building a model.

When developing an intelligent system, the methods and tools of functional design can be applied as a way to develop a specific system. In this work, this approach is applied on the example of the development of an intelligent military robot that operates in a specific subject area and solves the problem of defending and attacking a specific enemy.

**Keywords:** intelligence · modeling · intelligent system · knowledge representation · intelligent robots

## 1 Introduction

The problem of building and using intelligent systems today is one of the most urgent, since it is in this direction that a colossal resource for the development of human society is potentially laid. Throughout its history, man has constantly – sometimes faster, sometimes more slowly – learned how the world around him works, and how it can be used. With the advent of smart intelligent systems that allow solving problems, demonstrating intelligent behavior, independently finding and describing new structures in the world around us, creating artistic, musical, literary works, proving theorems, developing intelligent robots, the world around us has changed. And, above all, because the new intelligence that mankind has been looking for for hundreds of years in the Universe,

suddenly began to appear in artificial intelligence systems created by man and entering our lives.

Although the field of artificial intelligence has been developing only in the last 70 years, numerous results have been obtained in it, which only confirm the necessity and importance of these studies. And practical application begins to reveal the numerous facets of these studies and their role in shaping the conditions for the existence of modern society [1–4].

It turns out that artificial intelligence can be successfully implemented in numerous processes (production, scientific, cultural, informational) that define modern society. Moreover, the use of intelligent systems changes the nature of these processes, significantly raising their level. Humanity has received a new resource and new opportunities.

The study of this general trend in the development of society is not the subject of this work. Therefore, only one problem is considered below: the use of functional programming methods to build systems with some intelligence – intelligent systems. As an example, the development of one of the classes of modern robots, namely, robots with a military purpose, is considered. The general approach to the construction of intelligent systems is based on the definition of intelligence given in [5–8]. The language considered to implement functional programming in this work is Python, as it is one of the most widely used programming languages today and its use continues to grow, due to its simple syntax and that is open source.

The approach to the development of such systems assumes that a system can be considered as intelligent if its behavior is determined by properties that are defined as a manifestation of intelligence. And the stronger these properties are manifested, the higher the level of intelligence that the system possesses.

But even in the simplest cases, the intellectuality of actions is a necessary condition for the existence, development and use of any field of knowledge, both scientific and technical – industrial, technical, household, research, cultural and educational – for the development and application of systems that use intelligence as the basis of their behavior.

## 2 Intelligent systems

**Intelligence**, considered as a property of a subject or system, is determined by the ability using some logic, firstly, to model the subject area (SbA) in which this subject (system) operates, and, secondly, on the basis of this model, to effectively solve the tasks assigned to this subject (system) [5, 6].

A **system** with intelligence is hereinafter referred to as an intelligent system (IS). There are many different ways of modeling SbA [5, 9, 10], as well as approaches to solving problems in these models. In [6–8] it was proposed to use various logics to describe the tasks themselves, and options for drawing conclusions from the logical description of the problem condition to its solution.

In this paper, we consider the problem of applying functional programming representations to the development of IS. In this case, it is necessary to use an approach, which is based on the original definition of intelligence, to the creation of IS. This approach

will serve as the basis for developing a programming representation. It is assumed that the system has a set of organs that allow it to perceive the environment, and on the basis of this perception and the initial knowledge existing in the system, build a model of the environment. This model contains in advance the rules and structures related to the SbA in which it is supposed to use the IS or to customize its actions for the finished system. Unlike universal IS, which allow an independent choice of the method of modeling the environment, in the case under consideration, some forms of the SbA representation are predetermined, and a specific version of the IS is selected based on the information that it can collect. In other words, a certain class of SbA is built for IS, from which the system selects a specific environment, focusing on the information that it collects or that it previously receives from auxiliary sources.

This approach narrows the scope of IS, limiting the class of SbA in which the system operates, but increases the level of its efficiency, reducing the time spent on the detailed process of modeling and checking the model's compliance with the reality in which the system is located. For example, if an IS is the "brain" of a robot, then the conditions characteristic of the environment in which the robot must operate, that is, solve problems, are pre-set in it. Also, the means that the robot can use and which are included in its structure are also associated with this environment. The intelligent robot does not create these tools, but only chooses from what it already has to solve its problems.

In this case, the modeling process is divided into two parts. The first is the creation of a knowledge base that contains a description of the possible constituent elements of the simulated SbA, at the level of individual features and logical relationships between them, as well as the impact of these elements on other elements included in the modeled area. The base includes a description of the possible structures included in the SbA that determine the links between the elements of the domain, the assignment of links between the elements and structures. A separate component of the IS knowledge base are mathematical models that can be used in modeling the SbA and solving problems in it. At the same time, it is assumed that at the level of such a model, the internal "skeleton" of the considered SbA and the connections that exist in it can be represented.

When an IS enters the SbA, it first orients itself in it, highlighting individual elements and structures based on the collected features, and then turns to its knowledge base (KB), in which it looks for matches with the elements and structures identified by it in the environment. There is a process of recognition and identification of these elements and structures, as a result of which the SbA modeling is performed, which combines the perception of the IS and information from its knowledge base.

The process of recognition of the observed components and components of the KB is usually based on the selection of a group of observed features that are compared to a separate component, the formation of a neighborhood of these features that correspond to some generalized image and the search in the KB for a prototype whose features fit into the constructed neighborhood. If there are several such elements, then an extended (multiple) representation of the SbA is constructed, which is refined in the process of interaction between the IS and the SbA. The constructed SbA model further serves as the basis for solving problems that may arise or be put before the IS.

It is assumed that when developing an IS, the range of tasks that this system should solve in the SbA is also outlined in advance. Designing a set of tasks to be solved is one of

the common problems that IS solves. It is quite acceptable that the IS performs its actions on the basis of some general principles. But this is the next level of IS construction. For example, in the stories of A. Azimov, the behavior of robots was limited to three general principles that forbade robots from performing certain actions. A simpler variant can be specified under conditions when the range of tasks to be solved is defined for the IS, and the sequence of possible actions of the IS is specified for each task. Such program descriptions are a separate part of the knowledge base, which is associated with the IS.

Note that the solution of a specific problem may be associated with a certain sequence of actions, but, at the same time, such a solution may include some general algorithm for analyzing the situation, and only then a specific program of actions associated with this situation. Moreover, the algorithm of behavior may involve the choice of some optimal sequence of actions, which is based on an analysis of the situation.

Thus, the functional programming of an IS with this approach is divided into two parts: the first is the functional programming of a set of knowledge and operating with it, which provides modeling of the SbA. And the second part is the functional programming of the operational space of the IS, determined by the possibilities of solving the problems of the IS. Let us now consider the formal representation of each part separately. In the formalization of representations in the future, we will use the concepts of the Python programming language without additional references [11, 12].

In this work, the SbA and the tasks solved in it are considered on the example of the development of an IS that controls a military robot.

## 2.1 Subject Area Modeling

As discussed above, the intellectual behavior of the subject in the SbA is determined by two main factors that form the *intellect* of the subject interacting with the SbA: firstly, by its ability to model the environment – the SbA in which it is located, and, secondly, by its ability to base on this model to solve problems related to this environment.

The work [5] considered the modeling of the SbA based on the knowledge that the subject collects and accumulates, perceiving and analyzing this environment. Currently, there are various approaches to the use of knowledge in systems related to artificial intelligence. An example is different versions of knowledge-based systems [13, 14]. In this work, firstly, knowledge serves as the basis used in the process of modeling the SbA. They build a model in which problems are solved. And, secondly, this knowledge has a rather complex structure that combines various forms – descriptive, functional, logical and ontological – in a single knowledge, which makes it possible to build an SbA model in a representation which focused on multilateral problem solving. This property is characteristic of the intellect of a person who usually evaluates the problem being solved from different angles.

The process of forming knowledge about the elements of the SbA is the first intellectual component of the general modeling process, because the subject must, on the basis of this knowledge, recognize the object and link it with the ontology representing the SbA. At the same time, for an object, its links with other objects in the given SbA are indicated.

At the same time, it is assumed that the subject in the SbA can single out its separate parts, which are considered as generalized objects of this area, and these parts are also

characterized by a certain knowledge system, including features, connections, ontology. Separate features are associated primarily with the senses of the subject, or are derived from individual tuples or sets of features. A structured list of such features is the first component of knowledge about an object. The second component of knowledge is a logical expression that describes the object in the logic used by the subject to represent objects and to draw his conclusions in the process of solving problems in the SbA.

The third component of knowledge about an object from an SbA is a set of connections that exist and are found by the subject in the course of his perception and subsequent analysis of the object. These links define some neighborhood of the object, defining its interaction with other objects of the SbA. Connections and their compositions are a necessary condition for solving problems by the subject. From them, a solution to the problem is compiled, as a path connecting the condition of the problem with the expected result.

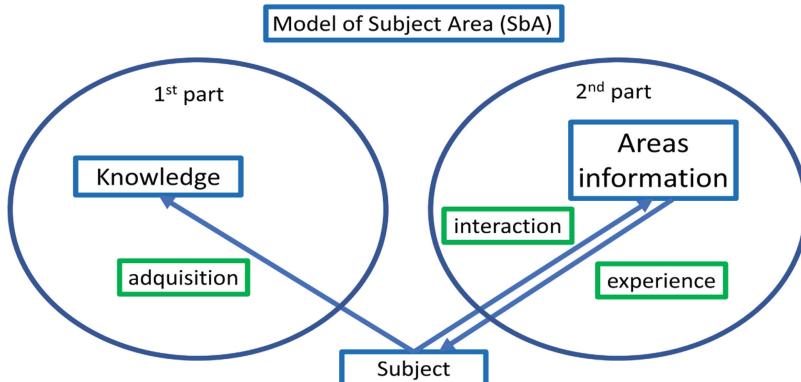
Finally, the fourth component of knowledge about an object is its ontological description, which assumes that information about individual objects of the SbA exists and is represented in some general catalog of possible objects. Such a catalog contains more detailed information that the subject may not find or obtain using only his senses. For example, a robot can observe individual elements of a tank in the SbA, which allow it to give an external characteristic of the observed object. But he will get a number of other characteristics if he can find the necessary prototype in the knowledge base and determine what type of tank he observes and what are its properties and capabilities, or means of protection. It depends on whether the robot will be able to destroy the tank with the weapons it has.

These four components of knowledge give some characteristic of the SbA object that is static in time. But this is not enough for ordinary real conditions, in which, for example, an unmanned vehicle or an aircraft, a robot or an autonomous control complex, a tank, a tractor, which are equipped with a control device and operate in a variable environment, are considered as an object, changing their position and their performance. Therefore, a pure function environment is needed, where stable, consistent, and predictable outcomes are given. Only for the case new data are given, new variable environment must be created.

In order to represent possible dynamic changes in knowledge, it is necessary to add at least one more component to them, which includes a set of operators ( $\Delta_1, \dots, \Delta_n$ ), potentially transforming the observed state of a particular object at each moment of time. In other words, operators set the dynamics of the individual components of the object, which the subject must take into account when solving the tasks assigned to him. For example, with the help of an operator, it is possible to describe the movement of a car in the form of a change in a part of the features associated with it or in the neighborhood in the SbA, which is determined by the car object. The fifth, operator component of knowledge allows a more complete representation of the SbA model, reflecting its potential dynamics and, consequently, expanding the range of tasks that an intellectual subject (ISb) can solve. This component can be enriched with the aggregation rules for the set of operators ( $\Delta_1, \dots, \Delta_n$ ).

The knowledge collected by the subject is only one part that is related to the general task of modeling the SbA. The second part consists of a body of knowledge containing information about the areas in which the subject can act and the possible components

that these areas consist of. This knowledge contains the accumulated experience that reflects the possible scenarios that a subject may encounter when entering a particular environment (see Fig. 1). For example, a person who is going to visit another country usually first gets acquainted with the rules, traditions, features of his stay in this country, which may differ from the same features of his own place of stay.



**Fig. 1.** Representation of model of subject area.

And if the ISb has to model some environment, then it is desirable that it has preliminary information about what type of environment it may encounter. For example, if a robot has to simulate a forest, then it needs one piece of information, and if it ends up in a desert, in a field, in a city, then another. Of course, one could consider the possibility of the robot independently forming its own idea of the environment. But, firstly, it would take additional time, and in conditions where the environment is not favorable for the robot, it could pose a threat to its existence. Secondly, in order to adapt to different environments in conditions when the necessary information needs to be collected, analyzed and built in different models, a sufficiently high level of intelligence and the ability to restructure one's perception are required. And this is due to a longer general period of training and adaptation of the robot to the SbA.

Therefore, if it is possible to preselect the main environment model from the knowledge base, then this should be used to ensure that the robot is quickly configured for a specific environment. Such a model includes the main elements that ISb can meet under the conditions during consideration, as well as a certain formal organization of these elements at the level of some mathematical structure [5].

And the further modeling process consists in setting up the selected main model, for example, a city, forest, steppe, jungle, desert, mountains, for specific conditions and objects in this main model, detailing this model, with the transition to solving already specific problems. Naturally, the choice of the main model also narrows the class of possible objects that the subject will put into the model based on his perception, and reduces the time required to develop the SbA model.

It is assumed that the ISb is located on our planet, but can fall into different climatic zones and at different times of the year, operate in different environments – on the

ground, in the air, under water, and has equipment for perceiving the environment and solving the tasks assigned to the subject. The main advantage of the subject (robot) is in a variety of means of perceiving the environment: the possibility of using a wide range from infrared to ultrasonic, the possibility of sensing the environment in various ways, obtaining information from flying drones and satellites. For example, a drone can be an integral part of a robot, which allows expanding its ability to obtain additional information about the SbA for each specific situation. In other words, such robots can collect information and build SbA models in more detail and perfection than a human does.

## 2.2 Functional Representation of the Modeling Process

Assuming that the robot has already collected data about its environment, transformed it into knowledge about individual objects, and it is given tasks that need to be solved based on this information, the SbA modeling process itself can be considered as the process of constructing and using a function that displays these knowledge into a structural model of the environment surrounding the robot.

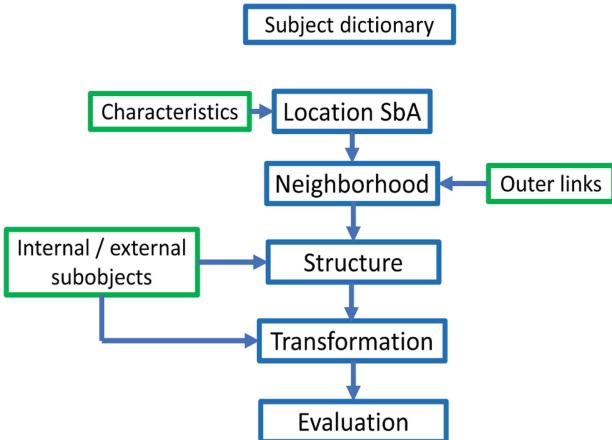
To formalize the representation of the description of the environment, we use the concepts of the Python programming language. This can be easily generalized to other functional programming languages. In this case, knowledge about the individual components of the SbA and its structures is specified in the form of lists and named dictionaries. At the same time, the dictionaries include names derived from the ontology, which is included in the knowledge and is associated with this area. For an SbA, there is a knowledge base that the robot can use to build knowledge based on its perception.

If knowledge about the object is incomplete, for example, it was not possible to fully identify the object based on the collected features, then the model for this object saves the collected information, and the process of knowledge formation is stretched in time. Based on the collected data, the robot can make several assumptions about what the analyzed object is. It may turn out that an exact classification of the object is not needed to solve the problem.

The dictionary associated with an individual object (Fig. 2) includes, firstly, information about the position of the object in the SbA and its main characteristics that determine this position. Secondly, possible external links with other, mostly neighboring objects. Thirdly, the proposed structure of the object and possible internal and external subobjects that are combined with this object of the SbA, but allow their independent existence and behavior. Fourth, the possible dynamic characteristics of internal subobjects, reflecting, for example, the movement of internal objects in the structure of the original object. Fifth, a possible assessment of the value of the object, which is given by the robot in relation to the SbA.

As a subobject, we consider some structure, which is considered as part of the original object. For example, in the course of military operations, people who take shelter in this building, or observation and firing points with which the building is equipped, or some structures with which the building is supplemented, can be considered as a subobject of a building.

The model itself is defined as a multi-layer system of maps, starting with maps that indicate the relative position of objects with varying degrees of their detail in the



**Fig. 2.** Subject dictionary components

representation, and ending with maps that show the possible dynamics of SbA objects reflected in the robot's information. Note that the map can include both a graphical representation of the environment and sets of programs corresponding to the actions that the robot can perform in a particular environment. Finally, the mapping program may include algorithms that determine and optimize the actions that the robot must perform to solve the problem.

The set of functions that transform the initial knowledge of the robot into such a system of maps, together with the functions that implement transitions between different layers of maps **upon** request from the subsystem of the robot that solves the tasks assigned to it, constitutes the functional system of modeling SbA.

### 3 Problem Solving

The solution of problems by an system is provided, firstly, by knowledge of the SbA in which the system must solve these problems. And, secondly, by structuring methods – dividing the task into parts that are associated with individual parts of the overall solution path. From these parts, a solution is formed, which is determined by the transition from the condition of the problem to its result. During the solution, individual transformations of the condition of the problem or intermediate results of the solution are performed in such a way as to ultimately obtain the desired result. As the intellectual experience of people who solve problems shows, this process can be quite complex, including dozens and hundreds of separate transformation steps. There are, for example, problems in mathematics that a person cannot solve for centuries, although thousands of attempts are made. Examples are Fermat's Last Theorem (proved only in 1994) or the Riemann Hypothesis. The latter has not yet been resolved.

The general way to solve problems for IS is to use inference to move from the condition of the problem to its solution, using the logic specified for the IS. But if the range of tasks is limited and known, then the developer can single out and independently propose

a set of methods for solving such a task. In this case, the result of the decision depends only on specific data. This approach can be used in the development of robots. Besides, other problem resolution strategies, such as non structure decision making, heuristic approaches or unsupervised learning requires a combination of fact based data analysis and highly subjective and even intuitive judgements about aspects of the problem.

Therefore, when developing an IS, especially in the form of robotic systems, it is desirable to build it in such a way that, under the conditions of using the given SbA, it would be possible to first formulate the classes of problems that such a system should solve. For example, considering the behavior of a military intelligent robot, we simultaneously assume that it must solve a given range of tasks for which this robot is built. One and the same task can be *solved* under different specific conditions and with different possibilities of interaction with the enemy. The intelligence of the robot is manifested in the fact that it manages to find a good solution to the task assigned to it for these conditions, when the general nature of the desired solution is known, but adaptation to the specific features of the environment is determined by the intelligence of the robot. This intelligence is manifested in the fact that, based on the general principles and classes of tasks known to the robot, it must select and solve a problem in a specific environment, based on these principles, using the methods of solution and optimization already known to it.

In many respects, this approach is determined by the fact that in the conditions of interaction with the enemy, the time for solving the problem is limited by the actions of the enemy. The robot does not have time to look for a variety of original solutions, while simultaneously collecting the necessary information that it lacks, analyzing numerous options for possible solutions, and looking for new options. On the other hand, the conditions for solving a problem always remain incompletely defined. Therefore, it is impossible to put one ready-made solution into the robot's memory in advance, removing the intellectual component for it.

For example, considering a military robot of the simplest structure – an intelligent agent with organs of perception and execution – we can formulate the following tasks. The task of ***reconnaissance of the situation*** on the line of contact with the enemy or in his rear, in the face of possible opposition from the enemy; the task of ***destroying military or technical means***, logistics, warehouses, command posts, personnel, in conditions where there are several goals and means of defense against attack; the task of ***countering offensive and reconnaissance*** actions of the enemy; the task of ***protecting personnel and equipment*** in an attack. These are classes of possible problems that may have different solutions depending on the specific conditions of the SbA.

Accordingly, the real range of specific tasks and their solutions that are available to such a robot is much wider. It is important that when solving problems, the robot can use methods and tools that are inaccessible to humans under similar conditions. At the same time, the robot goes through numerous solutions, uses mathematical models and a large amount of memory, which stores the experience of successful solutions. A situation arises, known in modern chess, when a person who owns a computer plus a high-speed computer with a large memory defeats a professional grandmaster.

Modern chess experience shows that in order to solve a narrow class of problems, it is always possible to create a computer that, after training in a certain SbA, will solve

these problems in it more efficiently than a person does. A similar situation occurs when playing poker, in which a trained computer beats a human.

The learning process that determines the creation of an SbA model is an integral part of the process of developing an intelligent robot. Under certain conditions, tasks can be combined with each other into some common complexes. For example, countering enemy reconnaissance activities can be considered as a task of sequentially solving two subtasks: first, the subtask of recognizing and finding enemy assets used for reconnaissance – radar stations, drones and unmanned devices, and other modern technical reconnaissance equipment. This subtask is a component of the main task of reconnaissance of the enemy.

Based on the information collected by the first subtask, the second subtask is solved – the choice of means and the destruction (or incapacitation) of enemy reconnaissance means: disruption or destruction of radar stations with special anti-radar missiles, destruction of drones and drones with missiles or a robot laser, data transmission to other complexes combating enemy reconnaissance assets. This may also include the destruction of command posts for collecting intelligence information with the help of missiles and firepower of the robot. Note that the second subtask can be considered as a component of other classes of problems identified above.

### **3.1 Functional Programming of Problem Solving**

The development of an intelligent robot implies the possibility of programming for it the processes of solving problems that the robot must solve in the course of its activity. In particular, such programming can be performed at the functional level. In this case, for each task, its own function is constructed, which transforms the condition of the tasks into some result, considered as a solution to this task. Let's consider how the functions associated with the above task classes are created.

Based on the fact that the robot has built a system of knowledge about the SbA and modeled this area using its knowledge and the modeling method based on this knowledge, any task from the classes of problems considered above is solved as a task of constructing a function that maps the data from the SbA model to the data corresponding to conditions that are determined by the corresponding class of problems.

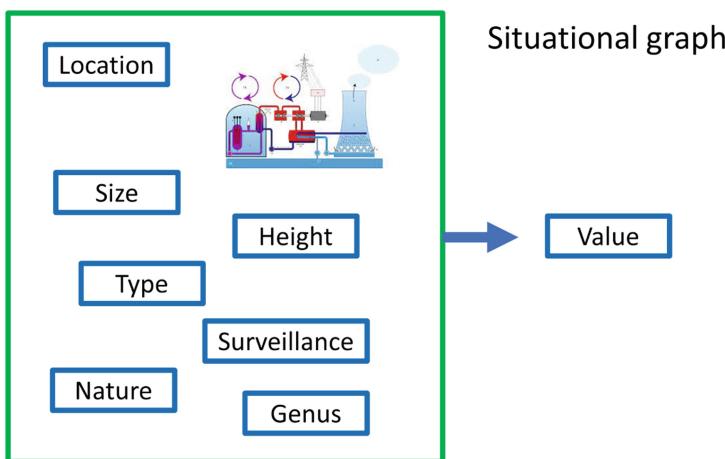
The collection of intelligence information by a robot assumes that using the SbA model it is possible to find the military and technical means of the enemy, which the robot recognized in the SbA. If the model information is presented in the form of a database consisting of named tuples or dictionaries, then we build a function that selects those names that are recognized as enemy assets. At the same time, we bind the found objects to the map of the area. The resulting map is considered as the result of exploration.

The second class of tasks is the destruction of control systems, technical means, enemy logistics or means that are installed on the robot and controlled by it, or by other military or technical means, but based on information received by the robot. This task, if solved by the robot, is divided into subtasks:

- select those enemy assets that need to be destroyed;
- choose their own means, which the robot uses to strike independently or to call stationary means of attack – artillery, missiles, aircraft, drones;

- determine the optimal order of strikes, based on the selected targets;
- calculate the order of actions and movements of the robot, providing it with safety.

It should be noted that when modeling an SbA on the basis of the information received by the robot, not only the formation of features takes place, but also the recognition of objects to which these features correspond. After recognizing an object, its ontology is formed, with which the assessment of this object is associated in the form of an object weight that characterizes the value of the object. It is assumed that the higher the weight, the higher the need to destroy the object. In addition, as an element of the ontology, the robot can specify the choice of means by which the object can be disabled. After that, a *situational graph* is built, the vertices of which are the objects that the robot needs to hit, and the directed arcs from the vertex  $x$  determine the objects that the robot can hit next. See Fig. 3 for an example of situational graph for a military target.

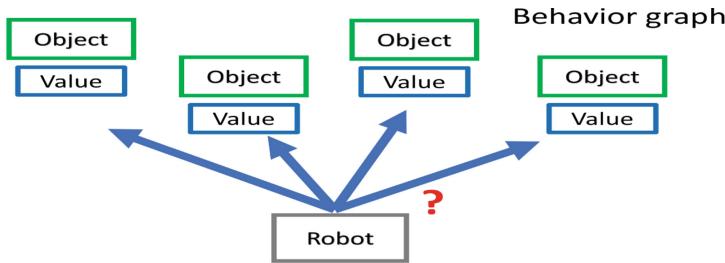


**Fig. 3.** Situational graph for a military target.

The situational graph is used to construct a *behavior graph* representing the optimal order in which the robot performs its tasks (see Fig. 4). When constructing this graph, an algorithm is used to select the next object, depending on the weight of the next object and the ability of the robot to reduce the time of the next strike. This possibility depends on the robot control system – capturing a new target, choosing a possible means of destruction, assessing the degree of enemy opposition, and moving on to the analysis of the next object.

Note that as separate conditions for constructing a behavior graph, it is necessary to take into account the possibility of moving the robot either to create favorable conditions for the next strike, or to ensure its own safety.

The function of constructing a behavior graph – static, in which the robot remains in place while solving its task – delivering successive strikes, or dynamic – in which the position of the robot changes after performing several individual strikes, is the main function used in solving this problem. At the same time, the robot can deliver several



**Fig. 4.** Behavior graph

blows to the same target, separated by time intervals, during which the robot performs other actions.

The following classes of tasks of countering the enemy and protecting people and military support are built as a combination of applying the solutions that were obtained for the intelligence and attack classes. At this level, both the behavior of an individual robot and the combination of several robots to solve one common task can be considered. In this case, the situational graph is represented as the sum of situational graphs of either individual robots or situational graphs in which several robots act synchronously, inflicting joint strikes on enemy objects. In this case, some generalized robot can be formally considered, consisting of several interconnected robots operating in parallel with each other.

The task of functional description is similar to the previous one, with the difference that the number of damaging agents (multiplied by the number of robots included in the association) that such a generalized robot disposes of is increased several times. True, in this case, another problem arises – a generalized robot can be divided into several smaller generalized ones of a smaller size, each of which solves its own separate tasks.

## 4 Conclusion

An intelligent system is a system that uses intelligence to shape its behavior in its environment. This intelligence is determined by two main factors: the first factor is the ability to model the environment, and the second is the use of this model to successfully solve problems that determine the behavior of the system in the environment. Environment modeling is based on the use of knowledge about the environment and its components. The system collects this knowledge using its sensors and sense organs, as well as the knowledge base, which stores information previously collected or incorporated during the development of the system. In particular, this information may include various forms of mathematical structures that form the basis of the environmental model. When solving problems, previous experience collected in the knowledge base can also be applied, simultaneously with a logical conclusion, based on the logic embedded in the intelligent system during its development.

Creating an intelligent system, you can apply functional programming methods. First, by using programming language structures to represent individual knowledge and knowledge bases. Secondly, by developing functions that transform this knowledge

into a representation of the environment model at three possible levels: at the feature level, at the inference level, and at the ontological level. Similarly, problem solving can be represented as a function that transforms the initial knowledge into system actions in the subject area to obtain the required form of representation. At the same time, changes in the system itself are also considered at the functional level. An example is the design of an intelligent military robot, whose task is to protect against an attack by actively influencing the enemy with the help of the means that the robot is equipped with. Note that the approach under consideration is quite universal, applicable in almost any environment.

## References

1. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson (2016)
2. Bostrom, N.: Superintelligence: Paths, Dangers. Oxford University Press, Strategies (2014)
3. Laurier, J.L.: Systems of Artificial Intelligence. Mir, Moscow (1991). (in Russian)
4. Letichevsky, A.: Theory of interaction, insertion modeling, and cognitive architectures. Biol. Inspired Cogn. Architectures **8**, 19–32 (2014)
5. Meitus, V.Yu.: Problems of Building Intelligent Systems. Intell. Model. Cyber. Syst. Analysis (4), 3–19 (2021). (in Russian)
6. Meitus, V.Yu.: Problems of building intelligent systems. Representation of knowledge. Cyber. Syst. Analysis (4), 3–14 (2019). (in Russian)
7. Meitus, V.Yu.: Introduction to the Theory of Intelligent Systems. Basic Representations. Palmarium Academic Publishing, Saarbrücken (2015). (in Russian)
8. Meitus, V.Yu.: Problems of building intelligent systems. Levels of intelligence. Cyber. Syst. Analysis (4), 32–44 (2018). (in Russian)
9. Khaikin, S.: Neural Networks: Full Course. Williams, Moscow (2006). (in Russian)
10. Glibovets, M.M., Gulaeva, N.M.: Evolutionary Algorithms: Assistant. NaUKMA, Kyiv (2013). (in Ukrainian)
11. Summerfield, M.: Python in Practice: Create Better Programs Using Concurrency, Libraries, and Patterns. Addison-Wesley (2013)
12. Lutz, M.: Programming Python. O'Reilly Media, Inc. (2010)
13. Smith, R.G.: Knowledge-Based Systems. Concepts, techniques, examples. Schlumberger-Doll Research (1985). <http://www.reidgsmith.com>
14. Chandrasekaran, B., Johnson, T.R., Smith, J.W.: Task-structure analysis for knowledge modeling. Commun. ACM **35**(9), 124–137 (1992)



# Model-Based Support for Collaborative Concept Mapping in Open-ended Domains

Junya Morita<sup>1(✉)</sup>, Masaji Kano<sup>1</sup>, Shigen Shimojo<sup>2</sup>, Yoshimasa Ohmoto<sup>1</sup>, and Yugo Hayashi<sup>2</sup>

<sup>1</sup> Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu 432-8011, Japan  
j-morita@inf.shizuoka.ac.jp

<sup>2</sup> Ristumeikan University, 2-150 Iwakura-cho, Ibaraki, Osaka 567-8570, Japan

**Abstract.** This paper presents a concept of model-based support for collaborative learning using concept mapping, which is a tool for externalizing learners' knowledge. The proposed support for this activity uses computational modeling of the process of constructing knowledge building. This paper first proposes cognitive models of individual and collaborative concept mapping based on the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture. We then discuss a method for supporting concept mapping in both individual and collaborative learning situations based on the proposed cognitive models.

**Keywords:** Collaboration · Concept map · Cognitive modeling · ACT-R

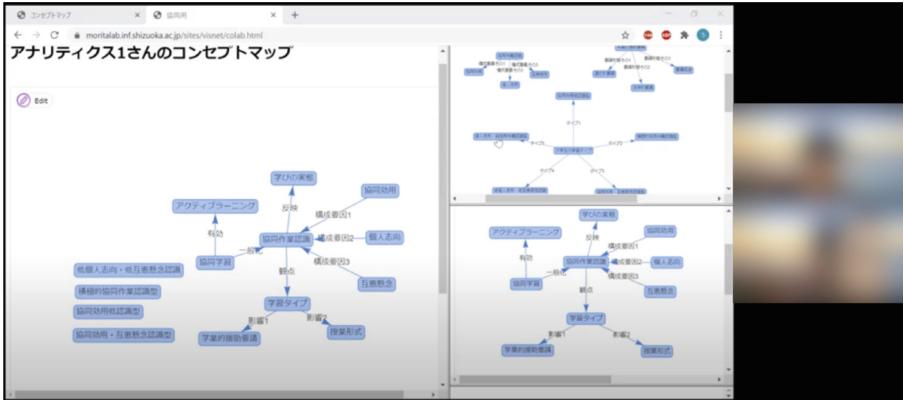
## 1 Introduction

Concept mapping is a tool for modeling and externalizing knowledge stored in learners' internal memory [11, 15]. It requires observing internal thought and clarifying and resolving contradictions in understanding. Research on collaborative learning using concept maps has also been conducted [5, 14].

As an example, Fig. 1 shows an environment developed in our previous study [9], where pairs of learners who have built their own concept maps (the right panels) in advance also interact through a collaborative concept map (the left panel) to merge their understanding on the topic. This activity facilitates detection of discrepancies in understanding between the learners because the collaborative concept mapping externalizes the learner's knowledge in an observable manner. For this reason, concept mapping is expected to be a valuable tool for collaborative problem solving, especially for open problems with no fixed solution.

Following this background, the present study aims to address the following questions:

1. How can the process of constructing collaborative concept mapping be modeled in a computational system?
2. How can the models thus developed be used to support the creation of collaborative concept mapping between learners?



**Fig. 1.** Example of collaborative concept mapping. The details of the activities are shown in our previous paper [9]

The challenge of our study is to explore model-based support in an open-ended domain. In the past, model-based support has usually provided hints to learners solving problems with specific correct answers, such as programming or mathematics problems (e.g., [3]). In contrast, concept mapping has been applied to activities for which it is hard to define one correct solution, such as exploring the causes of natural, social, or psychological phenomena [11, 15] or summarizing academic papers [9]. In this paper, we describe a cognitive model for constructing concept maps as well as the ideas and challenges of modeling this activity in the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture [2], in particular for realizing interventions based on the computational cognitive model. Before presenting our models and supports, we briefly introduce this cognitive architecture.

## 2 Cognitive Architecture and ACT-R

The term “architecture” conventionally indicates a blueprint that connects functions and structures as a plan for constructing an artifact (e.g., [4, 6]). In other words, it is a medium for creating a model in a top-down manner. Cognitive architecture is a type of software architecture that simulates the functions and structure of human cognition [2]. Among the several commonly studied cognitive architectures, ACT-R is both representative and the most frequently used (see [8] for an extensive review of cognitive architectures). Using ACT-R, a modeler can build a cognitive model that executes a task compatible with human participants by combining modules, as has been verified by previous psychological studies.

To model concept mapping activity, we use ACT-R because it features a *declarative module*, a rich model of memory that connects memory elements (chunks) as a semantic network. Figure 2 shows examples of chunks modeled in

---

```

1 (machine-learning isa node label "machine-learning")
2 (supervised-learning isa node label "supervised-learning")
3 (classification isa node label "classification")
4 (regression isa node label "regression")
5 (logistic-regression isa node label "logistic-regression")
6 (learning-method isa array letter "machine-learning"
7   letter2 "supervised-learning")
8 (type1 isa array letter "supervised-learning"
9   letter2 "classification")
10 (type2 isa array letter "supervised-learning"
11   letter2 "regression")

```

---

**Fig. 2.** Examples of chunks.

ACT-R. Each chunk consists of a chunk name as the first element (e.g., *machine-learning*, *supervised-learning*, and *type1*) and the following pairs of attributes and values (e.g., the pair of *isa* (attribute) and *node* (value), the pair of *label* (attribute) and “*machine-learning*” (value)). Among these, concepts such as *machine-learning* and *supervised-learning* are defined as nodes (lines 1–2); these are connected by other chunks such as *learning-method* (line 6–7).

Additionally, chunks in ACT-R are retrieved based on a numerical *activation* parameter. When several chunks match the retrieval cue, the most activated chunk is retrieved. This parameter is computed as a summation of several components, the main ones being *base-level activation* and *spreading activation*. The former is determined by recency and frequency effects (when the chunk was created and how often it is used, respectively). These effects are thought to be related to the personal importance of memory for the individual at the time of processing [1]. In contrast, spreading activation controls memory activation according to the context provided by the external world, such as visual stimuli or goal settings. Thus, we can summarize the mechanism of ACT-R memory theory as the balance of factors driving exploration (spreading activation) and exploitation (base level) in search of a semantic network.

### 3 Cognitive Modeling of Concept Mapping

To realize model-based support for collaborative concept mapping, we first need to consider how such activities are represented in a cognitive architecture like ACT-R. This section presents notions of modeling concept mapping for two situations: individual and collaborative mapping.

#### 3.1 Cognitive Modeling for Individual Concept Mapping

As mentioned earlier, concept mapping can be considered as an activity that externalizes knowledge structure stored in a learner’s memory. The declarative module of ACT-R is assumed to represent a concept map directly because ACT-R stores memory as a network of chunks. In fact, a tool for representing concept

maps such as in Fig. 1 actually holds an analogous data structure represented as in Fig. 2 (see [9]).

In a situation where learners engage in individual concept mapping in a leaning setting, they usually create a novel semantic network connecting existing knowledge elements that have not been explicitly related to each other before. In ACT-R, such a process is modeled by three sequential activities:

1. Input of chunks from a textbook or other source to the model.
2. Retrieval of relevant existing knowledge using input chunks as cues. The activation of chunks governs this process. Thus, memorized chunks with commonalities to the input chunk are activated by spreading activation while prioritizing frequently used recent items (the base-level component).
3. Construction of a new chunk connecting input and retrieved chunks, which is then stored in the declarative module.

This follows a standard process of text comprehension that makes situation models [7] and assumes that an idealized concept mapping activity is the externalization of the semantic network in the declarative module, although the granularity of the externalization depends on several factors (e.g., the time of the task, the user interface for the concept mapping tool).

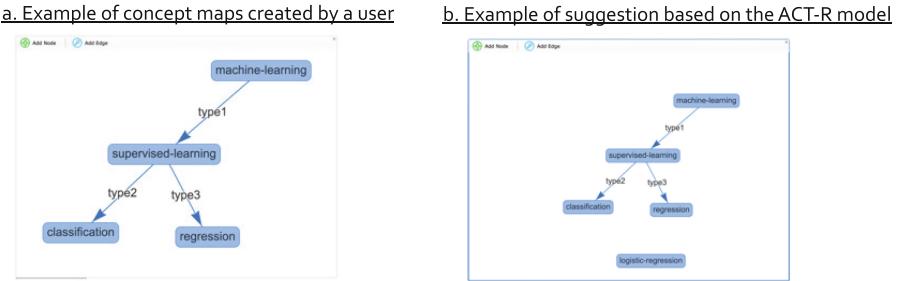
### 3.2 Cognitive Modeling for Collaborative Concept Mapping

As previously mentioned, we essentially do not distinguish between internal (declarative module) and external concept mapping. In this approach, idealized collaborative concept mapping is obtained based on an ACT-R model that merges declarative memories obtained from two learners. However, this merging process can be challenging for learners because of time and technical constraints.

Collaborating learners have to integrate their individual concept maps sequentially. This process requires a visual attention mechanism that selects information from the external world. It is natural to assume that the probability of picking specific visible items varies depending on the learners' characteristics. To model this individual difference, we can utilize models of visual saliency developed for ACT-R [12, 13]. This mechanism assumes the influence of low level visual features such as shapes, color, and on-screen locations on intentional process; also, individual differences of this mechanism come from different experiences stored in the declarative module.

## 4 Model-Based Support for Concept Mapping

So far, we have presented several ideas for modeling concept mapping activity based on ACT-R. This section discusses how this simple model can be used to support learners' concept mapping activities. We discuss two application settings, each utilizing the above individual and collaborative concept mapping models.



**Fig. 3.** Example of concept mapping modeling hint.

#### 4.1 Collaborative Concept Mapping Based on a Cognitive Model of the Learner

First, we consider a situation where a single learner collaboratively constructs a concept map with their model. To create the learner's model in ACT-R, we need to define two main components: chunks in the declarative memory and activation values attached to the chunks. Our approach is to model these components by collecting learning logs for the learner. A modern learning management system (LMS) stores fine-grained learning activities for each user. Chunks described in Fig. 2 can be easily extracted from simple concurrence analysis, leveraging past reports written by the learner.

By utilizing such a personalized cognitive model, we can set up a situation where learners construct their concept maps while receiving personalized hints. Such a situation is predicted to reduce mental costs related to externalizing internal knowledge. Figure 3 illustrates creating such a hint. This example assumes that the concept map presented in Fig. 3a was created by a learner who holds declarative knowledge as in Fig. 2. The ACT-R model holding knowledge in this format can provide a hint by retrieving a node indicating “logistic-regression” and utilizing the current concept map as a cue (Fig. 3b).

In addition to reducing mental cost, a personalized model can support memory retrieval in terms of explore-exploit balance. As noted in Sect. 2, ACT-R explains memory retrieval as the summation of elements relating exploration (spreading activation) and exploitation (base-level activation). The degree of prioritizing these elements is assumed to be varied between individuals. Therefore, collaboration with the model having different tendencies leads well-balanced memory search as presented in the past study [10]. For an over-exploiting learner, collaboration with a model prioritizing spreading activation might generate new insights by relaxing concentrations on already retrieved memory elements. Contrary, for an over-exploring learner, a model with high priority in the base-level component might regulate divergent memory search by presenting important concepts that frequently appear in the text.

## 4.2 Cognitive Model for Connecting Learners via Collaborative Concept Mapping

The support presented in Fig. 3 can also be applied to collaborative concept mapping. It is especially expected to play the role of facilitator in human collaborative concept mapping. In a collaborative situation, the ACT-R model has declarative knowledge constructed from both learners' learning logs. In a normal learning situation, learners could not directly access such joint knowledge (i.e., as a state of collaborative knowledge). Because of this lack of common ground, learners in collaborative learning tend to suffer several conflicts or lack of interaction. Our previous study [9] shows examples of such problems. Here, predictions by the cognitive model including the joint knowledge at each step of the task help resolve conflicts that have emerged during the open-ended collaborative concept mapping task.

## 5 Summary and Future Studies

This manuscript has presented models and supports for engaging in collaborative concept mapping. The ideas presented address the questions identified in the introduction as follows:

1. How can the process of constructing collaborative concept mapping be modeled in a computational system?
  - By externalizing declarative knowledge within constraints of time, tools, and cognitive tendencies (e.g., “explore” vs. “exploit” approaches).
2. How can the models thus developed be used to support the creation of collaborative concept mapping between learners?
  - By using nodes and links to build a model holding knowledge extracted from both participants.

The authors have already developed some of the system modules described here and have encountered some challenges in realizing the proposed support. In the prototype implementation, the ACT-R model retrieves a chunk by observing the current state of the concept map and then modifies the concept map based on the contents of the retrieved chunk (by either adding a node or connecting two nodes with a link). One of the technical challenges is placing nodes in an appropriate location on the screen. In human concept mapping, the location of nodes is important for representing hierarchical structure or grouping. However, the current implementation of the model does not include such higher-order reasoning but only simple perceptual and memory processes. Therefore, in the future, we need to explore method of including structural aspects of knowledge representation (i.e., placing higher-order concepts on the top of the screen) to elicit deeper insight into the learner's cognitive process.

The other challenge is the timing of interventions. Section 4.1 presents ways in which the ACT-R model collaborates with human learners. At present, however, using the model as an agent substituting for human collaborators is challenging.

Reproducing verbal communication, which is a key medium in conventional collaborative learning, is difficult using the ACT-R model alone. Even if we omit verbal communication and arrange for learners to collaborate with a nonverbal online agent, another technical difficulty arises: in order to intervene with appropriate timing, human internal processes must be simulated in real time. Although ACT-R possesses a function that estimates human execution time for each action in a task, it is still difficult to represent the complexities of internal processes in this task.

We believe these challenges can be met successfully through a recursive process of experiment and modification of the system. We need to use recent high-speed technological developments effectively. There are also several practical solutions for the aforementioned challenges (i.e., setting the hint button). In the future, we plan to conduct simulations and experimental studies based on the ideas presented in this paper.

**Acknowledgements.** This work was supported by JSPS KAKENHI Grant Number JP20H04299.

## References

1. Anderson, J.R., Schooler, L.J.: Reflections of the environment in memory. *Psychol. Sci.* **2**(6), 396–408 (1991)
2. Anderson, J.R.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press, Oxford (2007)
3. Anderson, J.R., Boyle, C.F., Reiser, B.J.: Intelligent tutoring systems. *Science* **228**(4698), 456–462 (1985)
4. Bell, C.G., Newell, A.: Computer structures: Readings and examples (1971)
5. Cañas, A.J., et al.: CmapTools: a knowledge modeling and sharing environment (2004)
6. Ford, N., Richards, M.: Fundamentals of Software Architecture. O'Reilly Media, Sebastopol (2020)
7. Kintsch, W.: Text comprehension, memory, and learning. *Am. Psychol.* **49**(4), 294 (1994)
8. Kotseruba, I., Tsotsos, J.K.: 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artif. Intell. Rev.* **53**(1), 17–94 (2018). <https://doi.org/10.1007/s10462-018-9646-y>
9. Morita, J., Ohmoto, Y., Hayashi, Y.: Integrating knowledge in collaborative concept mapping: cases in an online class setting. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 99–106. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-80421-3\\_12](https://doi.org/10.1007/978-3-030-80421-3_12)
10. Morita, J., Pitakchokchai, T., Raj, G.B., Yamamoto, Y., Yuhashi, H., Koguchi, T.: Regulating ruminative web browsing based on the counterbalance modeling approach. *Front. Artif. Intell.* **5**, 1 (2022)
11. Novak, J.D.: Concept mapping: a useful tool for science education. *J. Res. Sci. Teach.* **27**(10), 937–949 (1990)
12. Salvucci, D.D.: An integrated model of eye movements and visual encoding. *Cognit. Syst. Res.* **1**(4), 201–220 (2001)

13. Stewart, T.C., West, R.L.: Cognitive redeployment in ACT-R: salience, vision, and memory. In: The 8th International Conference on Cognitive Modelling, Ann Arbor, MI (2007)
14. Tohyama, S., Miyake, N.: The evaluation of ReCoNote summaries for learner-centered integration. In: 2011 Third International Conference on Intelligent Networking and Collaborative Systems, pp. 855–856. IEEE (2011)
15. White, R., Gunstone, R.: Probing Understanding. The Falmer Press, London (1992)

# **Augmented Intelligence in Healthcare Informatics**



# CPR Emergency Assistance Through Mixed Reality Communication

Manuel Rebol<sup>1,2,3</sup> , Alexander Steinmauer<sup>1</sup> , Florian Gamillscheg<sup>1</sup> , Krzysztof Pietroszek<sup>2</sup> , Christian Gütl<sup>1</sup> , Claudia Ranniger<sup>3</sup> , Colton Hood<sup>3</sup> , Adam Rutenberg<sup>3</sup> , and Neal Sikka<sup>3</sup>

<sup>1</sup> Graz University of Technology, Graz, Austria  
rebol@gwu.edu

<sup>2</sup> American University, Washington, D.C., USA

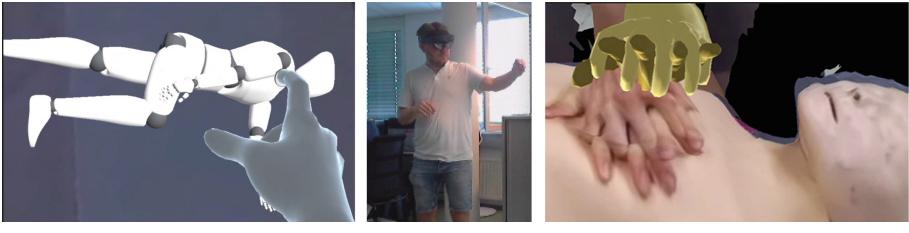
<sup>3</sup> George Washington University, Washington, D.C., USA

**Abstract.** We design and evaluate a mixed reality real-time communication system for remote assistance during CPR emergencies. Our system allows an expert to guide a first responder, remotely, on how to give first aid. RGBD cameras capture a volumetric view of the local scene including the patient, the first responder, and the environment. The volumetric capture is augmented onto the remote expert's view to spatially guide the first responder using visual and verbal instructions. We evaluate the mixed reality communication system in a research study in which participants face a simulated emergency. The first responder moves the patient to the recovery position and performs chest compressions as well as mouth-to-mask ventilation. Our study compares mixed reality against videoconferencing-based assistance using CPR performance measures, cognitive workload surveys, and semi-structured interviews. We find that more visual communication including gestures and objects is used by the remote expert when assisting in mixed reality compared to videoconferencing. Moreover, the performance and the workload for the first responder during simulation does not differ significantly between the two technologies.

**Keywords:** Mixed Reality · CPR · Remote collaboration

## 1 Introduction

In recent years interest and popularity in Augmented Reality (AR) devices have increased drastically and have attracted the attention of the research world and consumers alike [8]. AR has many possible applications that range from entertainment and education to design and manufacturing [8, 21, 28]. AR can be described as a technique that augments the real world with virtual computer-generated information or content with which the user is able to interact [4, 8]. As opposed to Virtual Reality (VR) approaches, where users are fully immersed in the virtual environment and have no connection to the real world, AR tries



(a) Pointing and 3D model (b) Instructor (c) Hand technique and position

**Fig. 1.** First aid assistance through mixed reality. The instructor (b) points at a 3D holographic model that demonstrates the recovery position (a) in the augmented field of view of the first responder. The instructor uses a 3D holographic hand model (c) for guidance when giving chest compressions to the patient.

to superimpose virtual objects upon the user's surroundings [8]. AR technology is not restricted to head-mounted displays (HMD) but is also used in handheld devices like smartphones and tablets [4]. The virtual objects that are embedded in the real world provide information to the user, which can help in performing tasks, such as by helping workers through the electrical wire system of an aircraft [8, 27]. In the past few years AR has progressed immensely and more and more AR applications and devices have entered the market. In medicine, AR technology has enormous potential, where HMDs are for example an essential tool in computer-aided surgery (CAS) [6]. In this field, AR can provide useful information to surgeons in a less distracting manner, compared to information on a monitor [6]. In addition to surgical settings, AR has been used in endoscopic procedures [26], needle puncture procedures [22] and in training for complex surgical procedures [12]. Aside from practical applications, AR has gained a lot of attention in the field of medical education like anatomy, where teaching methods have not changed significantly over the last century. [13]. In the medical field, learning predominantly involves workplace training, which is time-consuming, not very cost-efficient, and comes with some risk [21]. AR can offer a safer learning environment, where concepts can be practiced without the fear of making errors [21]. One area where proper training is particularly important is cardiopulmonary resuscitation (CPR) [9]. Cardiac arrest is a significant public health problem, with approximately 350,000 people per year in the US and Canada alone receiving CPR [36]. Well-executed and timely CPR has a significant impact on survival and neurological outcome, which is why in 2011 the American Heart Association (AHA) published a recommendation for mandatory CPR training starting at school-age [9, 10]. However, multiple studies have shown that even when carried out by healthcare professionals, CPR quality is often poor [1, 16]. With its visual feedback, AR technology can be a valuable tool for guiding responders.

In this research project, we introduce a mixed reality system, where the user is guided by a remote expert in performing first aid (Fig. 1). Our RGBD cameras allow for three-dimensional visual information that helps the user through the steps. Within this project, we conducted an evaluation with 30 participants separated into two groups. Both groups were given the same tasks. They had to give first aid to a lifeless person, bring the person into the recovery position and start with CPR after the person stopped breathing. We compared instruction via the mixed reality (MR) approach (group A) with video-based communication (group B). We analyzed objective metrics of CPR quality recorded by the CPR mannequin, and data from users including workload surveys and interviews. We conclude our main contributions as follows:

1. We introduce an MR communication system designed for remote first aid assistance.
2. We conducted a comparison between MR communication technology and video-based communication.
3. The project team measured workload and performance when giving assisted first aid in MR and videoconferencing.

The paper is structured as follows: Sect. 2 gives an overview of past research of Extended Reality (XR) applications for first aid assistance. In Sect. 3, we discuss the design of the MR communication system, focusing on the different views and their interaction with each other and the devices and software that were used. We evaluate the presented MR communication system in a research study described in Sect. 4. After introducing the study procedure, we present and discuss the results that were obtained. In the end, we summarize the work and give an outlook on future work.

## 2 Related Work

Extended Reality (XR) [32] has been a valuable communication tool especially in medicine. In medical education, AR-supported instructions assist students in demanding tasks by offering a realistic learning environment in which they can develop their theoretical knowledge through didactics, as well as improve their practical skills through interactive simulation [11,35]. In the field of telemedicine, a field that uses communication technology to diagnose and treat patients remotely, AR has developed great potential. Wang *et al.* [37] developed a telemedicine mentoring application using Microsoft’s HoloLens. Their application allows remote experts to guide trainees through complex medical situations with hand gestures that were then displayed in the AR environment of the trainee. Their study, which examined the usability in a trauma setting revealed that their AR setup was regarded as more immersive than other telemedicine approaches. Similarly, Lin *et al.* [25] used AR in a surgical setting in combination with a Microsoft HoloLens. Their study, which used a lower-leg fasciotomy as a training task, revealed that participants that used the headset received a higher performance score and reported higher usability. In surgical settings, AR allows

surgeons to better concentrate on the operating field in comparison to traditional telemedical approaches, where surgeons are often forced to shift their focus to screens for receiving assistance from an expert, which can result in more errors [2]. Andersen *et al.* [2] used a different approach, which features a tablet PC that is located between the local surgeon and the patient. The tablet captures live video that is sent to the remote expert. The remote expert can then annotate the video with graphical information, which is then sent back to the local surgeon's screen. The user study found that participants that used the AR approach completed the tasks with higher accuracy and less distractions. Besides surgical settings, XR technology has been also used for consultation purposes. Anton *et al.* [3] used a combination of AR and VR devices to build a telemedicine system for remote consultations. Their setup consists of an AR client, that captures 3D surface information of the environment, which is then transmitted to the remote expert via a communication module, which enables peer-to-peer connection. On the physician side, the VR client receives the streamed data and is responsible for the rendering on a 3D display, which allows the physician to examine the information and interact with it. In the field of postoperative care, Ponce *et al.* [29] used AR on mobile devices in their study to let physicians virtually examine the patient over a long distance. Their application allows users to interact with each other via mobile devices e.g. with visual annotations on the patient's screen. Their user study revealed that the 96% of the patients regarded the setup as useful, while physicians were slightly less satisfied with 89.6% of them expressing that the application was useful.

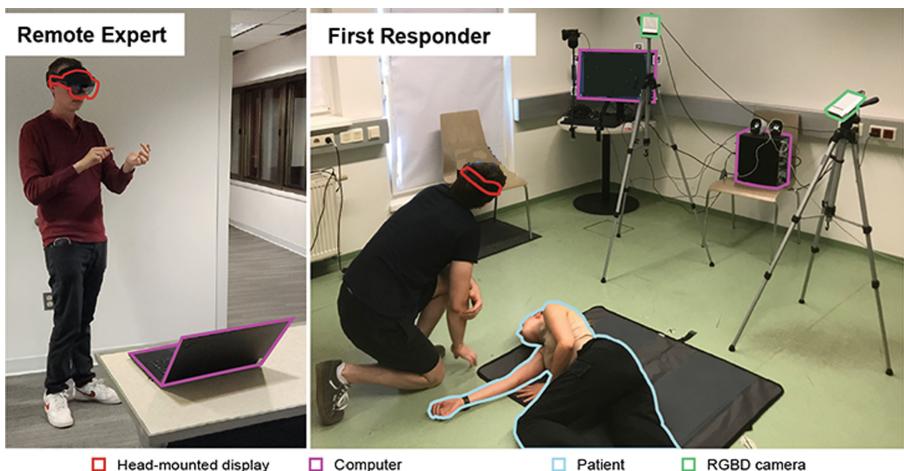
## 2.1 Extended Reality for First Aid Assistance

Multiple studies have demonstrated the potential of extended reality for assisting users in first aid. Already in 1997 Zajtchuk and Satava [38] pointed out the potential of VR for medical education. Kuyt *et al.* [23] included 42 articles in their review on the evolution of XR approaches for CPR training. Their study indicated that the number of both VR and AR applications for resuscitation training is rapidly growing. Due to the high rise of AR-related CPR publications and advancements in technology in recent years, the proportion of AR-related articles is expected to rise in the future. The study concludes that XR shows great potential for CPR training environments, which will likely result in innovations and novel applications. As an example, Girau1 *et al.* [17] proposed a VR-based training application for first aid, using an HTC Vive headset and a real mannequin to provide haptic feedback. Similarly, Blome *et al.* [7] also described a VR setup that uses a non-verbal approach for teaching and practicing reanimation. In contrast to VR, where users are fully immersed in their virtual environment, AR shifts the focus of the user's interaction to the real world [5]. It follows that AR applications can represent tasks in the real world more realistically than their VR-based counterparts. Some studies have already shown the potential of AR in medical education and first aid assistance. Fromm *et al.* [14] developed a concept for an AR application that teaches users first aid. The study revealed that an AR application would help users in emergency situations

in an intuitive and quick matter. Johnson *et al.* [20] used Microsoft's HoloLens for building *HoloCPR*, an AR application that provides real-time instructions for CPR. The subsequent evaluation revealed how the use of such devices can result in a better reaction time and improved accuracy. Frøland *et al.* [15] also used Microsoft HoloLens for developing a training environment for trauma first aid. In their study, they simulated an emergency, where a patient suffers from severe bleeding that had to be stopped by the participants.

### 3 Designing Mixed Reality Communication for First Aid Assistance

We designed the mixed reality (MR) communication system in consultation with a first aid instructor. The domain expert described the information required to understand the emergency situation. Moreover, the domain expert suggested how visual communication can be supported using virtual objects. The detailed system design approach and implementation details approach were described in [33] and [34], respectively.



**Fig. 2.** MR system components. The first responder's scene (right) is captured using RGBD cameras (green) and sent over to the remote expert (left). The remote expert uses the 3D information to guide the first responder using head-mounted displays (red). Computers (purple) handle the network communication and rendering. (Color figure online)

#### 3.1 Augmented Views

Both the expert and the first responder wear the Microsoft Hololens 2 head-mounted display (HMD) to communicate. In addition, to the view, the HMD

offers head and hand tracking. Head tracking is needed to anchor augmented objects in the physical environment. Thus, alignment between the views and visual interaction on a shared view is possible. Hand tracking allows for gesture communication between the users of the system.

*Remote Expert.* The remote expert's view is dominated by a mesh view of the local first responder's scene. The view is captured by Microsoft Azure Kinect RGBD cameras at the local scene. The remote expert can switch between two different RGBD cameras to get a volumetric view of the local scene. The RGBD cameras are positioned such that they capture the patient, the first responder, and the environment. In future versions of the system, the HMD-included camera can be used to capture the local scene instead of the separate RGBD camera. We used a separate Azure Kinect camera because the current Microsoft Hololens depth sensor does not provide sufficient quality.

In addition to the mesh view, the remote expert is presented with an augmented video feed from the local scene. The  $1920 \times 1080$  video feed provides the remote expert with a high-quality view of the local scene including details the mesh view misses because of limitations of the time-of-flight technology.

Besides the views of the local scene, the remote expert sees augmented objects which they can use to visually guide the first responder. The remote expert can manipulate the objects, *i.e.* resize, rotate, move, using gestures.

*First Responder.* The first responder's view is dominated by the physical environment. The augmented information for the first responder is kept to a minimum such that they can focus on the emergency. The augmented information for the first responder consists of augmented objects, the expert's augmented hands, and a video feed showing the remote expert. The first responder's MR communication role is passive. They see augmented objects and screens but do not actively manipulate them.

### 3.2 Interaction

The remote expert and the first responder can verbally communicate using audio. Moreover, the mixed reality system supports visual communication using gestures and case-specific augmented objects. For first aid, we provide a 3D holographic model that demonstrates the rescue position, hands illustrating the chest compression position, and an object to show the depth of the chest compressions. Only the remote expert actively manipulates augmented objects and screens in MR using gestures. The first responder manipulates the physical environment. We show the devices used given the research study setup in Fig. 2.

## 4 Research Study

We evaluated the proposed MR system in a research study in which we compare it against videoconferencing-based communication. Our research focused

on understanding how MR communication can be used in medical emergencies. We measured the workload between video-only and MR assistance and analyzed how it differs. Moreover, we were interested in whether higher performance can be achieved when using MR for emergency assistance compared to video-only communication in the context of first aid/CPR.



**Fig. 3.** CPR assistance through videoconferencing. A remote expert (left) guides the first responder (right) through the emergency using video communication.

#### 4.1 Study Design

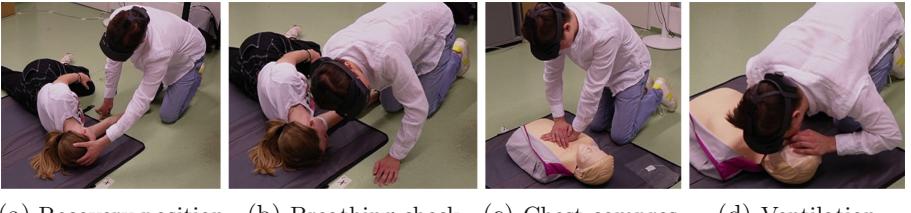
*Subjects.* Altogether, 30 participants were recruited as first responders. The study participants did not receive information about the content of the training prior to arrival at the test site, to prevent preparation and more closely simulate an unexpected emergency. Informed consent was obtained, and subjects were randomized to either test condition. All training sessions were conducted by a single expert who has over 10 years professional experience in first aid training.

*Room Setup.* Expert and first responder were placed in separate rooms to ensure that all communications occurred through the specified modality; room setup remained otherwise static for both conditions. Each room contained a patient actor who simulated an unconscious but still breathing patient, a CPR mannequin (Laerdal Little Anne QCPR mannequin) capable of measuring metrics of CPR quality that are aligned with AHA guidelines [24], and a mask for ventilation. We show the MR setup in Fig. 2, and videoconference setup (Cisco Webex) in Fig. 3.

*First Aid Training.* We illustrate the four main steps actively performed by the first responder in Fig. 4. The responder first assessed the patient actor to determine that they were unconscious but breathing, and placed the patient in the recovery position. The responder then continued to monitor the patient's breathing. After a defined period the patient actor held their breath to simulate cardiac arrest, at which time a CPR mannequin replaced the patient. The responder then

provided chest compressions and mouth-to-mask ventilation. In total each participant was instructed to give CPR for 4 min. Metrics of CPR quality such as rate and depth of compression were recorded from the mannequin during this time.

*Surveys.* At the conclusion of the simulation, participants completed a demographic survey, the NASA Task Load Index (NASA-TLX) [19] and the Simulation Task Load Index (SIM-TLX) [18], both metrics of cognitive workload. Finally, the MR group also completed an MR-specific questionnaire. The expert instructor completed NASA- and SIM-TLX once after completing the training.



(a) Recovery position (b) Breathing check (c) Chest compres. (d) Ventilation

**Fig. 4.** First aid steps. We present the 4 main steps of the first aid simulation. The first responder performs steps (a) and (b) on a patient actor and steps (c) and (d) on a CPR mannequin.

## 4.2 Results and Discussion

*Subject Demographics.* We illustrate the first responder demographics in Table 1.

**Table 1.** First responder demographics and prior experience.

	Video	MR
Age	23(7) y	23(4) y
Male/Female	11/4	11/4
Nr of prior first aid exposures	5(9)	4(6)
Nr of prior MR experiences	0	0

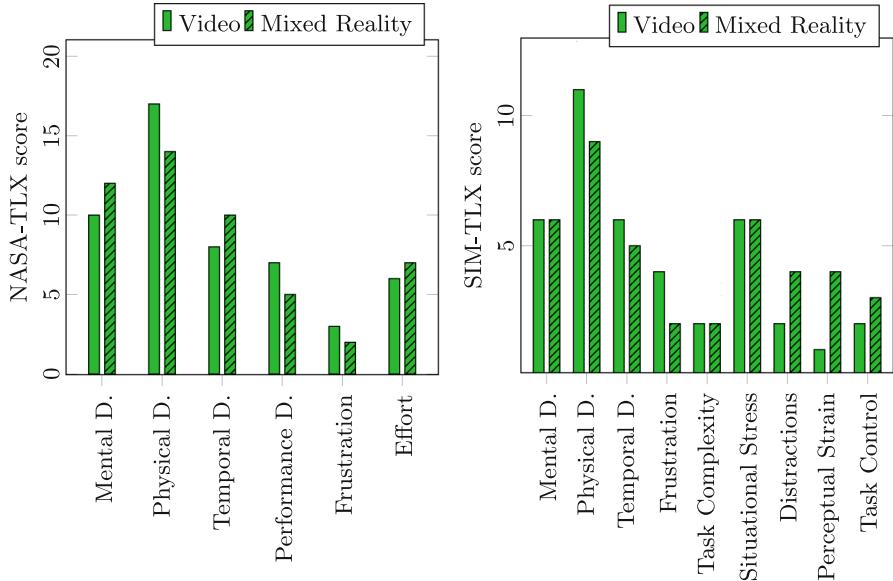
*CPR Performance.* Data for 25 learners (13 video and 12 MR) was included; 5 data samples were removed due to mannequin malfunction. There was no significant difference in CPR performance when comparing video and MR conditions for any CPR performance category (2 sample, 2 tailed t test with significance of  $\alpha = 0.05$ ). The data is presented in Table 2.

**Table 2.** CPR mannequin data. Mean ( $\bar{x}$ ) and standard deviation ( $\sigma$ ) for mannequin-derived CPR performance metrics, in Video and MR. The last column compares the two groups by presenting the p-value of a Student's t-test.

Category	Measurement [unit]	Video		MR		T-test p-val.
		$\bar{x}$	$\sigma$	$\bar{x}$	$\sigma$	
<b>Overall</b>	Score [%]	59%	20%	56%	23%	0.76
	Flow Fraction [%]	60%	6%	60%	10%	0.97
<b>Transition</b>	Heart	33	5	34	8	0.74
	Lung	1.0	0.7	1.3	0.9	0.31
<b>Compress.</b>	Compression Score [%]	82%	17%	68%	19%	0.07
	Number of Compressions [n]	253	33	266	40	0.41
	Avg (n/min)	109	17	114	18	0.48
	Correct Release [%]	92%	17%	80%	28%	0.22
	Correct Depth Rate [%]	96%	8%	94%	12%	0.66
	Compr. w/ adeq. speed [%]	42%	36%	44%	40%	0.87
<b>Ventilation</b>	Ventilation Score [%]	48%	31%	55%	25%	0.53
	Number of ventilations [n]	7	4	10	6	0.23
	Adequate ventilation [%]	99%	5%	99%	3%	0.99
	Hyperventilation [%]	1%	5%	1%	3%	0.99

*Workload.* We illustrate the overall results of both workload surveys, NASA-TLX and SIM-TLX, for first responders and the expert in Table 3. The expert reported a lower overall workload for both surveys when using the MR technology compared to videoconferencing. The first responders reported similar workload between the two technologies. A two-sample two-tailed t-test was not able to show a significant difference ( $\alpha = 0.05$ ) between the groups. The mean overall workload is similar when comparing the two technologies.

We present the per-category workload results for first responders in Fig. 5 to give a more in-depth insight. The physical demand was the highest during the first aid emergency. We posit that this is related to the physical exertion during chest compressions. The SIM-TLX signals three interesting trends. The video-assisted first responders reported higher frustration but lower distraction and perceptual strain than the MR communications group. We propose that the higher frustration with video results from two factors. First, instructions involving physical space must be communicated via 2D video and voice, which must then be translated into actions, rather than by using virtual objects to provide visual demonstration of the necessary action. Secondly, the first responders must switch gaze from the task at hand to the video, in order to compare their own progress with the instructor's directions. In MR, visual communication is more similar to in-person because the views are aligned, and pointing gestures and augmented object demonstrations can be used. The higher perceptual strain reported by the MR group may also be due to the relative increase in visual information presented to and processed by the learner in the MR mode. Higher



**Fig. 5.** First responder workload per category. We compare the NASA-TLX score (top) and the SIM-TLX score (bottom) presented as mean and (standard deviation) of the first responder when getting assistance through video (solid) and mixed reality (line pattern).

**Table 3.** NASA-TLX and SIM-TLX workload presented as mean and (standard deviation) for first responders and the expert during first aid emergency simulations using videoconferencing and MR technology.

	NASA-TLX		SIM-TLX	
	Video	MR	Video	MR
<b>First responder</b>	51(12)	49(14)	39(15)	42(17)
<b>Expert</b>	49	35	39	30

distraction ratings in the MR group may result from the limited field of view (FOV) of the Hololens 2 headset, and the need to locate or track instructor input outside the FOV. We argue that this can be improved by only showing them the remote hands when necessary, for example when the remote expert decides to point or gesture actively.

*Mixed Reality Survey.* The first responders noted in the open-ended questions that they especially liked the gestural communication including pointing by the remote expert. They highlighted the importance of visual communication in stressful situations. When asked about problems with the MR system, some participants noted that the visual instructions were not always given at the best position such that

they had to look around to see the augmented instructions. When asked about what they liked about the experience, the expert reported: "I really liked the possibility to have additional holograms illustrating hand positions or rescue positions. This helped to provide faster help and spend less time with instructions. Additionally, the different camera perspectives were great for evaluating the quality of the CPR (chest compressions and ventilation)." The negative aspects were: "The initial setup of the 2D and 3D areas was sometimes a bit cumbersome. When moving the cube (a handle for moving the location of the visual feeds) the area did not move adequately. Sometimes the connection got lost due to some reason but restarting the system was not a big deal." The expert was also asked about features that could be added: "Some participants complained that audio and visual instructions can be a bit overwhelming. Therefore, a visual way to show the frequency of chest compressions would be great. Especially for the rescue position, a 3D object with movable joints (knee, arm, etc.) would be helpful to show how to bring a person to the rescue position (instead of an object of the already correct position)."

*Discussion.* Although the application of MR devices seem to be futuristic within the context of emergency situations, it enables interesting perspectives. Many countries have successfully implemented guided first aid instructions over an emergency call. A (3D) video channel could provide additional information to the emergency call center regarding the set measures and their quality.

Although the MR and video groups did not differ significantly in terms of CPR performance in this study, use of MR with truly novice first responders may be more beneficial during complex maneuvers, for instance to evaluate and demonstrate head tilt-chin lift for ventilation.

Similar to AEDs, which are usually close by to an emergency situation in urban areas, MR devices could be co-located with them to initiate an emergency call or request for additional support from the emergency call center.

Analyzing the combined performance and workload results, we conclude that the technology is well suited for emergencies because the usage is intuitive. The remote expert only required brief guidance on how the MR technology works. Furthermore, the first responders put on their HMD only about 30 s before the actual simulation started. Moreover, assuming that the technology has a learning curve, better MR results can be expected after using the technology longer or more frequently.

*Broader Impact.* MR technologies are used for communication across industries. Since first aid emergencies are stressful and complex procedures, we believe that the technology can not only be used for other medical procedures including medical training, but also across domains. An example could be remote repair emergencies where a local operator needs to receive guidance from a remote expert to complete the task as quickly and accurately as possible. Similar to giving CPR, spatial information and visual communication including pointing, gestures, and augmented objects seem invaluable.

### 4.3 Limitation

One of the main concerns of this research study is its feasibility for deployment in real-life emergency situations. However, due to the rapid development and spread of AR technologies such devices and application areas are increasing. We believe that as HMDs become lighter and slimmer, sooner than later people will wear MR devices in their everyday lives. Thus, what now seems futuristic, will become relevant in the near future.

The current setup requires some time and experience to be installed. In emergency situations this is not possible. However, the setup can be faster and require fewer devices as technology progresses. For example, instead of Azure Kinect RGBD cameras, the Hololens 2 integrated RGBD cameras could be used to transmit spatial information to the remote expert. This would make the system more portable and easy to deploy in emergencies. Similarly, communication computers can be replaced by handheld phones or better HMD-integrated hardware.

The current research study was conducted on a local network. However, the MR system was designed for remote data transfer over the internet. Especially, given recent 5G developments, bandwidth and latency deficiencies have already been resolved in many parts of the world.

The limitation of low 3D mesh quality when the system is used outdoors because of the interference of natural infrared light with the time of flight sensor technology can be tackled by using prerecorded volumetric information. Moreover, machine learning can be used to fill in missing information and provide the remote expert with a higher-quality 3D mesh view.

## 5 Conclusion

We presented the design and the evaluation of a mixed reality (MR) communication system for first aid. The system allows a remote expert to guide a local first responder through giving first aid. Compared to help over phone and videoconferencing, the MR system allows for augmented visual instructions such as gestures, annotations, and object demonstrations. Moreover, the remote instructor is presented with a volumetric view which gives them spatial information, important for various medical procedures.

We evaluated the MR system in a research study in which we compared MR against videoconferencing. We found that overall, the results including objective CPR mannequin performance and subjective workload measures remains similar between the technologies. We identified many new opportunities that MR offers for an expert to visually guide a first responder. The results show that visual guides do help first responders.

In future work, we will analyze how MR can be used for other medical procedures. Moreover, we are interested in combining prerecorded procedural guidance including avatars [30, 31] for the local operator with active help from the remote expert in problematic situations. This would standardize the process and allow the remote expert to save energy by reducing repetitive instructing and

only intervene in critical situations, when the first responder needs additional guidance.

**Acknowledgements.** The work is supported by National Science Foundation grant no. 2026505 and 2026568.

## References

1. Abella, B., et al.: Quality of cardiopulmonary resuscitation during in-hospital cardiac arrest. *J. Am. Med. Assoc.* **293**, 305–310 (2005)
2. Andersen, D., et al.: An augmented reality-based approach for surgical telementoring in austere environments. *Mil. Med.* **182**, 310–315 (2017)
3. Anton, D., Kurillo, G., Yang, A.Y., Bajcsy, R.: Augmented telemedicine platform for real-time remote medical consultation. In: Amsaleg, L., Guðmundsson, G.P., Gurrin, C., Jónsson, B.P., Satoh, S. (eds.) *MMM 2017. LNCS*, vol. 10132, pp. 77–89. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-51811-4\\_7](https://doi.org/10.1007/978-3-319-51811-4_7)
4. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. *IEEE Comput. Graph. Appl.* **21**(6), 34–47 (2001)
5. Barsom, E.Z., Graafland, M., Schijven, M.P.: Systematic review on the effectiveness of augmented reality applications in medical training. *Surg. Endosc.* **30**(10), 4174–4183 (2016). <https://doi.org/10.1007/s00464-016-4800-6>
6. Birkfellner, W., et al.: A head-mounted operating binocular for augmented reality visualization in medicine - design and initial evaluation. *IEEE Trans. Med. Imaging* **21**(8), 991–997 (2002)
7. Blome, T., Diefenbach, A., Rudolph, S., Bucher, K., von Mammen, S.: VReanimate - non-verbal guidance and learning in virtual reality. In: Proceedings of the International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), pp. 23–30 (2017)
8. Carmignani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., Ivkovic, M.: Augmented reality technologies, systems and applications. *Multimed. Tools Appl.* **51**, 341–377 (2010)
9. Cave, D., et al.: Importance and implementation of training in cardiopulmonary resuscitation and automated external defibrillation in schools a science advisory from the American heart association. *Circulation* **123**, 691–706 (2011)
10. Cheng, A., et al.: Perception of CPR quality: influence of CPR feedback, just-in-time CPR training and provider role. *Resuscitation* **87**, 44–50 (2015). <https://www.sciencedirect.com/science/article/pii/S0300957214008351>
11. Christopoulos, A., Pellas, N., Kurczaba, J., Macredie, R.: The effects of augmented reality-supported instruction in tertiary-level medical education. *Br. J. Edu. Technol.* **53**(2), 307–325 (2022)
12. Coles, T., John, N., Gould, D., Caldwell, D.: Integrating haptics with augmented reality in a femoral palpation and needle insertion training simulation. *IEEE Trans. Haptics* **4**, 199–209 (2011)
13. Duarte, M., Santos, L., Guimarães Júnior, J., Peccin, M.: Learning anatomy by virtual reality and augmented reality. A scope review. *Morphologie* **104**(347), 254–266 (2020). <https://www.sciencedirect.com/science/article/pii/S1286011520300813>
14. Fromm, J., Mirbabaie, M., Stieglitz, S.: The potential of augmented reality for improving occupational first aid. In: Proceedings of the International Conference on Wirtschaftsinformatik (2019)

15. Frøland, T.H., Heldal, I., Ersvær, E., Sjøholt, G.: State-of-the-art and future directions for using augmented reality head mounted displays for first aid live training. In: Proceedings of the International Conference on e-Health and Bioengineering (EHB), pp. 1–6 (2020)
16. Gallagher, E.J., Lombardi, G., Gennis, P.: Effectiveness of bystander cardiopulmonary resuscitation and survival following out-of-hospital cardiac arrest. *J. Am. Med. Assoc.* **274**(24), 1922–1925 (1995)
17. Girau, E., et al.: A mixed reality system for the simulation of emergency and first-aid scenarios. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5690–5695 (2019)
18. Harris, D., Wilson, M., Vine, S.: Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). *Virtual Reality* **24**(4), 557–566 (2019). <https://doi.org/10.1007/s10055-019-00422-9>
19. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Hum. Mental Workload* **1**(3), 139–183 (1988)
20. Johnson, J., Rodrigues, D., Madhuri, G., Weibel, N.: HoloCPR: designing and evaluating a mixed reality interface for time-critical emergencies. In: Proceedings of the EAI International Conference, pp. 67–76 (2018)
21. Kamphuis, C., Barsom, E., Schijven, M., Christoph, N.: Augmented reality in medical education? *Perspect. Med. Educ.* **3**(4), 300–311 (2014). <https://doi.org/10.1007/s40037-013-0107-7>
22. Kim, M., et al.: Efficacy of VeinViewer in pediatric peripheral intravenous access: a randomized controlled trial. *Eur. J. Pediatr.* **171**, 1121–1125 (2012)
23. Kuyt, K., Park, S.H., Chang, T.P., Jung, T., MacKinnon, R.: The use of virtual reality and augmented reality to enhance cardio-pulmonary resuscitation: a scoping review. *Adv. Simul.* **6**(1), 11 (2021). <https://doi.org/10.1186/s41077-021-00158-0>
24. Laerdal: Little Anne QCPR (2022). <https://laerdal.com/us/products/simulation-training/resuscitation-training/little-anne-qcpr/>
25. Lin, C., et al.: A first-person mentee second-person mentor AR interface for surgical telementoring. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 3–8 (2018)
26. Mahmud, N., Cohen, J., Tsourides, K., Berzin, T.: Computer vision and augmented reality in gastrointestinal endoscopy. *Gastroenterol. Rep.* **3** (2015)
27. Mizell, D.: Augmented reality applications in aerospace. In: Proceedings of the IEEE and ACM International Symposium on Augmented Reality, Los Alamitos, CA, USA, p. 11. IEEE Computer Society (2000). <https://doi.ieeecomputersociety.org/10.1109/IWADS.2000.880913>
28. Nee, A., Ong, S., Chryssolouris, G., Mourtzis, D.: Augmented reality applications in design and manufacturing. *CIRP Ann.* **61**(2), 657–679 (2012). <https://www.sciencedirect.com/science/article/pii/S0007850612002090>
29. Ponce, B.A., et al.: Telemedicine with mobile devices and augmented reality for early postoperative care. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4411–4414 (2016)
30. Rebol, M., Gütl, C., Pietroszek, K.: Real-time gesture animation generation from speech for virtual human interaction. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA 2021. Association for Computing Machinery, New York (2021). <https://doi.org/10.1145/3411763.3451554>
31. Rebol, M., Gütl, C., Pietroszek, K.: Passing a non-verbal turing test: evaluating gesture animations generated from speech. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pp. 573–581 (2021)

32. Rebol, M., Pietroszek, K.: Artificial Reality Continuum. Springer, Heidelberg (2023)
33. Rebol, M., et al.: Collaborative system design of mixed reality communication for medical training. In: Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 418–427 (2023)
34. Rebol, M., et al.: Mixed reality communication for medical procedures: teaching the placement of a central venous catheter. In: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 346–354 (2022)
35. Tang, K., Cheng, D., Mi, E., Greenberg, P.: Augmented reality in medical education: a systematic review. *Can. Med. Educ. J.* **11** (2019)
36. Travers, A., et al.: Part 4: CPR overview: 2010 American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* **122**, S676–S684 (2010)
37. Wang, S., et al.: Augmented reality as a telemedicine platform for remote procedural training. *Sensors* **17**, 2294 (2017)
38. Zajtchuk, R., Satava, R.M.: Medical applications of virtual reality. *Commun. ACM* **40**(9), 63–64 (1997). <https://doi.org/10.1145/260750.260768>



# The Relative Importance of Cognitive and Behavioral Engagement to Task Performance in Self-regulated Learning with an Intelligent Tutoring System

Xiaoshan Huang<sup>1</sup> , Shan Li<sup>2</sup> , and Susanne P. Lajoie<sup>1</sup>

<sup>1</sup> McGill University, Montréal, QC H3A 0G4, Canada

xiaoshan.huang@mail.mcgill.ca

<sup>2</sup> Lehigh University, Bethlehem, PA 18015, USA

**Abstract.** Self-regulated learning (SRL) is essential in promoting students' learning performance, especially in technology-rich environments where learning can be disorienting. Student engagement is closely associated with SRL, although the regulation of engagement in SRL is still underexplored. In this study, we aimed to compare the relative importance of cognitive and behavioral engagement in the three SRL phases (i.e., forethought, performance, self-reflection) to learning performance in the context of clinical reasoning. Specifically, students were tasked to solve two virtual patients in BioWorld, an intelligent tutoring system. We measured student behavioral engagement as their time spent on diagnostic behaviors. Students' cognitive engagement was extracted from their think-aloud protocols as they verbalized their thinking and reasoning process during the tasks. We analyzed the relative importance of cognitive and behavioral engagement in the three SRL phases to diagnostic efficacy. Results suggested that the effects of engagement on student performance depend on task complexity. In the complex task, the six predictors (i.e., two types of engagement in the three SRL phases) explained 36.81% of the overall variances in learner performance. Cognitive engagement in SRL played a more significant role than behavioral engagement in predicting students' performance in clinical reasoning.

**Keywords:** Self-Regulated Learning · Cognitive Engagement · Behavioral Engagement · Relative Importance · Intelligent Tutoring System

## 1 Introduction

Intelligent tutoring systems (ITS) can promote learners' self-regulated learning (SRL), a pivotal competency for achieving better learning performance [1–3]. Meanwhile, student engagement is also found to be positively related to their learning performance within ITS [4]. The literature is well-documented about the measurement and effect of two types of student engagement, i.e., behavioral engagement and cognitive engagement, on student learning. Behavioral engagement is typically defined as on-task behaviors in

learning or problem-solving. Cognitive engagement involves specific learning-oriented strategies [5]. Effectively facilitating behavioral and cognitive engagement in SRL yields optimal learning performance. Nevertheless, limited studies have investigated the patterns of behavioral and cognitive engagement within SRL processes and their effects on learning performance. Particularly, to our knowledge, no study has examined the relative importance of behavioral and cognitive engagement to task performance in SRL with an ITS.

In this study, we explored medical students' cognitive and behavioral engagement in different SRL phases as they solved clinical reasoning tasks with an ITS. We attempted to understand the relative importance of behavioral and cognitive engagement, as they occurred in each phase of SRL, to student performance. In line with the literature, we conceptualized behavioral engagement as on-task behaviors. We used think-aloud protocols to capture students' thinking and reasoning activities in clinical reasoning, based on which we calculated students' cognitive engagement in SRL phases. We will explain the measurement of behavioral and cognitive engagement further in the method section. In sum, the overarching goal of this study is to unravel the relationships among SRL, behavioral engagement, cognitive engagement, and task performance as students solved clinical reasoning problems with an ITS. We also look into how task complexity affects those relationships. This study has practical and methodological implications. This study also informs the design of augmented intelligence in intelligent tutoring systems.

## 2 Theoretical Framework

SRL is an active, constructive process through which students monitor, regulate, and control their learning efforts in both external and internal dimensions [6]. SRL is a cyclical process consisting of three phases [7]: forethought, performance, and self-reflection. In the forethought phase, students set goals and plan the strategies and effort to achieve their goals. The performance phase involves self-observation and self-control, whereby students monitor and control their actions to approach their goals. In the self-reflection phase, students reflect on their learning trajectories and progresses, based on which they make adaptations towards desired outcomes. In technology-rich environments (e.g., ITS), SRL is pivotal to better performance, higher achievement, and a satisfied learner experience [8, 9].

Engagement was first defined in school contexts as a multidimensional concept consisting of behavioral, cognitive, emotional, and psychological components [5]. In this study, we were particularly interested in the behavioral and cognitive aspects of engagement. The differentiation between behavioral and cognitive engagement helps understand student effort in learning. Behavioral engagement is the observable behaviors of students in completing a learning task, whereas cognitive engagement shows the amount of mental effort students investigate in working on the task [5]. Researchers have reached a consensus that both behavioral and cognitive engagement is positively related to task performance. Instructors should design strategies or scaffoldings to promote behavioral and cognitive engagement. Nevertheless, rarely the relative importance of behavioral and cognitive engagement to task performance is studied. The research is

even more scarce when it comes to the context of self-regulated learning. Researchers have investigated the relationships between SRL and engagement in various contexts, e.g., collaborative learning [2], and flipped classroom [3]. For instance, findings from a qualitative study indicated that engagement varied across an SRL cycle [10]. However, no studies have compared the relative importance of the two types of engagement (i.e., behavioral engagement and cognitive engagement) within the SRL process. It is also noteworthy that the concepts of engagement and SRL could sometimes be less well-defined, which may lead to ambiguous interpretations [11]. The boundaries between behavioral and cognitive engagement are blurred. In this study, we make a clear distinction between the two constructs, and we compare their relative importance to task performance in SRL with an ITS.

We situate this study in clinical reasoning, a complex problem-solving setting that requires critical thinking and reasoning skills. In clinical reasoning, medical practitioners diagnose patients (or virtual patients) by inquiring about a patient's symptoms and life experience, performing medical tests, seeking for guidance, proposing diagnostic hypotheses, assessing the probability of each hypothesis, and making a final decision. As such, clinical reasoning asks medical practitioners to be behaviorally and cognitively engaged during diagnostic processes. Meanwhile, clinical reasoning consists of forethought, performance, and self-reflection activities. Therefore, clinical reasoning represents a good example of studying the relationships between engagement, SRL, and task performance. Specifically, this study aims to answer the following two research questions:

- (1) What is the relative importance of cognitive and behavioral engagement in the three SRL phases (i.e., forethought, performance, and self-reflection) to clinical reasoning performance?
- (2) Does task complexity affect the relative importance of cognitive and behavioral engagement to clinical reasoning performance?

### 3 Methods

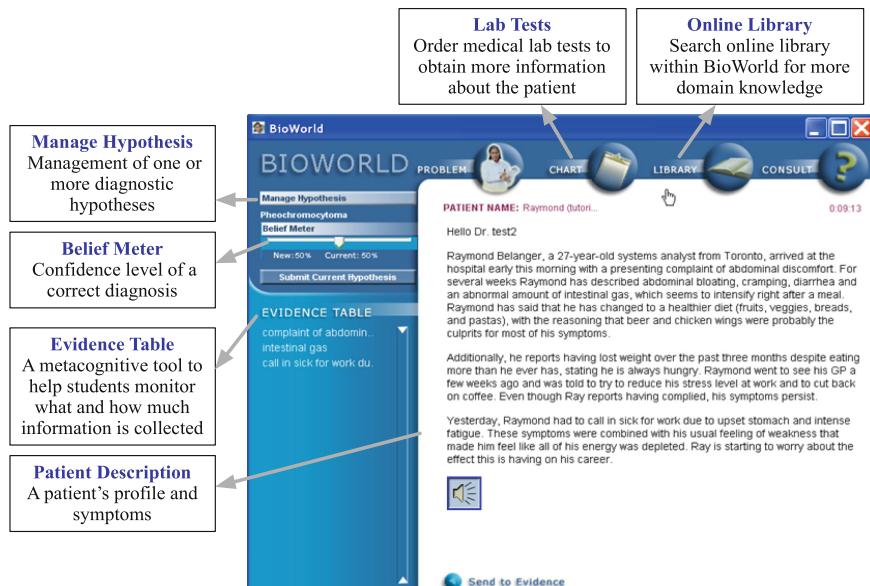
#### 3.1 Participants

Thirty-three medical students (23 Female, 10 Male;  $M_{age} = 23.24$ ,  $SD = 3.00$ ) from a large North American university volunteered to participate in this study. We had obtained ethical approval from the Research Ethics Board (REB) Office at the institute where the study was conducted. The participants signed a consent form before the experiment. They were also informed that they could withdraw anytime during the experiment without worrying out detrimental consequences. During the experiment, the participants were asked to diagnose two virtual patients (i.e., one easy task and one complex task). The students were asked to think aloud while diagnosing patient cases with an ITS, BioWorld. One participant only completed the think-aloud in the easy task, whereas the remaining 32 students did think-aloud in both tasks. Thus, we used 65 cases altogether to study learners' engagement within SRL phases.

### 3.2 Learning Contexts and Task

BioWorld is an intelligent tutoring system designed to assist medical students in practicing clinical reasoning skills in a simulated learning environment [12, 13] (see Fig. 1). Each case in BioWorld begins with a description of a virtual patient and pertinent symptoms. Students select useful evidence based on the presented case facts by recalling their prior knowledge of the disease. Then, students formulate one or more hypotheses. To validate or refute their findings, students can order medical lab tests or search an embedded medical library for additional explanations. After providing a final diagnosis, students are required to assess the relevance of the collected evidence to their hypotheses, justify the probability of each hypothesis, and summarize their clinical reasoning processes.

In this study, participants were assigned to resolve the Amy and Cynthia patient cases in BioWorld. The accurate diagnoses for the two cases were Type 1 diabetes and pheochromocytoma, respectively. The Amy case was designed to be simple, whereas the Cynthia case was designed to be difficult.



**Fig. 1.** The main interface of BioWorld [12, 14].

### 3.3 Procedure

Before the experiment, we provided a training session to teach participants how to use the BioWorld system. In addition, they had the opportunity to become familiar with the system by solving a sample case in BioWorld. After the training session, participants were required to independently solve two clinical cases (i.e., the Amy case and the

Cynthia case) for a total of 1.5 h. Students' diagnostic behaviors were automatically recorded in the log files of the BioWorld system in real-time. We randomized the order of the patient cases to counteract its effect on participant performance.

### 3.4 Measurements and Diagnostic Performance

Using the log file data, we identified eight types of diagnostic behaviors in the three SRL phases (i.e., forethought, performance, and self-reflection). Collecting evidence items (CO) was coded as the forethought phase. Three clinical reasoning behaviors were coded in the performance phase: raising/managing hypotheses (RA), adding tests (AD), and searching the library for information (SE). The self-reflection phase included categorizing evidence/results (CA), linking evidence/results (LI), prioritizing evidence/results (PR), and summarization for final diagnosis (SU). Specifically, we calculated the portion of students' behavior duration (i.e., how long a behavior lasts) in each SRL phase as their behavioral engagement. The duration of a diagnostic behavior was obtained by subtracting the timestamp of the behavior itself from the subsequent one. The portion of behavior duration in each SRL phase was the accumulated duration of each SRL phase divided by the total duration.

Moreover, we manually coded six SRL strategies in the three SRL phases (i.e., forethought: orientation and planning; performance: monitoring, evaluation, elaboration; reflection: self-reflection) based on a previous coding scheme (see Table 1, [15]). We used SRL strategies to indicate learners' cognitive engagement in each SRL phase. Three coders were assigned 30% of the protocols in a way that two raters coded each protocol. We used Cohen's Kappa to assess inter-rater reliability in that it is widely regarded as the gold standard of reliability indices [16]. The final inter-rater reliability for SRL activities was .813, indicating a high degree of agreement between coders [17]. We then calculated the frequency of each SRL phase and used the accumulated score of each SRL phase divided by the total SRL score to represent cognitive engagement in the relevant SRL phase.

**Table 1.** The Coding Scheme of Self-Regulated Learning Strategies in Clinical Reasoning.

Code	Sub-Code	Definition
Orientation (OR)	Hypothesizing (H)	Outlining a single or multiple diagnoses
	Identifying/repeating important information (IR)	Highlighting information in relation to the case
	Activating prior knowledge (APK)	Recalling information pertaining to the disease
Planning (PL)	Goal setting (GS)	Formulating a superordinate goal wherein multiple actions plans are executed

(continued)

**Table 1.** (*continued*)

Code	Sub-Code	Definition
	Sub-goaling: Form action plan (FAP)	Forming a plan to order lab tests/identify symptoms while formulating a diagnosis
	Sub-goaling: Looking for information (LI)	Forming a plan to search for information in the library while formulating a diagnosis
	Sub-goaling: Using external source to get an explanation (ES)	Requesting a consult from BioWorld while formulating a diagnosis
	Sub-goaling: Organizing thoughts by self-questioning (OSQ)	Asking questions to oneself in reference to the action plan being formulated
Monitoring (MO)	Claiming progress in understanding (CP -/+)	Mentioning that their overall understanding of the case improved (+). Alternatively, mentioning partial or lack of understanding (-)
	Found required information (FI -/+)	Mentioning that lab tests ordered were pertinent/abnormal (+) or non-pertinent/normal (-)
	Noticing unfamiliar words/terms (NU)	Mentioning an instance of confusion that is attributed to terminology
	Noticing inconsistency, confusion, checking plausibility (NI)	Mentioning an instance of confusion pertaining to formulating their diagnosis
Evaluation (EV)	Checking (CH)	Reviewing the evidence, symptoms, and vital signs
	Verifying (V)	Claiming/Proving/Disproving something exists or is true
	Justifying (J -/+)	Justifying the probability of a hypothesis based on evidence items. The evidence items can be used to increase the likelihood of a particular diagnosis (+) or to decrease it (-)
	Concluding (CO -/+)	Make a conclusion about the virtual patient

*(continued)*

**Table 1.** (*continued*)

Code	Sub-Code	Definition
Elaboration (EL)	Integrating information	Adding more information to existing information to create a more complex, emergent whole
	Connecting parts of the text by reasoning (CR -/+)	Connecting different pieces of information in a logical way to form a conclusion or judgment
Self-Reflection (RE)	Self-questioning for reflection	Asking questions to oneself to check their understanding of certain contents or procedures
	Causal attribution	Attributional judgments about the results, such as whether poor performance is because of one's limited ability or insufficient effort
	Self-satisfaction	Self-satisfaction involves perceptions of satisfaction or dissatisfaction and associated affect regarding one's performance

In terms of student diagnostic performance, we calculated the match between a participant and experts' solutions. It is worth mention that experts' solutions, including useful evidence items, medical tests needed for diagnosis, and a correct diagnosis, are prefigured in the BioWorld system. Once a participant submit their final diagnosis, the system will generate the performance score for the participant [3].

### 3.5 Data Analysis

We used a relative importance analytic method [18] to compare the relative importance of students' behavioral and cognitive engagement in the three SRL phases (i.e., forethought, performance, and self-reflection) to task performance. In other words, there were six predictive variables, i.e., two types of engagement in the three SRL phases. The relative-weight analysis deals with predictor intercorrelations using a variable transformation approach [19]. It is an alternative to the multiple regression technique, tackles the multicollinearity issue, and helps to determine the variables' rank in importance. It helps answer the question "which variable is the most essential and ranks variables based on their R-Square contribution." Relative importance weights are estimated for each predictor variable and offer a complete decomposition of the total predicted variance to the model  $R^2$  [19, 20]. This method is promising to deal with multiple predictors when determining which predictor is more important or useful than others to predict an outcome variable. The data analysis is conducted in R using the package "relaimpo" [21].

## 4 Results

- (1) What is the relative importance of cognitive and behavioral engagement in the three SRL phases to clinical reasoning performance?

Table 2 summarizes the relevant importance weights of all predictor variables on diagnostic performance. Relative importance weights sum to  $R^2$ , which represents the percentage of variance explained in the criterion that can be attributed to each predictor [20]. Regardless of the task complexity, the result showed that the predictors' overall explanation of the variance in diagnostic outcome (i.e., efficacy) was 27%.

**Table 2.** Relative Importance of Behavioral and Cognitive Engagement to Task Performance

Outcome Variable	Predictor Variables					
	Cognitive Engagement (CE)			Behavioral Engagement (BE)		
	FT	PE	RN	FT	PE	RN
Relative Importance	.03	.04	.04	.04	.05	.07

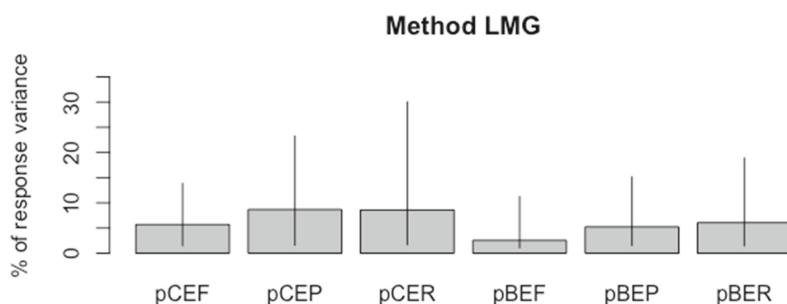
**Note.** FT: the forethought phase of SRL, PE: the performance phase of SRL, RN: the self-reflection phase of SRL

- (2) Does task complexity affect the relative importance of cognitive and behavioral engagement to clinical reasoning performance?

When distinguishing task complexity, a significant portion of explanation on the variances in efficacy was found in the difficult diagnostic case ( $R^2 = 36.81$ ) but not in the simple case ( $R^2 = 17.05$ ). The results indicated that the six engagement-related variables could explain 36.81% of the variances in learners' diagnostic efficacy when they solved a difficult task (Fig. 2). The relative importance indices for cognitive engagement in the three SRL phases (i.e., FT, PE, RN) are .06, .09, and .09, respectively. Regarding behavioral engagement in the three SRL phases, the relative weights are .03, .05, and .06, respectively. As shown in Fig. 2, cognitive engagement has greater contribution weights in all three SRL phases than behavioral engagement. Additionally, cognitive engagement and behavioral engagement in the self-reflection phase contributed higher predictive value on performance than engagement in the other two SRL phases.

In contrast, in the easy patient case, the predictive power of the six engagement-related variables were not as strong as it was in the difficult task (see Table 3). Specifically, the predictive power was less than half of the explanation power of a difficult task. It is suggested to run a bootstrapping procedure to determine the statistical significance of relative weights [18]. We applied 1,000 bootstrapping replicates to check if a particular type of engagement in a certain SRL phase is significantly more important than the other

## Relative importances for Efficacy with 95% bootstrap confidence intervals



$R^2 = 36.81\%$ , metrics are not normalized.

**Fig. 2.** The relative importance of six engagement-related variables to diagnostic efficacy. **Note.** pCEF: cognitive engagement in the forethought phase, pCEP: cognitive engagement in the performance phase, pCER: cognitive engagement in the self-reflection phase; pBEF: behavioral engagement in the forethought phase, pBEP: behavioral engagement in the performance phase, pBER: behavioral engagement in the self-reflection phase.

engagement. The result indicated that no specific engagement stood out from others when predicting diagnostic efficacy in the difficult task (i.e., all confidential intervals for difference include zero).

**Table 3.** Relative Importance of Behavioral and Cognitive Engagement in the Different Task

Task Complexity	Predictor Variables					
	Cognitive Engagement (CE)			Behavioral Engagement (BE)		
	FT	PE	RN	FT	PE	RN
Easy	.02	.02	.01	.03	.03	.06
Difficult	.04	.10	.16	.01	.01	.03

**Note.** FT: the forethought phase of SRL, PE: the performance phase of SRL, RN: the self-reflection phase of SRL

## 5 Discussions and Conclusions

In this study, we compared the relative importance of cognitive and behavioral engagement in the three SRL phases (i.e., forethought, performance, self-reflection) to diagnostic performance in clinical reasoning with an ITS. Findings from this study suggested that the effects of behavioral and cognitive engagement on student performance depends

on task complexity. Behavioral and cognitive engagement plays a more important role in predicting performance in a difficult task than in an easy task. Since we did not perform inferential statistics, this finding should be interpreted with caution. Despite this, our results highlighted the importance of behavioral and cognitive engagement in difficult tasks. In the easy task, the six engagement-related variables (i.e., two types of engagement in the three SRL phases) could explain 17.05% of the variance in students' diagnostic performance. A possible explanation is that other factors, such as prior knowledge [22], emotions [23], motivation [2], interact with students' SRL processes and outcomes. As task complexity increases, the importance of behavioral and cognitive engagement to performance increases. As a result, the six predictors explained 36.81% of the overall variances in learner performance in the difficult task.

While the six engagement-related variables were strong predictors of performance in a difficult task, there were no statistically significant differences among the variables regarding their relative importance to task performance. However, based on the descriptive analysis, students' cognitive engagement in all three SRL phases had a stronger influence on the variances of performance than behavioral engagement. This supports the importance of integrating cognitive engagement in SRL [24]. The result also suggested that instructors should emphasize both behavioral and cognitive engagement in complex clinical reasoning tasks, although more emphasis should be placed on cognitive engagement. In contrast, researchers and instructors should examine the factors other than engagement in easy tasks since engagement only accounts for a limited variance in student performance. For instance, prior knowledge may have a stronger influence on learners' performance when solving an easy case using ITS [25].

In sum, this study was one of the first to investigate the relative importance of behavioral and cognitive engagement in SRL phases to task performance with an ITS. Findings from this study underscored the factor of task complexity in studying the relationships between engagement and performance. As discussed above, this study has significant practical implications. This study also informs the design of intelligent tutoring systems. For example, the current intelligent tutoring systems can automatically analyze on-task behaviors, providing important insights about learners' behavioral engagement and behavioral patterns. It would be interesting to integrate augmented intelligence in an ITS so that it can grade the importance of behavioral engagement to learning performance in tasks of varying complexity.

**Acknowledgements.** This work is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) under the grant number of 895-2011-1006. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the SSHRC.

## References

1. Pérez-Sanagustín, M., et al.: A MOOC-based flipped experience: scaffolding SRL strategies improves learners' time management and engagement. *Comput. Appl. Eng. Educ.* **29**, 750–768 (2021). <https://doi.org/10.1002/cae.22337>

2. Duffy, M.C., Azevedo, R.: Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Comput. Hum. Behav.* **52**, 338–348 (2015). <https://doi.org/10.1016/j.chb.2015.05.041>
3. Lajoie, S.P.: Student modeling for individuals and groups: the BioWorld and HOWARD platforms. *Int. J. Artif. Intell. Educ.* **31**(3), 460–475 (2020). <https://doi.org/10.1007/s40593-020-00219-x>
4. Fisher, R., Perényi, Á., Birdthistle, N.: The positive relationship between flipped and blended learning and student engagement, performance and satisfaction. *Act. Learn. High. Educ.* **22**, 97–113 (2021). <https://doi.org/10.1177/1469787418801702>
5. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**, 59–109 (2004). <https://doi.org/10.3102/00346543074001059>
6. Pintrich, P.R.: A conceptual framework for assessing motivation and self-regulated learning in college students. *Educ. Psychol. Rev.* **16**, 385–407 (2004). <https://doi.org/10.1007/s10648-004-0006-x>
7. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. *Theory Into Practice* **41**, 64–70 (2002). [https://doi.org/10.1207/s15430421tip4102\\_2](https://doi.org/10.1207/s15430421tip4102_2)
8. Azevedo, R.: Reflections on the field of metacognition: issues, challenges, and opportunities. *Metacogn. Learn.* **15**(2), 91–98 (2020). <https://doi.org/10.1007/s11409-020-09231-x>
9. Fitzpatrick, C.: Optimizing Learning in a Technology Rich Environment: The Importance of Cognitive and Metacognitive Processing. Presented at the Society for Information Technology & Teacher Education International Conference March 2 (2009)
10. Järvelä, S., Järvenoja, H., Malmberg, J., Isohätälä, J., Sobociński, M.: How do types of interaction and phases of self-regulated learning set a stage for collaborative engagement? *Learn. Instr.* **43**, 39–51 (2016). <https://doi.org/10.1016/j.learninstruc.2016.01.005>
11. Pizzimenti, M.A., Axelson, R.D.: Assessing student engagement and self-regulated learning in a medical gross anatomy course. *Anat. Sci. Educ.* **8**, 104–110 (2015). <https://doi.org/10.1002/ase.1463>
12. Lajoie, S.P.: Developing professional expertise with a cognitive apprenticeship model: examples from avionics and medicine. In: Ericsson, K.A. (ed.) *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*, pp. 61–83. Cambridge University Press, Cambridge (2009)
13. Li, S., Zheng, J., Poitras, E., Lajoie, S.: The allocation of time matters to students' performance in clinical reasoning. In: Nkambou, R., Azevedo, R., Vassileva, J. (eds.) *ITS 2018. LNCS*, vol. 10858, pp. 110–119. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91464-0\\_11](https://doi.org/10.1007/978-3-319-91464-0_11)
14. Li, S., Zheng, J., Lajoie, S.P., Wiseman, J.: Examining the relationship between emotion variability, self-regulated learning, and task performance in an intelligent tutoring system. *Educ. Tech. Research Dev.* **69**(2), 673–692 (2021). <https://doi.org/10.1007/s11423-021-09980-9>
15. Lajoie, S.P., Lu, J.: Supporting collaboration with technology: does shared cognition lead to co-regulation in medicine? *Metacognition Learn.* **7**, 45–62 (2012). <https://doi.org/10.1007/s11409-011-9077-5>
16. Watson, P.F., Petrie, A.: Method agreement analysis: a review of correct methodology. *Theriogenology* **73**, 1167–1179 (2010). <https://doi.org/10.1016/j.theriogenology.2010.01.003>
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977). <https://doi.org/10.2307/2529310>
18. Tonidandel, S., LeBreton, J.M., Johnson, J.W.: Determining the statistical significance of relative weights. *Psychol. Methods* **14**, 387–399 (2009). <https://doi.org/10.1037/a0017735>

19. Dalal, R.S., Baysinger, M., Brummel, B.J., LeBreton, J.M.: The relative importance of employee engagement, other job attitudes, and trait affect as predictors of job performance. *J. Appl. Soc. Psychol.* **42**, E295–E325 (2012). <https://doi.org/10.1111/j.1559-1816.2012.01017.x>
20. Tonidandel, S., LeBreton, J.M.: Relative importance analysis: a useful supplement to regression analysis. *J Bus Psychol.* **26**, 1–9 (2011). <https://doi.org/10.1007/s10869-010-9204-3>
21. R: A Language and Environment for Statistical Computing (2014)
22. Trevors, G., Duffy, M., Azevedo, R.: Note-taking within MetaTutor: interactions between an intelligent tutoring system and prior knowledge on note-taking and learning. *Educ. Tech. Res. Dev.* **62**(5), 507–528 (2014). <https://doi.org/10.1007/s11423-014-9343-8>
23. Huang, X., Huang, L., Lajoie, S.P.: Relationships between SRL and teachers' emotions in technology integration. In: EDULEARN20 Proceedings, pp. 2249–2256 (2020). <https://doi.org/10.21125/edulearn.2020.0692>
24. Li, S., Lajoie, S.P.: Cognitive engagement in self-regulated learning: an integrative model. *Eur. J. Psychol. Educ.* **37**(3), 833–852 (2022). <https://doi.org/10.1007/s10212-021-00565-x>
25. Taub, M., Azevedo, R.: How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive SRL processes during learning with an intelligent tutoring system? *Int. J. Artif. Intell. Educ.* **29**(1), 1–28 (2018). <https://doi.org/10.1007/s40593-018-0165-4>



# Emotional Impact of Cognitive Priming on Alzheimer's Disease

Hamdi Ben Abdessalem<sup>(✉)</sup> and Claude Frasson

Département d'Informatique et de Recherche Opérationnelle, Université de Montréal,  
Montréal H3C 3J7, Canada

hamdi.ben.abdessalem@umontreal.ca, frasson@iro.umontreal.ca

**Abstract.** Subjective cognitive decline (SCD) is an early state of Alzheimer's Disease. As this disease continues to spread, there is still no efficient pharmacological treatment to stop it. Non-pharmacological treatment may be effective to slow down the progression of the disease. We propose in this paper a cognitive priming system which projects the name of forgotten celebrity names in a very fast way (29 ms) so only the subconscious part of the patients' brain can intercept it. Experiments with different SCD patients have been realized divided into 2 groups. The first group was to investigate the impact of the cognitive priming system, and the second one was to serve as a control group. Results show that SCD patients' frustration decreased when the cognitive priming system projected the missing information. In addition, results show that their memory performance increased when the system performed the cognitive priming technique. Thus, cognitive priming can reduce negative emotion and improve memory performance.

**Keywords:** Cognitive Priming · Alzheimer's Disease · Negative Emotions · EEG · Behavior and Cognition · Health Care · Wellbeing

## 1 Introduction

Alzheimer's Disease (AD) is one of the most critical diseases and affects millions of people every year, yet there is no effective pharmaceutical treatment to improve the course of the disease. AD patients suffer from negative emotions which affect their cognition and memory as well as their well-being [1]. The most vulnerable persons to the disease are those over 65 years old and the risk increases with age [2]. AD is a neurodegenerative disease which deteriorates gradually over time. We focus our research on older adults with subjective cognitive decline (SCD) as these individuals progress to dementia at a higher rate than those with no subjective impression of decline and are sometimes in the early stages of the disease [3].

Non-pharmacological interventions such as music therapy and animal-assisted therapy have positive effects on cognitive function of older adults suffering from SCD or AD [4]. Furthermore, relaxing environments such as Virtual Travel [5], Zoo Therapy [6], and Music Therapy [7] mixed with virtual reality have been demonstrated to be helpful to **reduce negative emotions** and improve cognitive performance and short-term memory [8].

Another non-pharmacological technique which could be used to improve the memory of SCD patients, is the cognitive priming method. The principle of cognitive priming is to project information very quickly so the conscious part of the brain cannot receive it, but the subconscious part can intercept it [9]. We believe that this technique could help restore forgotten memory of AD patients. Is there a link with the reduction of negative emotions?

In this goal we conducted experiments with SCD participants to verify the following research questions. **Q1: Can cognitive priming reduce negative emotions of SCD patients? Q2: Is there a relation between the effect of cognitive priming on memory performance and its effect on negative emotions?**

The rest of this paper is organized as follows. In Sect. 2 we present an overview of the related work. Section 3 presents our methodology. In Sect. 4 we present the experiments. Finally, in Sect. 5 we present and discuss the obtained results.

## 2 Related Work

### 2.1 Non-pharmacological Treatment for Alzheimer's Disease

Several researchers utilized non-pharmacological approaches to treat AD patients. One of these approaches is Animal Assisted Therapy (AAT). AAT is a treatment that concentrates on psychosocial stimulation and human-animal interactions for therapeutic purposes. AD patients can have difficulties interacting with other people. The nonverbal action of a dog can be easily interpreted. Animals can play a role as transitional objects, allowing humans to first establish a connection with them and then extend that connection with other humans. AAT has a relaxing effect on patients with dementia and Alzheimer's disease [10], particularly as therapy with dogs. Virtual Reality Zoo Therapy [6] incorporates the theory of AAT in addition to virtual reality to maximize its effect with immersion.

Furthermore, another relaxing approach used for therapeutic reasons is music. It is commonly used for different conditions, such as autism, brain injury, and dementia [11–13]. Byrns et al. proved that they can reduce AD patients' negative emotions by combining music therapy with EEG brain assessment in a virtual environment [7].

### 2.2 Cognitive Priming

Other than relaxing environment, cognitive priming could be an alternative solution to improve memory performances. In fact, priming is defined as an **unconscious** form of human memory. It affects the perceptual identification of words and objects and has been recognized as separate from other forms of memory or memory system [9].

Cognitive priming is a type of priming which aims to enhance specific cognitive processes such as decision-making toward the goal of indirectly enhancing knowledge acquisition [14]. Del Cul et al. were the first to utilize cognitive priming [15]. In addition, DeVaul used cognitive priming to better associate names to faces [16]. Moreover, Lowery et al. showed that cognitive priming can increase performance on midterm exams compared to neutral primes and that subliminal priming may have long-term effects on real-world behavior [17]. Finally, Chalfoun and Frasson used these techniques and discovered a clear memory and cognitive improvement [18].

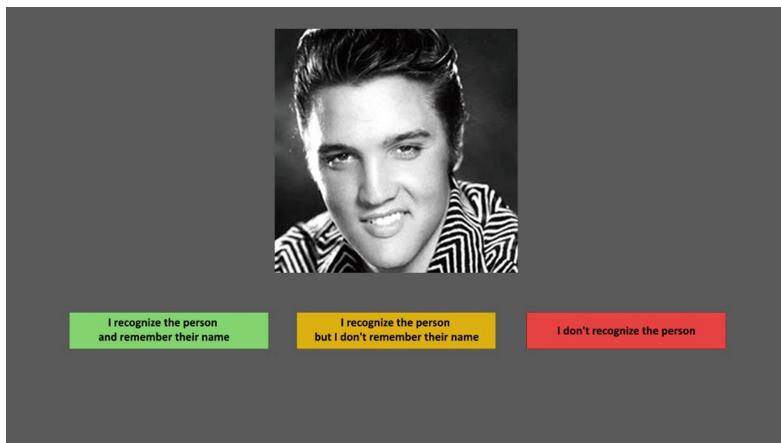
### 3 Methodology

Our proposal is to utilize the approach of cognitive priming technique to reduce SCD patients' negative emotion and restore missing information in their memory. Therefore, we created a cognitive priming system to identify the missing information and project it to SCD patients. The principle is to project an image containing the name of a celebrity that the participant did not recognize in addition to their picture during a short time (29 ms). This projection reaches the subconscious mind of the participant. The speed at which information is projected is crucial since there is a threshold for which the user can consciously perceive information (Del Cul et al., 2007 [15]).

Our system operates in **two phases**. In the first one, it presents celebrity pictures and for each one, the participant should select one of these three choices:

1. I recognize the celebrity and remember their name.
2. I recognize the celebrity but do not remember their name.
3. I do not recognize the celebrity.

Figure 1 illustrates a screen capture of celebrity Elvis Presley and the three choices listed above.



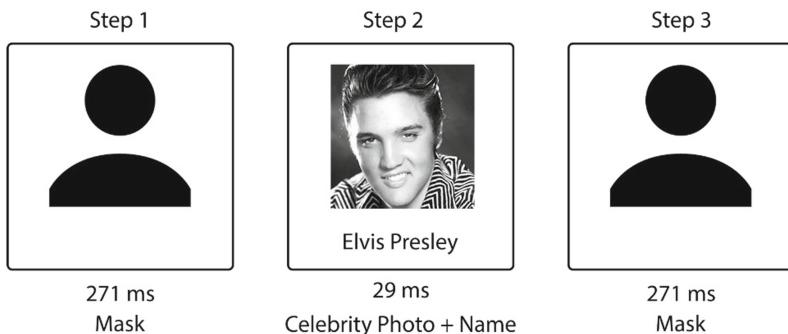
**Fig. 1.** Screen capture of the cognitive priming system

In the first phase, the system presents 25 celebrity pictures or more and stops when the patient selects the second or third option at least 5 times. After that, the participants take a break by looking at a gray screen for 3 min. Following that, the second phase begins in which the system presents again the celebrity that the patient did not recognize or remember the name during the first phase. During this phase, prior to presenting each picture, the system performs the cognitive prime method following these steps.

1. Step 1: the system presents a mask for 271 ms

2. Step 2: the system presents one of the unrecognized celebrities in addition to their name for 29 ms
3. Step 3: the system presents a mask for 271 ms.

Figure 2 shows an example of the sequence of cognitive priming method.



**Fig. 2.** Example of cognitive priming method

## 4 Experiments

We conducted experiments on 34 total participants with SCD. The groups of participants were divided into 2 groups. The first group includes 18 participants, and the second group includes 16 participants. The first group performed the cognitive priming session and the second one was a control group, thus they did not.

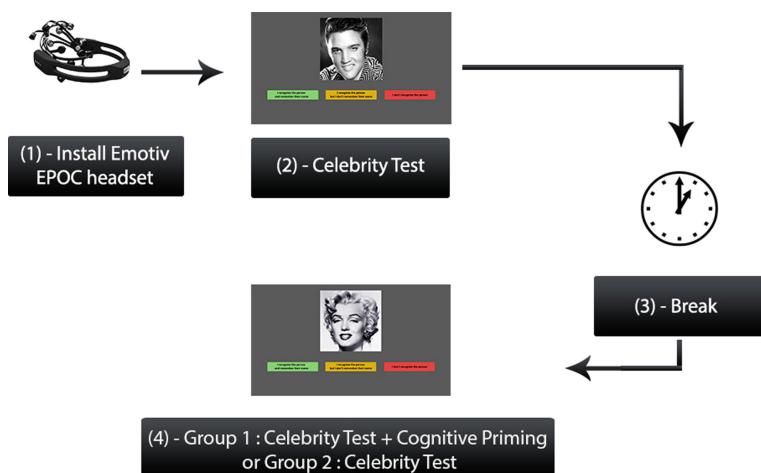
Before the experiments, all the participants took a pre-experiment session in which we make sure that they are SCD patients, and they are eligible to complete the experiments. Our eligibility criteria were the following:

- Older than aged 60 of age.
- Francophone.
- Normal or correct-to-normal vision.
- Normal hearing.
- Met the Consortium for the Early Identification of Alzheimer's Disease – Quebec (CIMA-Q) criteria for SCD:
  - Presence of a complaint defined as a positive answer to the following statements: “My memory is not as good as it used to be” “and it worries me.”
  - MoCA 20–30.
  - No impairment on the logical memory scale based on the education-adjusted CIMA-Q cut-off scores.

During the experiments of the first group, a series of photo appear one by one, and the participants must select if they recognize the person or not and know their name

or not (celebrity test). Following, they take a three-minute break. After that, a second series of photo containing the ones they did not recognize and their name appear again and the participant must select the same options. However, during the second series, the cognitive priming technique is used by the system as indicated in Fig. 2 (celebrity test + cognitive priming).

For the second group, the system presents a series of photo one by one, and the participants must select if they recognize the celebrity or not and know their name or not. Following, they take a three-minute break. Then, the system presents a second series of photo containing the ones they did not recognize, and they need to select the same options. During the second series, the system does NOT perform the cognitive priming technique, since this second group is our control group. Figure 3 details the process of the experiment.



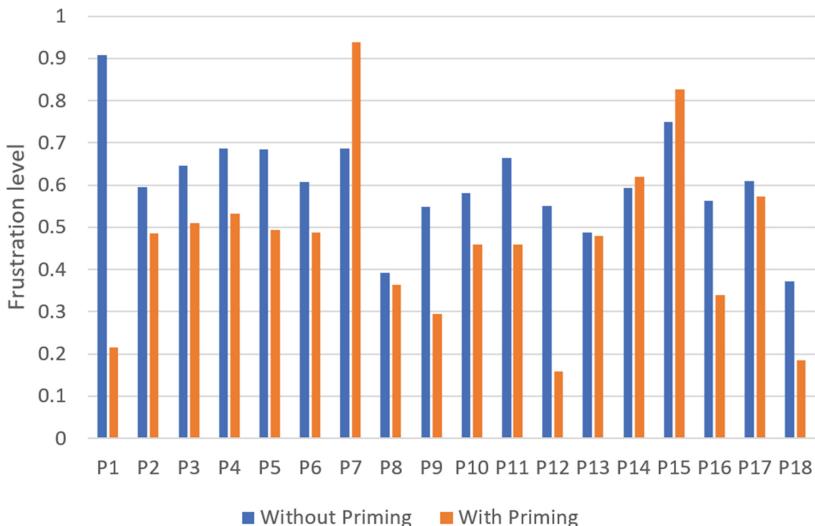
**Fig. 3.** Process of the experiment

To perform the experiments, we used an MSI GE75 Raider 95F which has a Full-HD display capable of 144 Hz refresh rate. This screen refreshes every ~6.94 ms, in addition, our cognitive priming system runs at 250 frames per second ( $>144$ ). Therefore, we are sure that the 29 ms cognitive priming will appear on the screen.

## 5 Results and Discussion

The first objective of this research is to explore if **cognitive priming reduces negative emotions of SCD patients**. To this end, we analyzed participants' frustration level (extracted from Emotiv EEG). We calculated the mean level of frustration for each participant comparing the results without the cognitive priming (before the break) with the results with the cognitive priming (after the break). Results show that the mean frustration without cognitive priming was 0.61 and with cognitive priming was 0.47. Thus, the participants' frustration decreased when using the cognitive priming method.

Individual results are shown in Fig. 4, in which we can observe a frustration decrease for 16 of the 19 participants. Only participants 7, 14 and 15 failed to show the effect. This can be explained by the fact that participants succeeded in recovering memory and developed more positive motions.

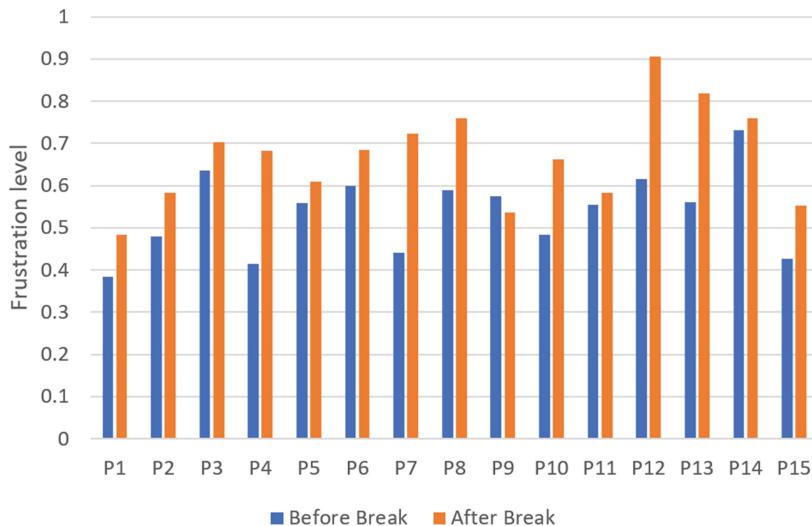


**Fig. 4.** Frustration level per participant before and after cognitive priming.

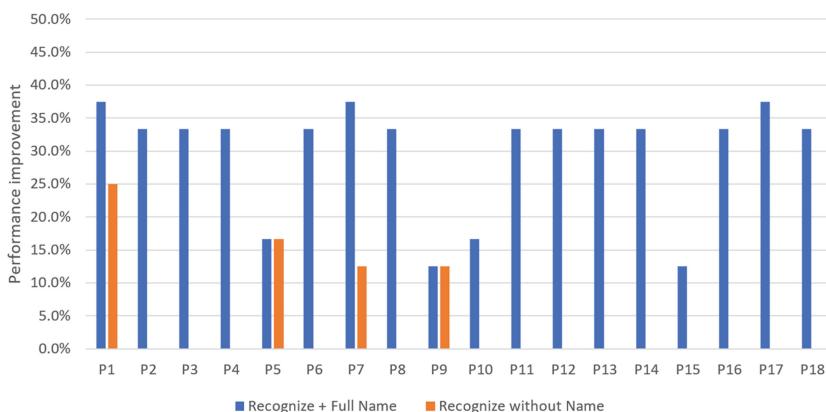
To ensure that the obtained results are due to the cognitive priming effect, we analyzed participants' frustration level of the control group. We calculated the mean level of frustration for each participant, comparing the results before the break with the results after the break. Results show that the mean frustration before the break was 0.54 and after the break was 0.67. Therefore, the participants' frustration increased after the break. Individual results are shown in Fig. 5, in which we can observe a frustration increase for most participants. This can be explained by the fact that the participants still cannot retrieve the name of the celebrities.

In addition, and to verify our second research question and see if there is a **correlation between the effect of cognitive priming on memory performance and its effect on negative emotions**, we proceeded by analyzing participants' responses before and after the cognitive priming. We calculated the improvement for each participant comparing the results before the cognitive priming with the results after. Figure 5 shows the individual obtained results in which we can notice a clear increase in performance. When the participants respond that they do not recognize the celebrity at all before the priming, and after the priming they select that they recognize them but do not remember their name, we consider this an improvement as well, however, these results are separated in Fig. 6 (orange).

To confirm that the observed results are an effect of the cognitive priming, we examined the percentage improvement of the control group as well. Figure 7 shows the individual obtained results in which we can observe a clear difference of performance



**Fig. 5.** Frustration level per participant without cognitive priming (control group).

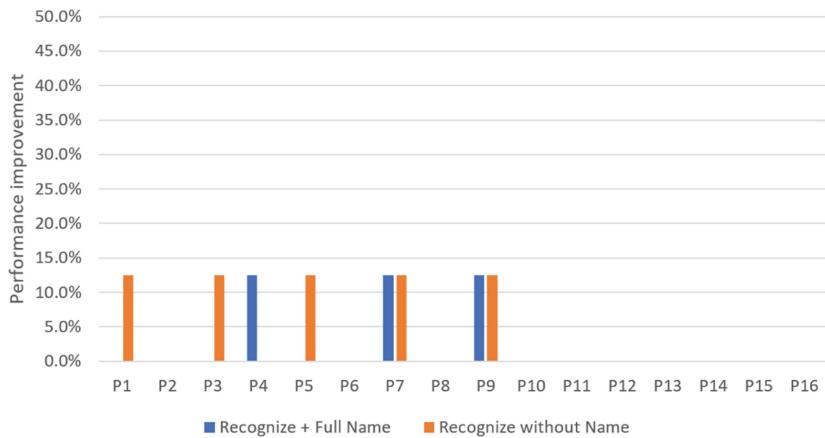


**Fig. 6.** Performance improvement per participant using cognitive priming. (Color figure online)

improvement comparing the improvement observed in Fig. 6. For the control group without the cognitive priming, most participants' memory did not improve.

Consequently, these results confirm that the use of cognitive priming technique can reduce negative emotions and improve memory performance. We believe that this reduction of negative emotions is due to memory performance improvement and satisfaction. As the participants retrieved missing information in their memory, they became less frustrated, and their negative emotions were reduced.

These results encourage us to use our cognitive priming system with learning applications. Since learners experience negative emotions such as stress, we believe that the use of our method to help them acquire missing information, can reduce their negative



**Fig. 7.** Performance improvement without cognitive priming (control group).

emotions and thus improve their learning capabilities. For instance, in the case of learning medical applications, medical students try to solve a medical case by reading the symptoms, performing analysis, choosing a diagnosis, and performing medical actions. We can modify our system to perform cognitive priming of the missing information when it detects that their negative emotions are increasing.

## 6 Conclusion

In this paper, we introduced a novel approach which uses cognitive priming as a non-pharmacological treatment for SCD patients. Thus, we created a cognitive priming system for SCD patients aiming to investigate the effect of the cognitive priming technique on their emotions and memory. The experiments were divided into two groups. For the first group, the system performed the cognitive priming technique. The second group was our control group, so the system did not perform cognitive priming. Results showed that our system can increase memory performance using cognitive priming. In addition, results showed that cognitive priming system can reduce negative emotions. Besides SCD treatment, this system could be used to help reduce learners' negative emotions by performing cognitive priming of their missing information.

**Acknowledgment.** We acknowledge NSERC-CRD (National Science and Engineering Research Council Cooperative Research Development), PROMPT, and Beam Me Up for funding this work. Thanks to Alexie Byrns who contributed to the experiments. We also acknowledge the contributions of Sylvie Belleville and Marc Cuesta from IUGM (Institut Universitaire de Gériatrie de Montréal).

## References

1. Benoit, M., et al.: Behavioral and psychological symptoms in Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* **10**, 511–517 (1999)
2. Boyle, P.A., et al.: Attributable risk of Alzheimer's dementia attributed to age-related neuropathologies. *Ann. Neurol.* **85**, 114–124 (2019)
3. Jessen, F., et al.: A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's Dementia* **10**, 844–852 (2014)
4. Tsoi, K.K., Chan, J.Y., Ng, Y.-M., Lee, M.M., Kwok, T.C., Wong, S.Y.: Receptive music therapy is more effective than interactive music therapy to relieve behavioral and psychological symptoms of dementia: a systematic review and meta-analysis. *J. Am. Med. Dir. Assoc.* **19**, 568–576 (2018)
5. Abdessalem, H.B., et al.: Application of virtual travel for Alzheimer's Disease. In: 9th International Conference on Sensor Networks, pp. 52–60 (2020)
6. Ben Abdessalem, H., Ai, Y., Marulasidda Swamy, K.S., Frasson, C.: Virtual reality zoo therapy for Alzheimer's disease using real-time gesture recognition. In: Vlamos, P. (ed.) GeNeDis 2020, pp. 97–105. Springer International Publishing, Cham (2021). [https://doi.org/10.1007/978-3-030-78775-2\\_12](https://doi.org/10.1007/978-3-030-78775-2_12)
7. Byrns, A., Abdessalem, H.B., Cuesta, M., Bruneau, M.-A., Belleville, S., Frasson, C.: EEG analysis of the contribution of music therapy and virtual reality to the improvement of cognition in Alzheimer's Disease. *JBiSE* **13**, 187–201 (2020). <https://doi.org/10.4236/jbise.2020138018>
8. Abdessalem, H.B., Byrns, A., Frasson, C.: Optimizing Alzheimer's disease therapy using a neural intelligent agent-based platform. *IJIS* **11**, 70–96 (2021). <https://doi.org/10.4236/ijis.2021112006>
9. Tulving, E., Schacter, D.L.: Priming and human memory systems. *Science* **247**, 301–306 (1990). <https://doi.org/10.1126/science.2296719>
10. McCabe, B.W., Baun, M.M., Speich, D., Agrawal, S.: Resident dog in the Alzheimer's special care unit. *West J. Nurs. Res.* **24**, 684–696 (2002). <https://doi.org/10.1177/019394502320555421>
11. Bradt, J., Magee, W.L., Dileo, C., Wheeler, B.L., McGilloway, E.: Music therapy for acquired brain injury. In: The Cochrane Collaboration (ed.) Cochrane Database of Systematic Reviews. p. CD006787.pub2. John Wiley & Sons, Ltd., Chichester, UK (2010). <https://doi.org/10.1002/14651858.CD006787.pub2>
12. Geretsegger, M., Elefant, C., Mössler, K.A., Gold, C.: Music therapy for people with autism spectrum disorder. *Cochrane Database Syst. Rev.* **2014**, CD004381 (2014). <https://doi.org/10.1002/14651858.CD004381.pub3>
13. Moreno-Morales, C., Calero, R., Moreno-Morales, P., Pintado, C.: Music therapy in the treatment of dementia: a systematic review and meta-analysis. *Front. Med.* **7**, 160 (2020). <https://doi.org/10.3389/fmed.2020.00160>
14. Jraidi, I., Chalfoun, P., Frasson, C.: Implicit strategies for intelligent tutoring systems. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 1–10. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-30950-2\\_1](https://doi.org/10.1007/978-3-642-30950-2_1)
15. Del Cul, A., Baillet, S., Dehaene, S.: Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol.* **5**, e260 (2007). <https://doi.org/10.1371/journal.pbio.0050260>
16. DeVaul, R.W., Pentland, A., Corey, V.R.: The memory glasses: subliminal vs. overt memory support with imperfect information. In: Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings. pp. 146–153. IEEE, White Plains, NY, USA (2003). <https://doi.org/10.1109/ISWC.2003.1241404>

17. Lowery, B.S., Eisenberger, N.I., Hardin, C.D., Sinclair, S.: Long-term effects of subliminal priming on academic performance. *SSRN J.* (2006).<https://doi.org/10.2139/ssrn.937950>
18. Chalfoun, P., Frasson, C.: Showing the positive influence of subliminal cues on learner's performance and intuition: an ERP study. In: Aleven, V., Kay, J., Mostow, J. (eds.) *ITS 2010. LNCS*, vol. 6095, pp. 288–290. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13437-1\\_47](https://doi.org/10.1007/978-3-642-13437-1_47)



# Combining XR and AI for Integrating the Best Pedagogical Approach to Providing Feedback in Surgical Medical Distance Education

Yves S. Nkulu-Ily<sup>(✉)</sup>

University of Kentucky, Lexington, KY 40506, USA  
lyshabo@hotmail.com

**Abstract.** This position paper maintains that learners' involvement and accurate representation through cues and stimuli are essential aspects of an AI-guided XR app for learning surgery. It then shows how and why it is crucial to conduct expert interviews when designing/developing such an app. A multidisciplinary team of 22 experts was interviewed to provide insights into new visual technologies' inter-communication (i.e., Extended Reality-XR, Artificial Intelligence -AI or Machine Learning-ML, and 360° videos) and the pedagogical aspects (i.e., educational theories, techniques, and practices) of designing an XR system supported by an ML-based subsystem. These insights are about instructional design elements to maximize the impact of simulation-based training (SBT). The paper guides educational games for medical surgery design and user experience theory and concepts. It shows how to capture the actual task's and environment's fundamental characteristics expressing realistic behaviors. It shows the significant issues (e.g., occlusion and the use of instruments) in capturing these tasks and possible options to overcome them (e.g., using video feeds from endoscopies and robotic surgery). It also provides practical opportunities for facilities in resource-constrained regions lacking internet access/reliability/speed. For qualifying impact, this paper provides the best way to measure learning experience return on investment for medical education.

**Keywords:** extended reality · artificial intelligence · intelligent system

## 1 Introduction

The higher level of realism in implementing appropriate pedagogical theories, practices, and techniques for medical education, the better the learning process is continuously modified during the performance and even post-hoc. At least, that is what is expected of orthopedic physicians training in an extended reality (XR)/artificial intelligence (AI) and distance education (DE) environment. I.e., the media (technology) cannot influence learning. Learning is influenced by pedagogical methods used in the media (or by a live teacher with the same amount and type of learning).

Making things hard is the trying to incorporate the right way to learn better, and desirable difficulties of Bjork best demonstrate how surface learning often fails in the

long run. Desirable Difficulties or a considerable but desirable amount of effort in the initial learning period is a research-based principle for teaching and learning, leading to improved long-term performance [5, 6]. This difficulty level should involve a small investment of time and effort in spaced activities, calling for generation and self-testing through feedback to translate into significant learning gains identified as a desirable level of difficulty opening a window of learning opportunity [20]. It also looks like the information processing theory and the descriptive theory of skill acquisition align with the Challenge Point Framework (CPF) [21] and desirable difficulties [9]. I.e., challenges in a learning environment should match learners' abilities. The CPF, coined by Guadagnoli and Lee [14], helps in the effects of various practice conditions' conceptualization in motor learning by relating practice variables to the individual's skill level, task difficulty, and information theory concepts. Hence, any raises in task difficulty may considerably impact learning potential with a reverse impact on performance, i.e., maximizing learning while minimizing any harm to performance in practice determines the optimal challenge point (OCP).

It is believed that such a learning environment can create a model of success in which learning becomes inherently rewarding and enjoyable. However, "students may not engage in retrieval practice in their natural educational settings due to lacking metacognitive awareness of the mnemonic benefits of testing." Retrieval information practice through testing the latter positively impacts long-term retention and learning [19, 25].

In different terminology, Metcalfe's [20] Region of Proximal Learning, Hitchcock and Mcallister's [17] OCP, and both Hartwig and Dunlosky's [16] Zone of Proximal Development (ZPD) and Desirable Difficulties refer to a similar concept, that students engage in independent learning and practice, which should be pretty tricky, but at an acceptable level. Students are challenged enough to maintain their focus, and they can learn new concepts with guided assistance and scaffolding. Then, as the learning proceeds, the support structure is slowly removed. While Verenikina [30] refers to the ZPD as scaffolding's theoretical foundations, De Guerrero and Villamil [15] refer to scaffolding as ZPD's metaphor. ZPD and scaffolding also seem synonymous in that they both suggest learning with adult guidance or more capable peers' collaboration exceeds independent learning. The ZPD, or OCP, is a tool instructor can use to make complex decisions about when to give extra scaffolding.

Reiser [23, 24] points out that scaffolding is possible by structuring and problematizing and argues the interest in taking advantage of software tools to scaffold learners' complex tasks. These software tools can be used in structuring learning tasks while guiding learners through critical elements that support their planning and performance.

Since learning strategies creating desirable difficulties in the learning environment are most effective for learning outcomes' long run, Biwer et al. [4] demonstrated that these findings are reliable for practical students' training in higher education (i.e., positive effects on knowledge about effective learning strategies and increased practice testing use). Finally, beyond OCPs' potential maximizing for surgical training, Gofton and Regehr [13] also stress the irreplaceable value of a mentor.

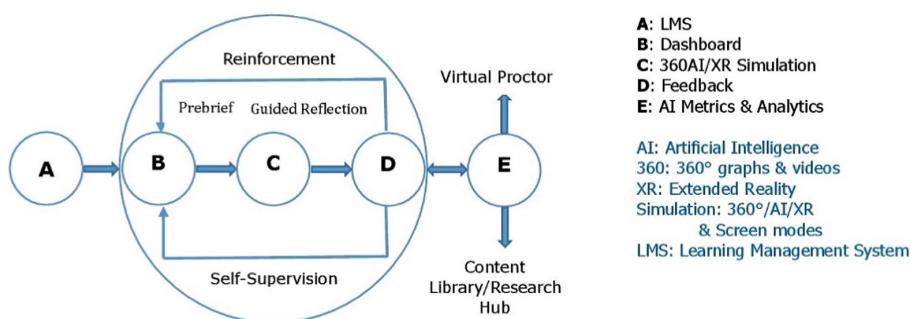
These approaches – i.e., OCP, desirable difficulties, scaffolding, and ZPD – seem related and should not be considered separately for their application in a medical or

surgical AI/XR environment where practicing is a must. Such an environment has a high cognitive load (CL) for learners unfamiliar with it. Software tools can be leveraged for that purpose, and the interactivity effect of CL theory elements can help determine when to achieve “Desirable Difficulties.” Hence, this paper aims to investigate how to integrate pedagogical aspects by combining AI, XR, and 360° videos to improve the learning curve in Orthopedic Surgery DE.

## 2 Background/Purpose

This study will explore the use of XR, AI, and 360° videos in orthopedic surgery DE, focusing on understanding the appropriate level of realism for different learning stages (1–5 categorization) and using deep learning (DL) algorithms for skill training. An XR/AI system for competency-based training and assessment in orthopedic surgery DE, with live instructors lead 360° Procedure training with prerecorded elements, programmatic skills modules using augmented reality (AR), and AI-driven models for visual performance assessment, will be developed.

As a potential further extension to the subsystem, a rule-based agent (i.e., CPF), with either tutor-specified or automatically learned rules, will be used to implement an ‘Adaptive Course Sequencing System (ACSS),’ such as Alzahrani et al. [2] model. Depending on students’ pedagogical capabilities and immediate needs, their intelligent pedagogical tutoring system can autonomously adjust its level of guidance; the instructor or students can do so too. The to-be-implemented ACSS will allow residents to grant the subsystem a level of autonomy in guiding them through cases of varying levels of difficulty and on different issues to assist them in becoming sufficiently competent in one subject before moving to another and suggesting appropriate sequences of points to optimize each resident’s learning experience.



**Fig. 1.** Learning Model

The focus is also on understanding how to use AI to provide feedback for reinforcement and assessment within the XR application and effectively communicate that feedback to users (see Fig. 1). In the trial, the intervention will be feedback to surgeons/medical students; 50% get feedback from technology and the other 50% from a human instructor.

### 3 Materials/Method

Interviews and user surveys, with follow-up interviews, were conducted via email or Zoom, and 22 pieces of feedback from experts from the industry and academia were gathered. A few interviewees provided detailed, discipline-centered answers (i.e., sims in surgery and medical education). Some also did not have the time or inclination to write/answer the immediate Ph.D. length response needed to answer these questions correctly. Some others probably gave helpful answers to about 25% of them to prioritize their current deadlines and paid work.

Based on the questions (see Appendices), user experience (UX) departments were contacted, especially for answers to human-computer interaction (HCI) questions related to AI and XR applied to surgical training. The other lines of inquiry touched on quite a few topics, including the application of AI (1) for pedagogical purposes, (2) for tracking a candidate's movements, procedures, etc., based on video observations, and (3) for assessment and feedback.

Twelve questions were more technical and back-end related, so data science or computer science experts were contacted for their perspectives. Nine out of the seventeen questions focused on medical and pedagogical aspects of the simulation, so experts from Educational Technology, Medical Education, and Orthopedic Surgery were sought for their perspective. Finally, industry experts and an international research organization were reached out to for synergistic expertise. (see Table 1).

**Table 1.** Summary of the number of experts per category.

Academia			International research organization (AI)	Industry (e.g., AI, XR, Robotic surgery, Games, etc.)
User Experience/UX	Data Science or Computer Science	Educational Technology, Sims in Surgery, Medical Education, & Ortho Surgery		
3	9	6	1	3

The need to use semi-structured user surveys/interviews with foundational questions was to propose insights improving XR medical education applications' design knowledge for non-Medical XR people. It was also unclear to which extent the XR application and AI would communicate with one another to integrate pedagogical aspects.

### 4 Practical Implications

The goal of developing an AI-guided XR tool for learning surgery (Orthopedic surgery) in DE is to understand how learning works in a medical education context. Most aspects of the learning trajectory are certainly similar for both an orthopedic and an appendix surgeon. The overall question is how to ensure the simulator training skills transfer to real life.

Procedural motor learning requires practice in a realistic environment. It also involves variation in the training and adaptability of the training difficulty to match the student's

progress and improve both long-term memory and transfer of learning. It is essential for motivation and efficiency of learning to keep challenge and ability in lockstep. Otherwise, people get bored or intimidated by the tasks being too challenging, both of which result in the individual losing motivation to complete the training.

Using the CPF adds variability in guidance that can be helpful to the learner. CPF simulations can be used to provide training examples so that the AI model can be trained. After training the AI model, the simulation may incorporate the model to show how well the AI model would work. Simulation is based on some formulas and assumptions, but actual operations from learners can be any type and in the wilderness/real-world (i.e., they are not controlled, and many unexpected operations/scenarios may occur).

OCPs should be determined for each level of the learner. Each surgical skill or task will have its own learning curve that will be individualized for every learner. A decision should be made about where the limits of simulation are. The setback is that, at some point, the OCPs will require actual patient encounters; learning in the simulation will be asymptotic and a waste of time. Fortunately, deliberate practice, when used in addition to mentorship, is an excellent method to improve learner performance while shortening the growth phase of the learning curve [7, 12].

#### 4.1 How Will the AI Determine a Failure or Successful Event?

In terms of how the AI will determine a failure or successful event, it may be helpful to define specific criteria or metrics that the AI can use to evaluate performance. These criteria or metrics could involve using data on the user's movements and actions within the simulation, as well as other relevant factors such as time taken to complete the task or the accuracy of the user's actions (e.g., through imitation by labeling images/videos and comparing "Ghost images" vs. "Real-world images," [11]). However, generally speaking, AI is only "intelligent" in finding patterns in data. With the right AI algorithm, good training data, and optimal parameters, the AI could effectively be trained to recognize a "failure" or "success" event by seeing enough examples of each event. I.e., DL will need to be trained on hundreds of hours of video labeled by experts. Good refers to quality, right relates to functionality, and optimal refers to a balance between different attributes, such as performance vs. safety.

The alternative would be that, rather than using a proper "AI," the designer/programmer could manually determine a set of rules or heuristics that a program can use to determine failure or success. AI systems are generally based on goal functions (also called utility functions or the results from "criteria or metrics"). To know how to react (no matter if the AI learns or reacts), the AI system must compare alternative "actions" or picks to make decisions. Even in some expert systems, where no goal functions seem to be implemented, there are goal functions in place, but they are binary (generate true or false values for each step). Most of the work could go into the design of such a function. Most code in a chess computer player is based on the chess rules, which is the foundation of how the goal function can assess each "game position."

The AI could then use this information to provide feedback to the user and assess their performance. It may also be helpful to consider how the AI can be integrated into the XR application in a way that is seamless and intuitive for users. Whether the trained AI evaluator eventually is to be built into the XR application would reduce friction (i.e.,

it would allow for a better user experience). Yes, integration could be advantageous if program complexity can be held at a minimal level. There is no need to integrate it into the XR application as long as it is connected to the internet. However, once the model is trained, it should be relatively light on resources, so building it in should be possible.

## 4.2 AI Integration into the XR Application

For the integration of the AI into the XR application, the considerations should weigh the AI's performance needs/power consumption, especially when running on low power or low battery capacities devices. Mobile phones/tablets or standalone Head-mounted displays (HMDs), and response times – both from the perspective of how fast the AI can compute on the target hardware, as well as any latency introduced by having the AI running off device (e.g., in the cloud). For low-resource countries, another consideration is internet access/reliability/speed. Though running this AI off-device/in the cloud (i.e., content library; see Fig. 1) in places where internet access is reliable and fast enough would not be advantageous, it would serve as a backup solution in low-resource countries.

Also, any hardware integration has its challenges. Using any XR with less than 60 Hz frame rate is not advisable because users will get nauseous. For the other applications, this determination will be based on the HCI and how close it is to the natural motion to which the skill will transfer. Holding a mouse or using a keyboard is a far-transfer skill for many of the skills taught unless they are cognitive or knowledge-based skills.

Why is XR HCI a critical consideration too? XR HCI is not natural to the inputs used in surgery. Although some surgical robots use Xbox-type controllers (Auris), most do not have any joystick similar to what real surgeons use. XR applications that use natural hand motion have not had excellent fidelity for inputs and again have the same issue with occlusion and instrument handling. Some folks are looking at 3D-printed controls that look similar to the real thing but have built-in sensors (accelerometer, gyroscopes, etc.). Many commercial entities in the orthopedic space using XR (e.g., OssoVR, Precision OS, etc.) have some data, but they have not yet been validated or transferred to surgical skill. XR has a high CL for learners unfamiliar with it and can actually detract from the learning.

How will the AI be integrated into the XR application to provide feedback? The guidance AI needs for the kind of feedback that will be provided is in terms of the values of the indicators established for data format or transfer process (i.e., the scales users are evaluated on). What is the point here? The guidance AI needs should be related to the type of categorization (e.g., Likert scale 1–5 or 1–10, etc.) for different learning stages.

The Likert scale is a simplification to allow for expert review. 1–10 does not allow for humans to really make a sound judgment and drops the inter and intra-rater reliability. AI does not have this problem. It will analyze the experts (with a tight, narrow definition) to create the algorithm. The economy of motion has been the most significant area researched that can distinguish between novices and experts. That is on a continuous variable and not Likert.

Given enough high-quality data and training time, a DL model should be able to learn an association between a video of a novice and a video of an expert. The question is really whether those ratings are satisfactory. With a clear definition of success states for each scale/level, the experience levels can be gated; i.e., users must succeed in the

first level (or scale) to go to the next one. In each part of the experience, a decision must be made on what constitutes success and if the user can repeat the level to achieve a higher score.

Users will be evaluated on a scale of 5 (skilled) to 1 (novice). However, scaling is dynamic according to needs, so any scaling method can be used depending on the AI system design. The same scale or different scales can be used. An attractive alternative could be to use raw data as input for the AI machinery to estimate probabilistic states [28]. A continuous variable would be used instead of a Likert one. However, if different scales are used, they must be rescaled to keep values consistent, i.e., the AI scale must be mapped with the one used in XR.

### **Realism and Sophisticated Behaviors**

Based on a Likert scale measurement, how much fidelity would be needed? The question of how much fidelity is needed is hard to answer because the measurement of performance or learning is pretty blunt (1–5 categorization). Many things are being learned: anatomy, clinical knowledge, motor skills, procedures, and problem-solving. These can all be measured more precisely than 1–5 categorization. Each stage has specific goals (say, instrument familiarization and location in the training device).

Moreover, there is no quick fix to design for anyone to capture and analyze sophisticated behaviors in real time for skill training. Learning is influenced by teaching methods provided in the media or by a live teacher with the same amount and type of learning. Teaching is complex, though, and no ‘best practice’ can be transferred from one situation to another. Expert teachers intuitively assess each learner’s current cognitive knowledge and affective state through engagement with the learner. How will this complexity be addressed? The app, through the ACSS, will automatically use the ZPD or OCP to make complex decisions about when to give extra scaffolding.

Given that teaching is complex, what useful strategies could then be used to design the application to capture and analyze sophisticated behaviors in real-time? To design the application to capture and analyze sophisticated behaviors in real-time, high-speed videography and motion capture technology may be considered to replicate a tactile experience with real haptic feedback.

The preceding reasoning could also involve using 3D assets and hardware that provide haptic feedback to the user, such as haptic gloves or other wearable devices. However, haptic feedback devices are very cool. Unfortunately, this technology is still in its infancy to date, and therefore, there are not yet any well-supported, commercially available haptic feedback devices that would work for this kind of app.

Moreover, AI models cannot currently learn these sophisticated and tacit processes. The only way forward is to develop AI-human interfaces that help teachers and/or learners with these complex interactions. AI will catch the behavior of the user and see if the behavior is at or close to the right move. DL part will also require a lot of labeled data with many example videos (perhaps hundreds of hours) labeled by experts.

There are commercial entities actually working on this aspect. With computer vision, movements can be tracked; the most significant issues obviously come with occlusion and the use of instruments. Some companies are working on overcoming these. The most successful applications (like Proximie) use video feeds from endoscopies and robotic surgery where only instruments are tracked, not the hands. The preceding is a precious

line of research, in any case. To allow anyone to capture, equipment and setting must be standardized. This is why laparoscopy and robotics are the first areas to look at this.

An attractive alternative would be to add educational games with stealth assessment and built-in learning dashboards for instructors to provide real-time feedback [27]. The continuous interaction with stealth assessment produces rich sequences of actions as data points that the stealth assessment captures in log files. Stealth assessment rubrics automatically score these aggregated data in real-time by Bayesian networks (or other statistical models). This process shows evolving mastery levels on targeted competencies [27].

### 4.3 AI-Human Interfaces or Educational Games' Best Pedagogical Practices

In determining the best pedagogical practices for these AI-human interfaces or educational games, the best practice is scaffolding the learner depending on their abilities and performance. The best application will keep them engaged at the edge of their abilities and increase the difficulty over time. The best formative will give them subtle feedback, and the experience guide them in what is right and wrong (similar to video games). Real-time feedback has to be immediate, succinct, actionable, and straightforward. Education around performing the task should be embedded in the experience. Summative performance data should delineate their skill level in separate domains (e.g., accuracy, efficiency, etc.).

In learning applications, pedagogical agents are usually used. I.e., some characters communicate with each other with the user. This agent will also provide feedback later. In some applications, for example, a teacher or scientist character guides the user. AI needs to learn from each domain of the skill the novice's delta from an expert. For example, they may be accurate but not efficient. This information needs to be communicated in some way to both the AI to evaluate and feedback to the user.

In terms of using AI to provide feedback for reinforcement and assessment, it would also be essential to consider what information is most relevant and valuable for the users to receive. It may also be that showing/not showing this information is more appropriate/valuable at different stages in the educational pipeline. This decision may depend on the stage of learning and the specific skills being trained [26]. However, formative feedback is best served immediately and summative after each experience or set of experiences.

For example, novice learners may benefit more from detailed explanations of how to perform a task, while more experienced learners may benefit more from feedback on their specific performance. It may also be helpful to consider different ways of communicating feedback to users, such as through visual or auditory cues within the simulation or more detailed reports or analyses provided later. Feedback should be multimodal (sounds, lights, vibrations, etc.) and nearly immediate. There are UX guidebooks on best practices. It should not distract from the experience but be very clear and succinct to the user, as in video games.

At first, it would be desirable to try to make it accessible later. If it is built within the simulation, then only with sufficient data can the AI model tell if anything is wrong on the fly – the goal is to correct behaviors in real-time. Likely all the information listed will

be needed, which means the context is leaning towards game design and user experience theory and concepts here—like a car racing simulator or Game, such as Forza 7.

Regarding the relevance of the Game's use, a few points must be considered carefully. Though serious games for medical education and surgical skills training can apply here, there are several concepts related to using games in medical learning. These concepts include:

1. The use of games in medical practice is still limited, and their use is only limited as a companion to the leading learning media. The understanding is that only AR and Virtual Reality (VR) expose medical models, and only a few have interactions (such as surgical practice).
2. Referring to the articles on JMIR Serious Games, the use of games is still limited to understanding the material, not the practice of surgical operations.
3. There are several VR applications with educational content on human anatomy, but there are still very few specifically related to surgery.
4. There are some medical games, but they are only for fun, not entirely educational games.
5. The experience must be embedded in an environment that reinforces skills through the gameplay without blindly putting on badges, scoring, and leaderboards. These should be meaningful to the experience.
6. Game-based stealth assessment seems valid, reliable, and improves learning [27].
7. For DL, medical expert systems are suitable, and Ripple-Down Rules (RDR) may be more relevant [10]. I.e., combining ML and RDR, particularly RDR systems that recognize a case outside their competence. RDR is easier to implement on insufficient data than AI.

#### **4.4 A Pedagogical Approach to Errors in a Simulated Environment**

Educators teach students about errors to help avoid them. Residents can make mistakes in the simulated environment since that is how they learn, so they should fail in a simulated environment rather than in real life. However, in general, the understanding of the research is that teaching about errors tends to increase the mistakes, not decrease them. In medicine, many (some say most) errors are due to missed condition/action situations. For example, forgetting to do something, and the second most frequent error is “doing something wrong,” missing steps in a procedure, or implementing the wrong approach for the conditions. In both cases, the error is sometimes due to the instructor skipping over a step or steps in a procedure/assessment that misses the student’s failure to remember the actions or process. Adequate instruction, practice, and corrective feedback should help avoid future errors. In some cases of coaxers, students become distracted and cause errors despite solid education.

#### **4.5 Multi-device Considerations When Designing/Developing the App**

When developing an app. to be used for multi-devices, then one consideration that should be taken into account is whether this app can be available for multiple devices or require XR devices. The only natural way to understand it is to actually trial it since these

devices have many pros and cons (See Appendices, Table 2). However, potential issues with any preferences can be factored in at the start of a process. I.e., eliminating specific approaches due to technical limitations, pedagogic goal matching, budget constraints, and assessing possibilities.

Here again, the cost is the most significant consideration. Also, any hardware integration has its challenges. Using any XR with less than 60 Hz frame rate is not advisable because users will get nauseous. For the other applications, this determination will be based on the HCI and how close it is to the natural motion to which the skill will transfer. Holding a mouse or using a keyboard is a far-transfer skill for many of the skills taught unless they are cognitive or knowledge-based skills.

Broadly processing/computational power differs across all the devices, and the delivery modality (i.e., User Interface or UI and UX) will (or at least should) differ depending on whether mobile, XR headset, desktop, etc. The Oculus Rift is a high-powered 3D headset, whereas a Google Cardboard uses phones, so its power will depend on the quality of the user's phone. VR headsets can provide 3D video, whereas smartphones, tablets, etc., will not.

Fortunately, real-time engines (e.g., game engines Unity or Unreal) provide XR, support multiple platforms, and can connect to MATLAB by socket connection. So, an educational experience can be deployed to multiple platforms if the app is developed correctly to ensure it runs on multiple platforms.

## 4.6 Data on Decisions and Actions Source

Event data is excellent depending on how the analytics are hooked up moment by moment. Still, a data plan is needed to clean, analyze, and visualize meaningful data. Secondary validation information like the Objective Structured Assessment of Technical Skill (OSAT) and expert reviews can help validate the AI assessments. I.e., the outputs will need to be compared to some established, validated form of data. XR and AI materials are all predicated on what is being done.

Moreover, AI is good at classification and regression. Generation based on such technologies as stable diffusion is also in fashion. However, it is bad at identifying what is not known as unknown. It even tells seemingly plausible lies. Simply, it is lax, which is AI's limitation today. When 100% guaranteed information is a requirement, it is too harsh for AI.

Reinforcement learning algorithms are probably appropriate when inputting users' scales to train AI systems is required. Getting information on a finer scale would be better, but if it is difficult to require that of humans, it would be advisable to use 1–5 scales.

## 4.7 Used AI Methodology/Realized Experimentation in Neurosurgery

The Force sensor (i.e., educational system) lays a basis for integrating AI and VR simulation into surgical education with an algorithm distinguishing between two groups of expertise, objective feedback based on proficiency benchmarks, and instructor input [22]. Its rich digital output can establish a robust skill assessment and sharing platform for orthopedic surgical performance and training. Beyond distinguishing between

groups of expertise, objective feedback based on proficiency benchmarks, and instructor input, the force sensor (i.e., using the support vector machine learning model) [22], SmartForceps data app [3], and tool-tissue forces in Newton (i.e., using deep neural networks) [1] can help improve training in orthopedic surgery and allow avoidance of these errors and resulting complications. The latest deep learning models (i.e., deep neural networks) can capture more parameters for performance metrics and perform more realistic simulations.

#### **4.8 Learning Experiences in Healthcare Return on Investment (ROI) for Education – Measure**

“The level of expertise of surgeons performing procedures should be reported” to determine whether the technique or surgeon’s poor execution – of the technique – led to poor results. Not only is comparing results between surgeons with significantly different levels of expertise invalid, but this critical information is also generally missing from published articles. [29] A well-defined expertise definition includes (1) Non-specialist, (2) Specialist – less experienced, (3) Specialist – experienced, (4) Specialist – highly experienced, and (5) Expert [29]. Whether expertise or performance is measured, performance is usually measured by OSATs. However, surgical procedures’ performance and improved learning curves are rarely measured [7].

The technology vs. human comparison is a ‘teaching method one vs. teaching method 2’ study of the kind Clark [8] discouraged many years ago – on the ground, such comparisons are confounded, always. In this case, the question is: how does the feedback differ between the two conditions (other than the delivery via machine vs. human)? If the feedback has the same information, frequency, and structure, there should be no difference in outcomes. If they are different in any of these regards, those differences presumably underlie the results. If, for example, the machine can provide more or more frequent feedback and/or practice, then this is perhaps the basis of the difference, not the circumstance of machine vs. human.

A suggestion to make the subsystem more “intelligent” would be by noting residents’ responses to individual cases in a way previously done for students learning computer programming [18]. Their system records every student interaction with it, providing a rich data set for analysis. Then, statistical pattern recognition and ML techniques automatically identify the “common mistakes” for each type of case. This improved subsystem should respond appropriately to a more comprehensive range of resident errors, including the possibility of mistakes that even an experienced tutor may not have anticipated.

### **5 Conclusion**

This qualitative interview is not a general survey, and these are complicated questions that require either substantial experience in this area of research or research investigation in their own right. Many of them are complex and not necessarily “solved” – many are thesis topics on their own that are probably better answered by someone with a background in education (i.e., educational technology or medical education).

This paper describes early-stage research that is setting out to investigate how AI, XR, and 360° videos should communicate to integrate the best surgical medical education approach from the perspective of an Instructional Systems Designer. However, the biggest question in this research is where do learners need to be on the learning curve before and after the XR experience. Can they start at zero (XR is their first experience with that surgery), or do they need some base knowledge? Different interactive elements of XR will support different levels (i.e., a learner just needing to know anatomy does not need a highly interactive, game-based experience to build foundational knowledge, but a more advanced learner will). Ultimately at which point is XR not enough for the learner? When do they need actual cases to advance their learning?

According to Wang [31], no perfect product allows the fusion of AR, AI, and machine translation technologies. So, the intention should not be to fuse the models/technologies. AI and XR may be used together collaboratively without being fused. The RDR can be combined with the CPF and stealth assessment with built-in learning dashboards to help residents avoid mistakes while improving long-term retention and transfer of learning.

Rules should be used to establish WHEN and HOW to deliver feedback to the learner. Different types and timings of feedback are applicable to different users. Finally, a straightforward way to design a goal function is to use the error between the resident's operation and the instructor's operation – because recorded operation steps and trajectories, and those from the residents are already available, they may be compared to get the errors. Other kinds of goal functions may also be used.

## Appendices

**Table 2.** Hardware's pros and cons.

Other Devices				XR Devices	
Smartphones/Tablets:		Desktop/Laptop:		HMDs (e.g., Oculus Rift, HTC Vive, or Google Cardboard):	
Pros:	Cons:	Pros:	Cons:	Pros:	Cons:
<ul style="list-style-type: none"> <li>• Widely accessible</li> <li>• Most people are familiar with them,</li> <li>• Lots of sensors, and</li> <li>• Interaction methodologies</li> </ul>	<ul style="list-style-type: none"> <li>• Low performance,</li> <li>• Low battery life,</li> <li>• The form factor is not conducive to certain types of interaction (e.g., having to hold the device in one hand while using the other hand to interact)</li> </ul>	<ul style="list-style-type: none"> <li>• Widely accessible,</li> <li>• Most people are familiar with them,</li> <li>• Best performance, battery life</li> </ul>	<ul style="list-style-type: none"> <li>• Fewer sensors and interaction methodologies,</li> <li>• Generally suited for 2D "flat surface" experiences using a keyboard and a mouse</li> </ul>	<ul style="list-style-type: none"> <li>• Immersive,</li> <li>• The best form factor for some experiences (e.g., both hands are free to interact, best support of natural hand interaction)</li> </ul>	<ul style="list-style-type: none"> <li>• Less accessible,</li> <li>• If untethered worst performance and battery life,</li> <li>• A novel technology for most people, which means they will perform worse but still prefer it.</li> </ul>

*Sample Questions (Interviews/Questionnaire).*

The questions are way too detailed and the application domain very specialized. A lot of this set of opening questions are early questions, which means that refinement or more context will be given to answer where possible.

1. How can I design for anyone to capture and analyze sophisticated behaviors in real-time in applying the field of high-speed videography of movement and its analysis by DL algorithms to skill training?
2. Data format or transfer process: Users are evaluated on a scale of 5 (skilled) to 1 (novice). Should that be the same scale used in values sent to the AI?
3. Would it be better for the trained AI evaluator eventually be built into the XR application?
4. What is the best practice regarding pedagogy here? I'm also not sure exactly what information I will have to show the users - is it information on how to perform the task, their performance in the task, or something else?
5. Would it be advisable that this information be shown to the users within the simulation, or would it be better if it is accessible later?
6. How will AI be used in application development to provide feedback for reinforcement and assessment?
7. What is the best pedagogical approach to providing feedback, and how best to communicate that feedback?
8. How will AI determine a failure or successful event since AI needs some guidance for the kind of feedback that will be provided?
9. Which "serious games" for medical education and surgical skills training would apply here, please?
10. What type of data do you think will be helpful for my study, and what collection method would you recommend? Additionally, can you outline the development process for producing XR/AI materials, please?
11. I want the app to be available for Smartphones, Tablet, PC / MAC (desktop/Laptop), moreover, will also require the use of VR devices such as a Head-Mounted Device (HMD) like the Oculus Rift, HTC Vive, or Google Cardboard. What could be the potential issues with my preferences, please?
12. Please outline the development process for producing materials using the Challenge Point Framework (CPF) (graph or diagram).
13. Or can you please share some practical applications of the CPF to simulation-based medical education and training?
14. For example: How can this framework be incorporated into a simulation?
15. What are some of its potential limitations, and how can we address them?
16. What 3D- assets, Motion capture technology, and hardware are there that could be implemented to replicate a tactile experience with real haptic feedback?
17. Using a technology-authoring environment, how can you mock-up or storyboard a prototype for a type of learning experience you may envision learners going through to answer a research question related to AR/VR, please? (i.e., an Example or Sample in any aspect of surgery to give me an idea of such a prototype)

## References

1. Albakr, A., Baghdadi, A., Singh, R., Lama, S., Sutherland, G.R.: Tool-tissue forces in hemangioblastoma surgery. *World Neurosurg.* **160**, e242–e249 (2022)
2. Alzahrani, A., Callaghan, V., Gardner, M.: Towards Adjustable Autonomy in Adaptive Course Sequencing. In: Intelligent Environments (Workshops), pp. 466–477, Jul 2013
3. Baghdadi, A., Lama, S., Singh, R., Hoshaymanesh, H., Razmi, M., Sutherland, G.R.: A data-driven performance dashboard for surgical dissection. *Sci. Rep.* **11**(1), 1–13 (2021)
4. Biwer, F., oude Egbrink, M.G., Aalten, P., de Bruin, A.B.: Fostering effective learning strategies in higher education—a mixed-methods study. *J. Appl. Res. Mem. Cogn.* **9**(2), 186–203 (2020)
5. Bjork, R.A.: Memory and metamemory considerations in the. *Metacogn.: Knowing about Knowing* **185**(7.2) (1994)
6. Bjork, E.L., Bjork, R.A.: Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. *Psychol. Real World* **2**, 59–68 (2011)
7. Cafarelli, L., El Amiri, L., Facca, S., Chakfé, N., Sapa, M.C., Liverneaux, P.: Anterior plating technique for distal radius: comparing performance after learning through naive versus deliberate practice. *Int. Orthop.* **46**(8), 1821–1829 (2022)
8. Clark, R.E.: Reconsidering research on learning from media. *Rev. Educ. Res.* **53**(4), 445–459 (1983)
9. Chen, O., Castro-Alonso, J.C., Paas, F., Sweller, J.: Undesirable difficulty effects in the learning of high-element interactivity materials. *Front. Psychol.* 1483 (2018)
10. Compton, P., Peters, L., Edwards, G., Lavers, T.G.: Experience with ripple-down rules. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp. 109–121. Springer, London (2006)
11. Chinthammit, W., et al.: Ghostman: augmented reality application for telerehabilitation and remote instruction of a novel motor skill. *BioMed Res. Int.* (2014)
12. Delbarre, M., Diaz, J.H., Xavier, F., Meyer, N., Sapa, M.C., Liverneaux, P.: Reduction in ionizing radiation exposure during minimally invasive anterior plate osteosynthesis of distal radius fracture: naive versus deliberate practice. *Hand Surg. Rehab.* **41**(2), 194–198 (2022)
13. Goffon, W., Regehr, G.: Factors in optimizing the learning environment for surgical training. *Clin. Orthopaed. Relat. Res.* **449**, 100–107 (2006)
14. Guadagnoli, M.A., Lee, T.D.: Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *J. Mot. Behav.* **36**(2), 212–224 (2004)
15. De Guerrero, M.C., Villamil, O.S.: Activating the ZPD: mutual scaffolding in L2 peer revision. *Mod. Lang. J.* **84**(1), 51–68 (2000)
16. Hartwig, M.K., Dunlosky, J.: Study strategies of college students: are self-testing and scheduling related to achievement? *Psychon. Bull. Rev.* **19**, 126–134 (2012)
17. Hitchcock, E.R., Mcallister Byun, T.: Enhancing generalisation in biofeedback intervention using the challenge point framework: a case study. *Clin. Linguist. Phon.* **29**(1), 59–75 (2015)
18. Hunter, G., Livingstone, D., Neve, P., Alsop, G.: Learn programming++: the design, implementation and deployment of an intelligent environment for the teaching and learning of computer programming. In: 2013 9th International Conference on Intelligent Environments, pp. 129–136. IEEE (2013)
19. Karpicke, J.D., Butler, A.C., Roediger, H.L., III.: Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory* **17**(4), 471–479 (2009)
20. Metcalfe, J.: Desirable difficulties and studying in the region of proximal learning. Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork, pp. 259–276 (2011)

21. Mills, B.: The role of simulation-based learning environments in preparing undergraduate health students for clinical practice (2016). Retrieved from <http://ro.ecu.edu.au/theses/1786>. Accessed 11 Oct 2020
22. Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., Del Maestro, R.F.: The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS ONE **15**(2), e0229596 (2020)
23. Reiser, B. J. (2002). Why scaffolding should sometimes make tasks more difficult for learners
24. Reiser, B.J.: Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *J. Learn. Sci.* **13**(3), 273–304 (2004)
25. Roediger, H.L., III., Karpicke, J.D.: The power of testing memory: basic research and implications for educational practice. *Perspect. Psychol. Sci.* **1**(3), 181–210 (2006)
26. Shute, V.J.: Focus on formative feedback. *Rev. Educ. Res.* **78**(1), 153–189 (2008)
27. Shute, V.J., Rahimi, S.: Stealth assessment. Rapid Community Report Series. Digital Promise and the International Society of the Learning Sciences (2022). <https://repository.isls.org/handle/1/7671>
28. Shute, V., Ke, F., Wang, L.: Assessment and adaptation in games. In: *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, pp. 59–78. Springer, Cham (2017)
29. Tang, J.B., Giddins, G.: Why and how to report surgeons' levels of expertise. *J. Hand Surg. (European Volume)* **41**(4), 365–366 (2016)
30. Verenikina, I.: Understanding scaffolding and the ZPD in educational research (2003)
31. Wang, Y.: Analysis of the combination of AR technology and translation system. In: *2021 International Conference on Social Development and Media Communication (SDMC 2021)*, pp. 180–184. Atlantis Press (2022)



# Adversarial Learning for Improved Patient Representations

Bharath Shankar and Carol Anne Hargreaves<sup>(✉)</sup>

National University of Singapore, 6 Science Drive 2, Singapore 117546, Singapore  
[bharath.shankar@u.nus.edu](mailto:bharath.shankar@u.nus.edu), [carol.hargreaves@nus.edu.sg](mailto:carol.hargreaves@nus.edu.sg)

**Abstract.** In recent years, there has been an explosion in the amount of patient Electronic Health Records (EHR) made publicly available. This presents an opportunity to create predictive models that leverage the large amount of data to help guide healthcare worker’s decision-making capacity. However, Patient EHR data is often high-dimensional, sparse, temporal and multimodal. This presents an issue for predictive modeling with Machine Learning models, as many ML model archetypes are not built to handle these types of data. The temporality of EHR data also presents a complicating factor, as it contains multiple time series at different resolutions, which most ML models are not built to handle. This brings us to the topic of patient representation, which is the process of converting this raw EHR data into a dense, mathematical representation. Previous work in this field, however, has not leveraged the full potential of the data, since they opt to only deal with a single modality of data, or do not leverage the temporality of the data. In this paper, we attempt to create a network that creates a multimodal representation of EHR data by modeling it as a multiple sparse time series fusion task. We leverage transformers for sparse time series data, using a custom time-based positional encoding. We then fuse the data into a low-dimensional vector, which serves as our representation. We train the model on 2 separate tasks – Mortality prediction using a classification head, and we attempt to leverage a form of adversarial learning to improve the quality of the representation. We show that the patient representation extracted is meaningful and useful for downstream classification tasks. To our knowledge, this is the first attempt to leverage both adversarial learning and multimodality to create a patient representation. Source code can be found at: <https://github.com/BharathSShankar/Patient-Rep-UROPS>.

## 1 Introduction

There has been an increase in the amount of publicly available Electronic Health Record data over the years. This increase has been instrumental in many recent advances in Healthcare Information Systems, where healthcare workers have their decisions supported by large amounts of easily accessible data [1]. A key result of this increase in publicly available data has been a rise in attempts to use Artificial Intelligence to aid healthcare workers in their decision-making process [2]. More specifically, deep learning-based approaches have shown their utility in this area [3].

There has been a great deal of work done in the field of Healthcare and AI [4]. Deep-Learning based approaches have yielded great success in the domains of ECG signal

classification [5], EEG signal interpretation [6], medical image segmentation [7], among many other sub-domains in healthcare [4]. All of these advances in machine learning and AI were driven by advances in Deep Learning Architectures built to handle modalities of data such as images and text, increased compute capacity due to the advent of GPUs (and more recently, TPUs), as well as the aforementioned increase in publicly available medical data.

### 1.1 Need for Patient Representation

However, EHR data presents a unique challenge for machine-learning models. EHR data often is high-dimensional, temporal, sparse, multimodal, heterogeneous, filled with null values, and often is extremely large scale. This presents issues for machine learning algorithms of all types.

Thus, we need some way to condense all the information in the EHR into a dense, mathematical representation of the data. That way, downstream predictive models can use that dense representation of the data, instead of the raw patient data. This problem is known as the problem of patient representation [8].

A good patient representation should ideally be usable for downstream predictive tasks, and the representation should be meaningful. A meaningful representation implies that similar patients should have similar representations. Ideally, we would also want each individual feature to be semantically meaningful – i.e. – it corresponds to a real-world concept.

This leads to why patient representation matters to healthcare workers. Trust is an important factor in the adoption of any system in any industry, especially the healthcare industry. Explainability, i.e., being able to explain why a certain decision was made by a Machine Learning model in terms understandable to healthcare workers is key in encouraging the adoption of Machine Learning systems, as well as to help eliminate any biases or errors that may exist in the model.

Unfortunately, the explainability of any model is inversely proportional to its complexity. This gets even worse when dealing with deep learning models, which are often considered to be black boxes. A semantically meaningful patient representation offers us a way to both leverage the unparalleled feature extraction capacity of neural nets, as well as being able to offer some level of explainability for any output of the model, thus offering the best of both worlds.

### 1.2 Past Work

Past work in the field of patient representation has leveraged many different forms of neural nets, and varied mathematical objects to embody it. The first studies in this area used artificial neural networks to extract patient representations from data as seen in Che et al. [9]. However, basic Artificial Neural Networks do not handle temporality, sequential data, nor unstructured data very well. This then necessitated a change in approach. We then saw autoencoder-based models, which consist of an encoder-decoder pair come into use for patient representation. Miotto et al. [10] used a stacked autoencoder (Multiple autoencoders stacked on each other) to come up with a general patient representation.

However, these autoencoder-based architectures still suffered from a lack of feature extraction. Concepts of locality or temporality aren't encoded into the features of an ANN, which necessitates modifications to the architectures. This brings us to the advent of CNN-based architectures. CNNs use convolution operations to create features from neighboring points in the data. Stacking these convolutional layers then create higher level features, and so on and so forth. Thus, CNNs have shown significant improvements over artificial neural networks in the realm of patient representation in particular, due to their ability to consider neighboring events as being associated with each other [11]. Modifications to CNNs, to explicitly accommodate temporality in the data have also been shown to improve the performance of patient representation.

Clinical Notes are an essential component of patient data. They record key information that cannot otherwise be recorded. Thus, text analytics is another key front in the search for effective patient representation. Word2Vec approaches for word embeddings was a key innovation in the field of text representation. Word2Vec showed its efficacy in creating low-dimensional embeddings, as shown by Choi et al. in 2016 [12].

Recently, the advent of Transformer-based architectures, such as BERT, we have the possibility of dealing with sequence-based data extremely effectively. Transformers work by splitting up the data into tokens. It then creates a position embedding (usually based on the index of the datapoint in question), adds it elementwise to each of the tokens, and passes it through the attention mechanism. As seen above, the fact that each token is treated in relation to all the other tokens makes transformers a promising avenue for patient representation. Starting with Vaswani et al. [13], transformers have consistently been at the forefront of tasks ranging from generating text embeddings, as seen in BERT [14], Computer Vision (utilizing vision transformers [15]), among many, many other domains. Needless to say, transformers have shown significant progress in the realm of patient representation learning.

Transformer-Based architectures have been used to cluster related concepts from clinical test data in the 2020 paper by Li et al. [16]. They have also shown great strides in dealing with the sparse time series data that is a key feature in clinical data, as evidenced in Song et al. [17]. We also have seen transformers being able to learn the underlying graphical nature of Electronic Health Records, using Graph Convolutional Transformers [18].

A key challenge of dealing with Patient data is leveraging the multimodality of the data. Multimodal data fusion is a problem that crops up in multiple disciplines, from image-text matching [19], to video classification [20]. Needless to say, there have been various studies attempting to fuse different modalities of patient data for both patient representation and other downstream tasks. Multimodal fusion techniques have been applied to fields ranging from cancer diagnosis [21], survival prediction [22], and even for diagnosis of mental illnesses [23].

Adversarial Learning has been shown to improve the quality of extracted representations [24]. The idea is that forcing a model to effectively spoof random data, while simultaneously attempting to maximize classification accuracy would provide more consistent representations of each patient. Patient representation has been shown to improve by the application of adversarial learning, as evidenced by De Bois et al. [25], as well as by Zhang et al. [26]. However, to our knowledge, there has not been an attempt to use

both adversarial learning and multimodal fusion on patient data, which is the gap that the following paper attempts to mediate.

## 2 Methods

### 2.1 Data

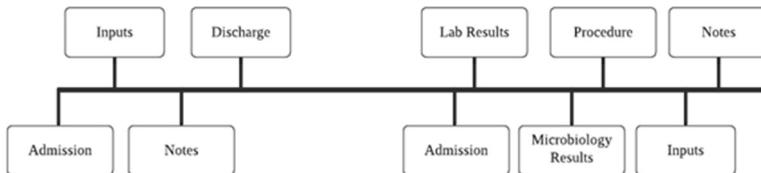
#### Source

MIMIC-III is a major clinical dataset for patient EMR [27]. It contains data from over 40,000 patients at the Beth Israel Deaconess Medical Centre in Boston, USA, collected over the course of 10 years, from 2001 to 2012. MIMIC-III has been used in a variety of studies in this field, and the data as a whole is well-documented and understood, making it an invaluable benchmark for the task of patient representation.

### 2.2 Preprocessing

#### A Patient as a Sparse Time Series

Data in the MIMIC dataset comes in the form of events at specified times. Thus, a natural way to represent a patient to feed into the model is as a sparse time series. However, as part of the anonymisation process, every observation for each patient has been shifted by a consistent and arbitrary amount. Given that we know each patient's date of birth, we can then calculate the age at which each event occurred. Thus, we can use age as a consistent timestamp (Fig. 1).



**Fig. 1.** A representation of a patient as a sparse time series.

### 2.3 Model

#### Time2Vec

When examining the original transformer architecture, we see that we need to provide each input token vector with a positional encoding vector [14]. In the original paper on transformers, the positional encoding was derived from a function of the index. However, in our case, that positional encoding can no longer use the indices, due to the irregularly spaced nature of the time series. Thus, we need a way to convert the timestamp of the

observation to a position embedding. The key insight of Time2Vec [28] is recognizing that the positional embedding is just that: a representation of the position of the token with respect to others. Hence, the timestamp is used to calculate the embedding as seen below:

$$v_t(x) = \sin(Wx + b)$$

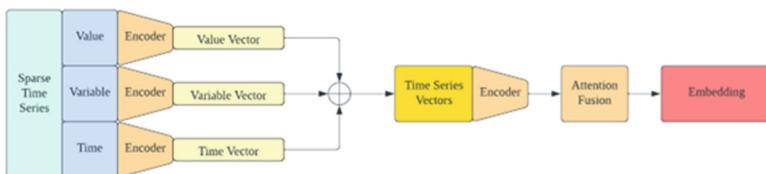
where  $W$  and  $b$  are learnable parameters,  $x$  is the timestamp and  $v_t$  is the vector representation of the timestamp.

### STraTS

Out of all 7 modalities of data considered, 4 of them were of the form (variable, value, timestamp). We needed a way to consider both the value and variable being measured at each time stamp.

Thus, we used a variant of STraTS [29] to encode the time series. STraTS is a method of using transformers to encode time series of differing lengths. STraTS encodes (variable, value, timestamp) triplets as 3 separate vectors by passing a numerical representation of each through a linear unit. We then add each of these 3 vectors element wise, which becomes the initial triplet embedding. We then pass these embeddings into multiple attention layers, followed by feed forward and norm layers. These allow for each of the individual embedding vectors to be considered in the context of the other embedding vectors, thus recovering multiple contextual embedding vectors. Finally, we fuse the embedding vectors using an attention fusion network, to recover a final embedding that represents the sparse time series as a whole.

For the attention fusion layer, we consider vectors  $v_1, v_2, \dots, v_n$  as being the vector representation of each time point in the sparse time series. To get the weights  $\alpha_1, \alpha_2, \dots, \alpha_n$  for each of the vectors, we pass them through a soft-maxed feedforward layer. The final fused vector is then  $v_{emb} = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n$  (Fig. 2).



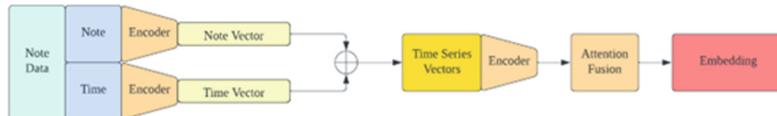
**Fig. 2.** A Description of a STraTS encoder

For the microbiology data, we have 2 categorical variables representing the specimen and organism available, and a response interpretation variable. This, along with the timestamp now makes up a quartet embedding. We use a slightly modified version of the Triplet Embedding alongside the transformer encoder layers and attention fusion layers to create embeddings for the microbiology data.

### Text Data

We use a sentence embedding model (using the pre-trained Albert [30] model from

huggingface) to create embedding vectors for each clinical note made for each patient. It is then added elementwise to the embedding of the time data, to create a full embedding for each note event. Then, it is passed through an attention fusion layer to create a single embedding representing all notes taken of the patient (Fig. 3).

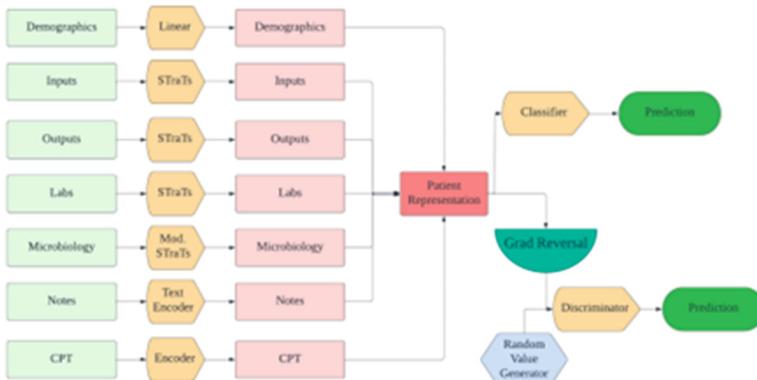


**Fig. 3.** Text Data Encoder

### Adversarial Learning

For the adversarial portion of the network, we used a simple MLP as a discriminator. We then randomly roll for an integer, either 1 or 0. This decides if we send in an actual representation of the patient data, or if we send in a random sequence instead. The discriminator is appropriately penalized, and we reverse the gradients applied to the discriminator for our encoder layer. In other words, we penalize our encoder layers when our discriminator performs well, and vice versa. This is accomplished by adding a gradient reversal layer just prior to the discriminator. The gradient reversal layer acts like an identity layer on the forward pass, but multiplies all incoming gradients during backpropagation by  $-1$ . We used the pytorch implementation of the same by Ganin et al. [31].

### Overall Architecture



**Fig. 4.** A Full Model Architecture

### 2.4 Training

The model was written using the Pytorch-Lightning library, and tested on google colab. We used encoders with 4 attention units, with a hidden dimension of 512. We trained the

model on Nvidia Tesla® V100-32GB [32] GPUs, made available by NUS HPC [33]. The loss function used was Binary Cross Entropy, which is standard for classification tasks. We also tried out adding different weights to our cross-entropy loss to our discriminator. The higher the weight, the more the model prioritizes spreading out data points over classification accuracy. We aim to strike a balance between the two objectives, to optimize both the usage and accuracy of the patient representation. Thus, our loss  $L_{\text{net}}$  can be represented as (Fig. 4)

$$L_{\text{net}} = L_{\text{class}} + \beta L_{\text{disc}}$$

where  $\beta$  is the weight given to the discriminator  $L_{\text{disc}}$  over the classification loss  $L_{\text{class}}$ . We trained each model with the Adam optimiser, with learning rate 10-5 and weight decay of 10-3. We used weight decay to add a constraint on the L2 norm of the weights of the model. In total, 3 models were trained at values of  $\beta$ : 0, 0.2 and 2. We employed a random train-test-validation split of 40,000: 4,000: 2505 patients. We employed the ASHA [34] (Async Successive Halting) scheduler for early stopping and used Ray [35] for hyperparameter tuning.

## 3 Results & Discussion

### 3.1 Classification Task

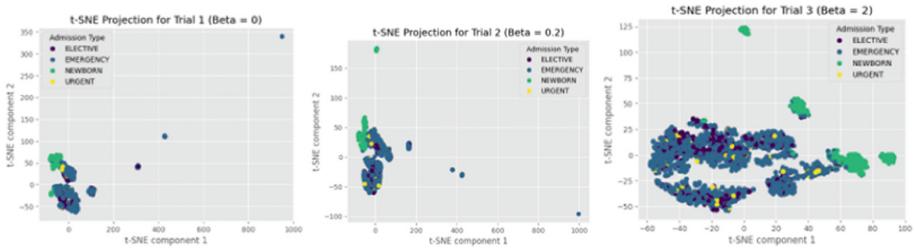
Beta Value	F1 Score	Precision	Recall	AUC-ROC
0	0.642	0.513	0.846	0.8198
0.2	0.6	0.444	0.923	0.8186
2	0.569	0.569	0.569	0.7766

We see that as the weight on the adversarial task increases, the prediction accuracy drops. This is to be expected, as the tendency for classifiers is to maximize the distance between various types of data. However, as mentioned earlier, this tendency causes the representation as a whole to suffer. This is because it becomes easier for the representation to place points in arbitrary locations (effectively overfitting for the task).

### 3.2 Patient Representation

We now examine the representations created by the model. The data used to create the above plots were the test set of MIMIC data. Since the resulting representations are 120-dimensional, we use t-SNE, a non-linear dimensionality reduction algorithm to reduce the data down to 2 dimensions for the purposes of visualization (parameters: perplexity = 40, PCA initialization). We can see that as we increase the weight of the discriminator loss, the latent space is more efficiently used by the representation (Fig. 5).

To assess the meaningfulness of the representation, the data were colored by the hospital admission type. This was not a feature that was fed to the model to create the representation. Since certain categories of admission are extremely different to others (Newborns vs. all others), observing natural clustering based on this criteria would imply that the representation has “understood” some facet of the patient. As seen in the third image above, the newborn patients are clearly separated from the rest of the patients.



**Fig. 5.** Patient representations visualized via t-SNE

## 4 Future Work

Another avenue for future work is to try to use another dataset, such as the eICU synthetic dataset [36] or MIMIC-IV [37] dataset. Also, running more hyperparameter tuning would also be an avenue to improve the performance of the model. Additionally, we can use fusion via mutual information [38] maximization to further improve the quality of the extracted patient representation. Also, adding a reconstruction task, as shown in the paper by Mai et al. [39], can aid in improving the quality of the extracted representations. Additionally, we can attempt to use attention map visualizations to help in improving the explainability of the model. Finally, testing the performance of the representations on other downstream tasks such as length-of-stay prediction is also required.

## 5 Conclusion

To sum up, in this project, we trained a model that successfully took into account multiple modalities of patient data to predict mortality. Additionally, we showed that leveraging adversarial learning for the patient representation leads to better quality representations. We also showed that the representations are meaningful, as evidenced by the representations naturally separating different types of patients along the reason why the patient was admitted to the hospital, which was not a feature used in the data to train the model. Thus, we have successfully created a method for multimodal patient representations, aided by adversarial learning.

## References

1. Holroyd-Leduc, J.M., Lorenzetti, D., Straus, S.E., Sykes, L., Quan, H.: The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *J. Am. Med. Inform. Assoc.* **18**(6), 732–737 (2011). <https://doi.org/10.1136/amiajnl-2010-000019>
2. Jiang, F., et al.: Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* **2**(4), 230–243 (2017). <https://doi.org/10.1136/svn-2017-000101>
3. Norgeot, B., Glicksberg, B.S., Butte, A.J.: A call for deep-learning healthcare. *Nat. Med.* **25**(1), 14–15 (2019). <https://doi.org/10.1038/s41591-018-0320-3>

4. Alloghani, M., Baker, T., Al-Jumeily, D., Hussain, A., Mustafina, J., Aljaaf, A.J.: Prospects of machine and deep learning in analysis of vital signs for the improvement of healthcare services. In: Yang, X.-S., He, X.-S. (eds.) *Nature-Inspired Computation in Data Mining and Machine Learning*, pp. 113–136. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-28553-1\\_6](https://doi.org/10.1007/978-3-030-28553-1_6)
5. Che, C., Zhang, P., Zhu, M., Qu, Y., Jin, B.: Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Med. Inform. Decis. Mak.* **21**(1), 184 (2021). <https://doi.org/10.1186/s12911-021-01546-2>
6. Hosseini, M.-P., Tran, T.X., Pompili, D., Elisevich, K., Soltanian-Zadeh, H.: Deep learning with edge computing for localization of epileptogenicity using multimodal rs-fMRI and EEG big data. In: 2017 IEEE International Conference on Autonomic Computing (ICAC), pp. 83–92 (2017). <https://doi.org/10.1109/ICAC.2017.41>
7. Lai, M.: Deep Learning for Medical Image Segmentation (2015). <http://arxiv.org/abs/1505.02000>
8. Si, Y., et al.: Deep representation learning of patient data from Electronic Health Records (EHR): a systematic review. *J. Biomed. Inform.* **115**, 103671 (2021). <https://doi.org/10.1016/j.jbi.2020.103671>
9. Che, Z., Kale, D., Li, W., Bahadori, M.T., Liu, Y.: Deep computational phenotyping. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 507–516 (2015). <https://doi.org/10.1145/2783258.2783365>
10. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**(1), Article 1 (2016). <https://doi.org/10.1038/srep26094>
11. Cheng, Y., Wang, F., Zhang, P., Hu, J.: Risk prediction with electronic health records: a deep learning approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining (SDM), pp. 432–440. Society for Industrial and Applied Mathematics (2016). <https://doi.org/10.1137/1.9781611974348.49>
12. Choi, Y., Chiu, C.Y.-I., Sontag, D.: Learning low-dimensional representations of medical concepts. In: AMIA Summits on Translational Science Proceedings, pp. 41–50 (2016)
13. [1706.03762] Attention Is All You Need (n.d.). <https://arxiv.org/abs/1706.03762>. Retrieved 5 November 2022
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). <https://doi.org/10.48550/arXiv.1810.04805>
15. Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). <https://doi.org/10.48550/arXiv.2010.11929>
16. Li, Y., et al.: BEHRT: transformer for electronic health records. *Sci. Rep.* **10**(1), Article 1 (2020). <https://doi.org/10.1038/s41598-020-62922-y>
17. Song, H., Rajan, D., Thiagarajan, J., Spanias, A.: Attend and diagnose: clinical time series analysis using attention models. In: Proceedings of the AAAI Conference on Artificial Intelligence, **32**(1), Article 1 (2018). <https://doi.org/10.1609/aaai.v32i1.11635>
18. Choi, E., et al.: Learning the graphical structure of electronic health records with graph convolutional transformer. *Proc. AAAI Conf. Artif. Intell.* **34**(1), 606–613 (2020). <https://doi.org/10.1609/aaai.v34i01.5400>
19. Srivastava, N., Salakhutdinov, R.: Learning Representations for Multimodal Data with Deep Belief Nets **8** (n.d.)
20. Jiang, Y.-G., Wu, Z., Tang, J., Li, Z., Xue, X., Chang, S.-F.: Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Trans. Multimedia* **20**(11), 3137–3147 (2018). <https://doi.org/10.1109/TMM.2018.2823900>

21. Cheerla, A., Gevaert, O.: Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**(14), i446–i454 (2019). <https://doi.org/10.1093/bioinformatics/btz342>
22. Vale-Silva, L.A., Rohr, K.: Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **11**(1), Article 1 (2021). <https://doi.org/10.1038/s41598-021-92799-4>
23. Meng, Y., Speier, W., Ong, M.K., Arnold, C.W.: Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inform.* **25**(8), 3121–3129 (2021). <https://doi.org/10.1109/JBHI.2021.3063721>
24. Mai, S., Hu, H., Xing, S.: Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion (2020). <https://doi.org/10.48550/arXiv.1911.07848>
25. De Bois, M., El Yacoubi, M.A., Ammi, M.: Adversarial multi-source transfer learning in healthcare: application to glucose prediction for diabetic people. *Comput. Methods Programs Biomed.* **199**, 105874 (2021). <https://doi.org/10.1016/j.cmpb.2020.105874>
26. Zhang, X., et al.: Learning robust patient representations from multi-modal electronic health records: a supervised deep learning approach. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pp. 585–593 (2020). <https://doi.org/10.1137/1.978161976700.66>
27. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**(1), Article 1 (2016). <https://doi.org/10.1038/sdata.2016.35>
28. Kazemi, S.M., et al.: Time2Vec: Learning a Vector Representation of Time (2019). <https://doi.org/10.48550/arXiv.1907.05321>
29. Tipirneni, S., Reddy, C.K.: Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series (2022). <https://doi.org/10.48550/arXiv.2107.14293>
30. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (2022). <https://doi.org/10.48550/arXiv.1909.11942>
31. Ganin, Y., Lempitsky, V.: Unsupervised Domain Adaptation by Backpropagation (2015). <http://arxiv.org/abs/1409.7495>
32. NVIDIA Tesla V100|NVIDIA (n.d.): <https://www.nvidia.com/en-gb/data-center/tesla-v100/>. Retrieved 21 November 2022
33. HPC – NUS Information Technology (n.d.): <https://nusit.nus.edu.sg/hpc/>. Retrieved November 21, 2022
34. Li, L., et al.: A System for Massively Parallel Hyperparameter Tuning (2020). <https://doi.org/10.48550/arXiv.1810.05934>
35. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A Research Platform for Distributed Model Selection and Training (2018). <http://arxiv.org/abs/1807.05118>
36. Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**(1), Article 1 (2018). <https://doi.org/10.1038/sdata.2018.178>
37. Alistair, J., Lucas, B., Tom, P., Steven, H., Leo Anthony, C., Roger, M.: MIMIC-IV (2.1). PhysioNet (n.d.). <https://doi.org/10.13026/RRGF-XW32>
38. Multimodal Data Fusion Based on Mutual Information. IEEE Journals & Magazine. *IEEE Xplore* (n.d.). <https://ieeexplore.ieee.org/document/6095545>. Retrieved November 21, 2022
39. Mai, S., Hu, H., Xing, S.: Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion (2019). <http://arxiv.org/abs/1911.07848>

# **Augmented Intelligence in Games, Serious Games and Virtual Reality**



# User-Defined Hand Gesture Interface to Improve User Experience of Learning American Sign Language

Jindi Wang<sup>1</sup>(✉) , Ioannis Ivrissimtzis<sup>1</sup> , Zhaoxing Li<sup>1</sup> , Yunzhan Zhou<sup>1</sup> ,  
and Lei Shi<sup>2</sup>

<sup>1</sup> Department of Computer Science, Durham University, Durham, UK  
`{jindi.wang,ioannis.ivrissimtzis,zhaoxing.li2,yunzhan.zhou}@durham.ac.uk`  
<sup>2</sup> Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK  
`lei.shi@newcastle.ac.uk`

**Abstract.** Sign language can make possible effective communication between hearing and deaf-mute people. Despite years of extensive pedagogical research, learning sign language remains a formidable task, with the majority of the current systems relying extensively on online learning resources, presuming that users would regularly access them; yet, this approach can feel monotonous and repetitious. Recently, gamification has been proposed as a solution to the problem, however, the research focus is on game design, rather than user experience design. In this work, we present a system for user-defined interaction for learning static American Sign Language (ASL), supporting gesture recognition for user experience design, and enabling users to actively learn through involvement with user-defined gestures, rather than just passively absorbing knowledge. Early findings from a questionnaire-based survey show that users are more motivated to learn static ASL through user-defined interactions.

**Keywords:** Human Computer Interaction · Sign Language · User Study

## 1 Introduction

According to the World Health Organization, around 2.5 billion people will have some degree of hearing loss by 2050<sup>1</sup>, and at least 700 million of them will require some kind of hearing rehabilitation. The use of sign language, as well as several other alternative sensory approaches, such as voice reading, writing with the hands, or vibration sensing, are all part of the rehabilitation training courses for people with hearing loss. Although sign language is the most popular means of communication for the deaf, most persons who do not have hearing loss have never taken sign language classes, making communication between these two groups difficult. Thus, in an effort to remove communication barriers between

<sup>1</sup> <https://www.who.int/zh/news-room/fact-sheets/detail/deafness-and-hearing-loss>.

various groups, learning sign language has emerged as a major research area in education.

The majority of the most recent approaches to the teaching of sign languages [9, 10, 16] employ predefined gestures, while approaches based on user-defined interaction technology are more sparse in the literature. For example, Adamo *et al.* [1] proposed the development of a new immersive 3D learning environment to increase the mathematical skills of deaf children by teaching mathematical concepts and American Sign Language (ASL) math terminology through user interaction with fantasy 3D virtual signers and traditional interaction environments. Schioppo *et al.* [20] proposed a sign language recognition method using features extracted from data acquired by a Leap Motion controller from an egocentric view. The method was tested on the 26 letters of the ASL alphabet. In a related development, Phan *et al.* [17] used motion tracking to trial a number of different methods for providing user feedback in a sign language learning system.

Regarding research on the processes by which users can define themselves a vocabulary of hand gestures, Piumsomboon *et al.* [18] conducted research on hand gesture guessability in an Augmented Reality (AR) environment. They invited users to make gestures corresponding to certain tasks, and created user-defined gesture sets to guide the designers in implementing user-centred hand gestures for AR. To the best of our knowledge, there are no studies on learning ASL through user-defined interaction techniques. Hence, the purpose of this paper is to investigate if user-defined interaction techniques can enhance users' learning of ASL. We believe that this is an important research topic since most such systems use hand gestures created by system designers, which do not always reflect user intention.

To accomplish our research goals, we developed a simple system with a user-defined gesture interface for learning static ASL. In the system design, we have identified and taken into account shortcomings of prior systems, including the small data sets used to train the gesture recognizer, the absence of a realistic environment, and most importantly, the user's difficulty in engaging with the system for an extended period of time. With inspiration from Bragg's ASL Sea Battle [5], a sign language game created to help gather user data, we created and integrated a Whack-a-Mole style game with a user-defined hand gesture interface into the system, aiming at boosting user motivation. Finally, we conducted a user study based on a survey designed according to Schrepp's [21] recommendations and concentrated on user experience analysis.

Summarising, the main research question motivating our work, "***Can user-defined interaction techniques enhance user motivation to learn static ASL?***", was looked into within the context of a gamified environment for learning static ASL. Our main contributions are as follows:

1. We implemented a user-defined hand gesture interface for ASL learning with a Whack-a-Mole type of game.
2. We conducted a user study to examine if user-defined interaction affected users' experience. The initial results indicate a positive user attitude towards gamified learning environments and a strong interest of the users in user-defined interactions.

The rest of this paper is organised as follows. The prior work on technology-assisted sign language learning is reviewed in Sect. 2. The basic design and the features of the proposed sign language learning environment are presented in Sect. 3. The design of the user study is described in Sect. 4, while the results are presented and analysed in Sect. 5. We discuss the main findings in Sect. 6 and briefly conclude in Sect. 7.

## 2 Related Work

The back-end of the proposed system supporting ASL learning with user-defined interaction mainly consists of hand gesture detection and a recognition model. Hence, we review prior research on hand gesture detection and recognition in ASL and user interfaces for creating user-defined hand gestures.

### 2.1 Sign Language Detection and Recognition

Real-time detection of dynamic hand gestures from video streams is a challenging task since: (i) there is no indication when a hand gesture starts and ends in the video; (ii) a performed hand gesture should only be recognized once; and (iii) the entire system should be designed considering memory and computational power constraints. Bheda *et al.* [3] proposed a method based on deep convolutional neural networks (CNNs) to recognize images of the letters and digits in ASL. Kim *et al.* [13] proposed a novel sign language recognition method, which employs an object detection network for a region of interest segmentation to preprocess the input data. Battistoni *et al.* [2] described a method for ASL alphabet recognition based on CNNs, which allows for monitoring the users' learning progress. Jiang *et al.* [12] proposed a novel fingerspelling identification method for Chinese Sign Language via AlexNet-based transfer learning and evaluated four different methods of transfer learning. Camgoz *et al.* [6] introduced a novel transformer-based architecture that jointly learns Continuous Sign Language Recognition and Translation while being trainable in an end-to-end manner. Zhang *et al.* [25] proposed MediaPipe Hands, a real-time on-device hand tracking pipeline to compute hand landmark positions from a single RGB camera frame for AR/VR applications. Goswami *et al.* [11] created a new dataset for ASL recognition and used it to train a CNN-based model for hand gesture recognition and classification. Finally, Pallavi *et al.* [15] presented a deep learning model based on the YOLOv3 architecture, reporting high recognition rates on the ASL alphabet.

Having reviewed the existing work on sign language recognition, we concluded that Mediapipe is the most suitable tool for the purposes of this paper, and thus, we used it for sign language recognition, benefiting from its highly accurate, real-time detection of hand landmark points. Moreover, as an open-source hand gesture detection framework from Google, it is well-documented and supported.

## 2.2 User Interfaces for User-defined Hand Gesture

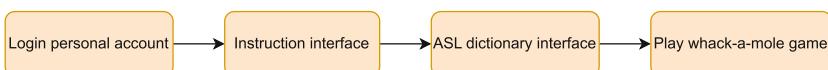
A lot of work has already been done on user-defined hand gesture user interfaces, but most of them support limited functionalities, such as letting the user select one out of two established hand gestures as the one they want to use. For example, Wu *et al.* [24] proposed an interface for users to customize hand gestures and apply them to VR shopping applications in 2019, while they [23] proposed a user-defined hand gesture interface that could be used on in-vehicle information systems in 2020. Besides, conventional means of accessing visual communication markers (VCM) rely on input entry methods that are not directly and intimately tied to expressive nonverbal cues. Koh *et al.* [14] addressed this issue, by facilitating the use of an alternative form of VCM entry: hand gestures. Moreover, to fill this gap Dai *et al.* [7] presented a training system, called CAPG-MYO, for user-defined hand gesture interaction. Takayama *et al.* [22] perform two user studies to derive a user-defined gesture set allowing 13 types of table manipulation.

To address the issue that pre-defined gestures do not completely reflect user intent, we evaluated earlier work on user-defined gestures. As a result of these studies, we were also motivated to consider whether the addition of user-defined gesture interaction will reduce sign language learners' weariness and boost their motivation for learning sign language.

Therefore, the primary goal of our research is to investigate if user-defined gesture interaction affects ASL learning. In order to give users an immersive experience, we developed a VR-based system. To stimulate users' curiosity and boost their motivation, we also included a simple game with a user-defined hand gesture function. In addition, because there is a lack of user research on the subject, we used a questionnaire to survey users to investigate whether customised gesture interactions can actually inspire more people to learn sign language. Our main objective was to critically assess our system and gather user feedback on how their interaction with our system affected their learning experience.

## 3 System Components

This section provides an overview of the key components of the proposed system. The system's recommended workflow is shown in Fig. 1. When the user enters their account information through the login interface, the system initialises their location to the Instruction interface. After users familiarize themselves with the user introduction information of the Instruction interface, they study the ASL dictionary for five minutes and then visit the sign language game interface to play the game and increase their understanding of sign language through it.



**Fig. 1.** The workflow of ASL learning system.

### 3.1 Learning Environment

The learning tools and the game for learning the numbers 0 to 9 in ASL are displayed in Fig. 2, with the user's viewpoint tilted 45° to the left. The entire scene was created in Unity (2020.3.32f1). Regarding the user's engagement with the system, we used the eye-tracking functionality of the HTC Vive Pro and enable clicking or picking an object after 3 s of the user's fixed attention. An inbuilt camera that was connected to the PC via openCV (version 3.4.2) [4] was used to acquire the images. Regarding gesture detection and recognition, Mediapipe is used to detect the user's hand and extract a series of 21 points matching corresponding landmarks on the detected hand. The feature vector from this sequence is then supplied as input to the classifier, which is an MLP with three fully connected layers, implemented in Tensorflow 2.6.0 [8] and Python 3.6 [19]. We used an RTX3080 GPU on a standard PC to train the classifier. The study's objectives were satisfied with an overall recognition accuracy rate of over 90%, which is expected to offer a generally positive user experience.

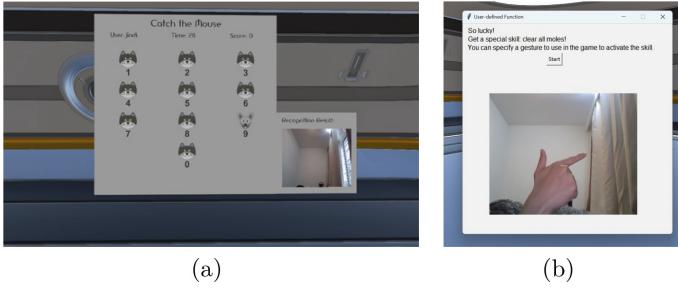


**Fig. 2.** The implemented ASL learning environment. **From left to right:** the Whack-a-Mole game; the ASL dictionary for the 0–9 digits; the Instructions interface.

### 3.2 Whack-a-Mole Game and User-Defined Interface

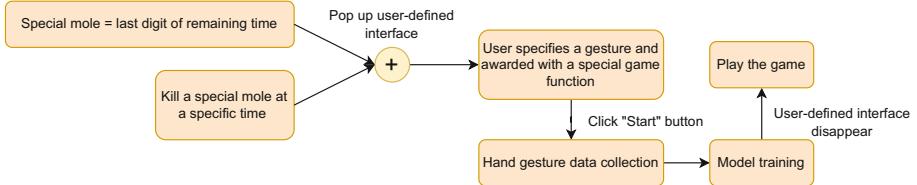
We adopted the Whack-a-Mole game and implemented a sign language-based version of it, aiming to make learners more interested in the material, increasing their motivation and, eventually, their engagement with the process. In our game, as shown in Fig. 3(a), each location is marked by a unique numeric identifier. If the user signs correctly the current position of the gopher, one point is added; otherwise, no point is awarded. The total duration of the game is 30 s.

The user-defined interface is a feature that is hidden from the user while they are playing the game. It will only appear when a mole is killed and trigger a hitherto hidden functionality, calling for the user to specify a wake-up gesture to be utilized later in the game. In the example shown in Fig. 3(b), the user is given the special game skill to “Clear all Moles”. At the end of one iteration of the course, the system will collect user-defined gesture data for 5 s, retrain the recognition model, and the user-defined interface will be hidden again. Now that



**Fig. 3.** (a) The Whack-a-Mole game for ASL learning; (b) the user defined interface.

the user has picked their special skills, they may start playing the game again by clicking “Start” on the game screen, and they can use their newly acquired special game skill. The workflow of the user-defined hand gesture interface is shown in Fig. 4. Notice that the special mole sequence must match the last digit of the remaining play time for the user interface to be activated. In addition, the user must be able to recognise the special mole’s number at a specific moment. When both requirements are satisfied, the user-defined interface will be displayed and the user will be awarded their special game skill.



**Fig. 4.** The workflow of the user-defined hand gesture interface.

## 4 Experiments

To evaluate the system design, we adopted the user survey scheme proposed by Schrepp *et al.* [21], which comprises six evaluation factors: **Attractiveness**, **Efficiency**, **Perspicuity**, **Dependability**, **Stimulation**, **Novelty**. Each factor is further divided into six or seven more specialised issues. Table 1 displays the specific issues associated with each factor. Based on the users’ scores on a scale of 1.00 to 5.00 on particular issues, we assessed the merits of the system in each factor.

We invited 15 users ( $M = 8$ ,  $F = 7$ ; aged between 19 and 21) to engage with our system, aiming at gathering user feedback to serve as the study’s data source. The majority of the users had either very limited or no prior understanding of ASL, or any other sign language. They were instructed to explore the system, adhering to the instructions in order to learn ASL in three stages: learning

signs from the dictionary interface; improving comprehension at the learning interface; and assessing their learning three times on the game interface. As it can be challenging for some beginners to pick up so many motions quickly, users were merely required to learn the ASL 0–9 numerals.

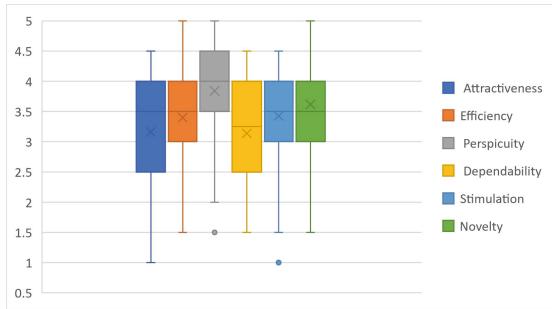
**Table 1.** System evaluation questionnaire.

Attractiveness	Efficiency
Do users like or dislike the product?	Is it easy to understand how to use the product?
annoying/enjoyable	Is it easy to get familiar with the product?
good/bad	not understandable/understandable
unlikable/pleasing	easy to learn/difficult to learn
unpleasant/pleasant	complicated/easy
attractive/unattractive	clear/confusing
afriendly/unfriendly	
Perspicuity	Dependability
Is it possible to use the product fast and efficient?	Does the user feel in control of the interaction?
Does the user interface looks organized?	Is the interaction with the product secure and predictable?
fast/slow	unpredictable/predictable
inefficient/efficient	obstructive/supportive
impractical/practical	secure/not secure
organized/cluttered	meets expectations/does not meet expectations
Stimulation	Novelty
Is it interesting and exciting to use the product?	Is the design of the product innovative and creative?
Does the user feel motivated to further use the product?	Does the product grab user's attention?
valuable/inferior	creative/dull
boring/exiting	inventive/conventional
not interesting/interesting	usual/leading edge
motivating/demotivating	conservative/innovative

## 5 Result Analysis

The user evaluation is summarized in Fig. 5, the box-plots showing the Minimum, First Quartile, Median, Third Quartile, and Maximum, while the Mean is shown by an ‘x’. The score distribution reflects generally positive feedback on the evaluation factors, all of which received mean scores greater than 3.00, while the overall system achieved a satisfactory average score of 3.42 ( $SD = 0.88$ ) over the six factors. It is also interesting to note that some low scores ( $<2.50$ ) were given in all factors, the possible causes of which are discussed below.

**Attractiveness:** as shown in Fig. 6(a), the average score over the 7 questions is 3.16 ( $SD = 0.90$ ). Each question has some scores lower than 2.50, possibly reflecting some lack of interaction with the users. For example, some users reported that animations should show up when the hand gesture was recognised correctly. Warnings should also be shown if no hand was detected, or when the hands were too close to the camera. Future improvement plans include the addition of more interactive features, such as moving backgrounds, scene changes, and animations.



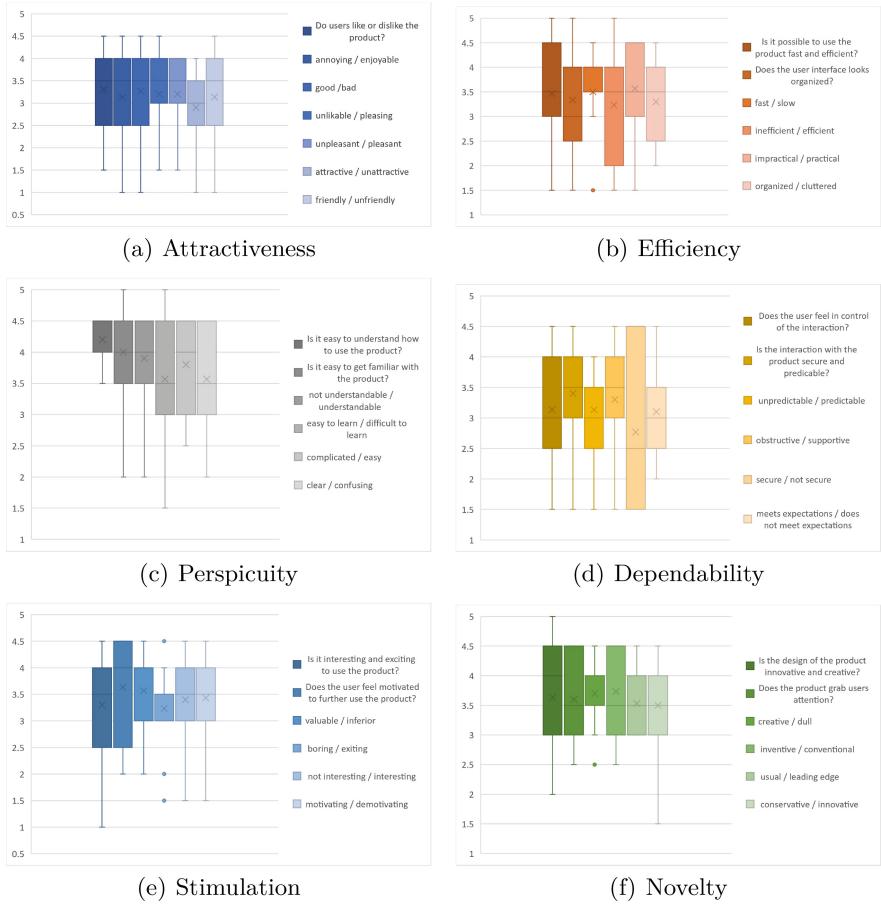
**Fig. 5.** Box-plots of the score values for each of the user survey's factors.

**Efficiency:** as shown in Fig. 6(b), the average score is 3.40 ( $SD = 0.93$ ). There were 3 users who gave a score of 5.00 on some questions. However, nearly one-third of them gave scores lower than 2.50 on each question, indicating that there is still room for increasing the system's efficiency. According to user feedback, the practice interface was not so convenient to use, as the users had to click the button “Start” to check for correctness. Improvements could be made to automate this process, thus requiring less activity of this type from the users.

**Perspicuity:** as shown in Fig. 6(c), the average score of 3.84 ( $SD = 0.77$ ) is the highest among the six factors, indicating that most users saw the system as easy to use, perhaps because of the simple design of the interface, which made it easy to use. Nevertheless, one user complained about the scene navigation system, overlooking apparently the navigation instructions button of the main menu. Thus, when this particular user tried to move to the ASL 0–9 dictionary, they did not know how to do it until we told them to look at the button. In a future study, the users will first be taught how to navigate the scene, before going into the main study of gestures.

**Dependability:** as shown in Fig. 6(d), the average score is 3.14 ( $SD = 0.90$ ). Although most users gave scores higher than 3.50, a small number of users gave scores of 1.50. Perhaps this was because, in the practice interface, some users thought that they did the correct gesture but were judged as being wrong. The reason behind this can be performance issues of the gesture recognition model in the back-end, perhaps because the training data for some gestures might have been of poor quality, thus leading to low recognition accuracy rates for these gestures. Future improvements will aim at training recognition models that will be able to better generalize to natural physiological differences in users' hands.

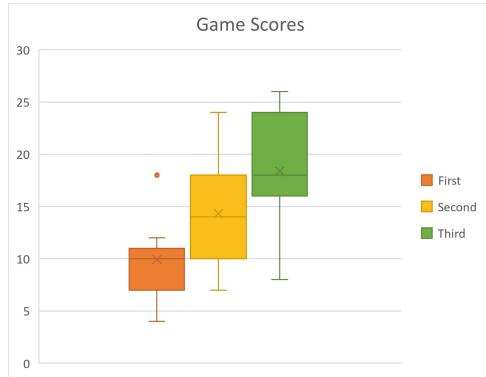
**Stimulation:** as shown in Fig. 6(e), the average score for this factor is 3.43 ( $SD = 0.85$ ), while the mean value of every question under this factor is above 3.25. It is also worthwhile to note that the first quartiles of all questions start from just below 3.00, and most of the users gave scores between 3.50 and 4.00. Nevertheless, there were also some low scores in all questions, showing that there is still room for improvement. In particular, the interaction of the system could be designed more creatively, aiming at better inspiring the users.



**Fig. 6.** Box-plots of the scores for each subdivision of the six factors.

**Novelty:** as shown in Fig. 6(f), the average score of this factor is 3.62 ( $SD = 0.70$ ), suggesting a broadly positive reception, with the mean score in all questions at 3.50 or above. The second and the fourth questions have relatively tight distributions, with interquartile range between 3.00 and 4.50. Overall, the results on this factor suggest that most of the users regarded the system as being innovative, with only a few of them perceiving it differently.

Figure 7 shows that users typically had poor game scores at their initial attempt, with the exception of one outlier with a score of 18, who had the good fortune to activate the user-defined interface and obtain the hidden game skill on their first attempt. We note that the average, lowest, and highest user score all gradually increased at the second and third attempt, showing that the user's sign language proficiency increased. Additionally, several users claimed that this hidden feature might stimulate their interest in the game, and, implicitly, help them advance their sign language skills.



**Fig. 7.** Box-plots of the game scores for each of the three attempts.

## 6 Discussion

Using VR technology, we developed an immersive environment for learning ASL. We looked into whether a user-defined way of interaction could boost users' motivation to learn sign language. We evaluated this issue using the survey method proposed by Schrepp *et al.* [21].

For the user evaluation of the system we used six assessment factors. The survey's results on these six metrics demonstrate how well the user-defined interface for ASL learning operates and that it can genuinely satisfy user needs. Besides, because user-defined interactions are more experiential, most users seem to accept them, according to the analysis of environmental experience. However, the system still needs to be further optimised and adjusted for some functions to improve user experience, even though the majority of users are happy with user-defined interactions.

### 6.1 Limitations

Our work still has a number of limitations in terms of system design and implementation, as was already mentioned. Here, we summarise them for each assessment factor separately. **Attractiveness:** when sign language was correctly interpreted, some users complained that there weren't any animated clues. Animations in suggestions could enhance user experience. **Efficiency:** because users have to manually touch the start button to play the sign language game, some users complained that the user interface wasn't sufficiently automatic. **Perspicuity:** some users have complained that they were unsure of how to navigate the scenario. **Dependability:** while playing the game, some users claimed to have used the correct sign language, but the algorithm determined that they had not, giving them a lesser score. **Stimulation:** a small percentage of users gave low scores on the stimulation factor study, which suggests that there is still room for our system to be designed more creatively. **Novelty:** a small proportion of users

felt that the system wasn't inventive enough, possibly because they thought the learning model was too simple.

On the other hand, our study has a number of methodological drawbacks. The user study and included only 15 participants. The invited people were between the ages of 19 and 21; there is no research on users in other age groups. To further test our methodology, we intend to enlist more individuals in future studies, who should come from a wider range of backgrounds (e.g., age). Ultimately, there is no specific evaluation indicator in the questionnaire survey on sign language acquisition. A uniform and standardised questionnaire-based assessment of sign language acquisition is needed for the next research.

## 7 Conclusion

A virtual environment that allows users to learn ASL through the use of user-defined hand gestures has been developed by our team. The user interface that is embedded in the virtual environment made it possible to most of the users to readily comprehend the workflow of the system, as well as each stage of the ASL learning process. The results of a user questionnaire that we carried out ( $N = 15$ ) revealed that participants were, in general, pleased with the digital ASL learning system that we developed. In conclusion, the overall results provide credence to our original hypothesis, which stated that an increase in users' motivation to learn can be attributed to the usage of user-defined interaction modalities.

In the future, we will include in the system more interactive components, such as backdrop movement, scene changes, and animation prompts. In order to reduce the amount of human involvement required to control the system, we will also add more automatic settings. To help the user understand how to manipulate the objects in the scene, a follow-through user interface will be developed. In addition, a stronger gesture recognition model will be developed, enabling the inclusion of more sophisticated sign language instruction materials.

## References

1. Adamo-Villani, N., Carpenter, E., Arns, L.: An immersive virtual environment for learning sign language mathematics. In: ACM SIGGRAPH 2006 Educators Program, p. 20-es (2006)
2. Battistoni, P., Di Gregorio, M., Sebillo, M., Vitiello, G.: AI at the edge for sign language learning support. In: IEEE HCC Conference, pp. 16–23 (2019)
3. Bheda, V., Radpour, D.: Using deep convolutional networks for gesture recognition in American sign language. arXiv preprint [arXiv:1710.06836](https://arxiv.org/abs/1710.06836) (2017)
4. Bradski, G., Kaehler, A.: OpenCV. DDJ Softw. Tools **3**, 120 (2000)
5. Bragg, D., Caselli, N., Gallagher, J.W., Goldberg, M., Oka, C.J., Thies, W.: ASL sea battle: gamifying sign language data collection. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2021)
6. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10023–10033 (2020)

7. Dai, Q., Li, X., Geng, W., Jin, W., Liang, X.: CAPG-MYO: a muscle-computer interface supporting user-defined gesture recognition. In: Proceedings of the 9th ICCCM, pp. 52–58 (2021)
8. Dillon, J.V., et al.: TensorFlow distributions. arXiv preprint [arXiv:1711.10604](https://arxiv.org/abs/1711.10604) (2017)
9. Empe, N.A.A., Echon, R.C.L., Vega, H.D.A., Paterno, P.L.C., Jamis, M.N., Yabut, E.R.: SimboWika: a mobile and web application to learn Filipino sign language for deaf students in elementary schools. In: R10-HTC, pp. 1–6. IEEE (2020)
10. Estrada-Cota, I., Carreño-León, M.A., Sandoval-Bringas, J.A., Leyva-Carrillo, A.A., Quiroz, H.X.C.: Design of a web tool for teaching-learning of states and capitals of México through the Mexican sign language. In: International Conference on Inclusive Technology and Education (CONTIE), pp. 174–179. IEEE (2021)
11. Goswami, T., Javaji, S.R.: CNN model for American sign language recognition. In: Kumar, A., Mozar, S. (eds.) ICCCE 2020. LNEE, vol. 698, pp. 55–61. Springer, Singapore (2021). [https://doi.org/10.1007/978-981-15-7961-5\\_6](https://doi.org/10.1007/978-981-15-7961-5_6)
12. Jiang, X., Hu, B., Chandra Satapathy, S., Wang, S.H., Zhang, Y.D.: Fingerspelling identification for Chinese sign language via AlexNet-based transfer learning and Adam optimizer. Sci. Program. (2020)
13. Kim, S., Ji, Y., Lee, K.B.: An effective sign language learning with object detection based ROI segmentation. In: 2018 Second IEEE IRC, pp. 330–333. IEEE (2018)
14. Koh, J.I., Cherian, J., Taele, P., Hammond, T.: Developing a hand gesture recognition system for mapping symbolic hand gestures to analogous emojis in computer-mediated communication. ACM TiiS **9**(1), 1–35 (2019)
15. Pallavi, P., Sarvamangala, D.: Recognition of sign language using deep neural network. IJARCS **12**, 92–97 (2021)
16. Patricks, A.: Developing an accessible learning application for sign language (c) (2022)
17. Phan, H.D., Ellis, K., Dorin, A., Olivier, P.: Feedback strategies for embodied agents to enhance sign language vocabulary learning. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, pp. 1–8 (2020)
18. Piomsomboon, T., Clark, A., Billinghurst, M., Cockburn, A.: User-defined gestures for augmented reality. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013. LNCS, vol. 8118, pp. 282–299. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-40480-1\\_18](https://doi.org/10.1007/978-3-642-40480-1_18)
19. Python, W.: Python. Python Releases for Windows **24** (2021)
20. Schioppo, J., Meyer, Z., Fabiano, D., Canavan, S.: Sign language recognition: learning American sign language in a virtual environment. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2019)
21. Schrepp, M., Hinderks, A., Thomaschewski, J.: Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In: Marcus, A. (ed.) DUXU 2014. LNCS, vol. 8517, pp. 383–392. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-07668-3\\_37](https://doi.org/10.1007/978-3-319-07668-3_37)
22. Takayama, Y., Ichikawa, Y., Shizuki, B., Kawaguchi, I., Takahashi, S.: A user-based mid-air hand gesture set for spreadsheets. In: Asian CHI Symposium, pp. 122–128 (2021)
23. Wu, H., Wang, Y., Liu, J., Qiu, J., Zhang, X.L.: User-defined gesture interaction for in-vehicle information systems. Multimed. Tools Appl. **79**(1), 263–288 (2020)
24. Wu, H., Wang, Y., Qiu, J., Liu, J., Zhang, X.: User-defined gesture interaction for immersive VR shopping applications. BIT **38**(7), 726–741 (2019)
25. Zhang, F., et al.: MediaPipe hands: on-device real-time hand tracking. arXiv preprint [arXiv:2006.10214](https://arxiv.org/abs/2006.10214) (2020)



# Distraction Detection and Monitoring Using Eye Tracking in Virtual Reality

Mahdi Zarour, Hamdi Ben Abdessalem<sup>(✉)</sup>, and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal,  
Montréal H3C 3J7, Canada

{mahdi.zarour, hamdi.ben.abdessalem}@umontreal.ca,  
frasson@iro.umontreal.ca

**Abstract.** Effective learning is highly affected by attention levels. Hence, Intelligent Tutoring Systems and other technologies for learning should be able to monitor the attention levels of learners and detect distractions in real-time to improve the learning process. We study the feasibility of detecting and monitoring visual distraction of participants, while they complete cognitive tasks, using Eye Tracking in a Virtual Reality environment. We also investigate the possibility of improving the attention of participants using relaxation in Virtual Reality. The Eye Tracking distraction model we developed correctly predicts the distraction state of participants with an F1-score of 86%. We also found that the most appropriate window size to detect distraction ranges from three to six seconds. Furthermore, results suggest that our relaxation method significantly decreased the visual distraction of the participants.

**Keywords:** Eye Tracking · Virtual Reality · Distraction · Attention · Human Interaction

## 1 Introduction

### 1.1 Distraction and Intelligent Tutoring Systems

**Attention**, **concentration**, and **distraction** are interrelated cognitive processes that determine our ability to focus and efficiently complete tasks. Attention serves as the basis for concentration, which is the ability to sustain attention on a particular task or thought. Distraction on the other hand, diverts attention away from the primary task, and disrupts our ability to concentrate and can lead to reduced productivity and effectiveness [1].

Intelligent tutoring systems (ITS) are educational technologies that provide personalized instruction and feedback to students are designed to adapt to individual learners' needs, providing personalized feedback and scaffolding to optimize learning outcomes [2].

Studies have shown that attention directly influences learning outcomes, with students who are better able to concentrate achieving higher academic performance [3].

Research has also shown that distractions negatively impact learning outcomes, as they divert attentional resources from the learning process and impair information processing [4]. Thus, one critical aspect to improve ITSs' effectiveness is attention management, which involves monitoring and guiding learners' attention and minimize distractions to ensure they remain focused on the learning task [5].

## 1.2 Virtual Reality

Virtual Reality (VR) is an advanced technology that simulates environments realistically. It lies in the intersection of many fields including electronic engineering, simulation, and computer graphics [6]. Many VR headsets come equipped with Eye Tracking technology nowadays.

While numerous experiments were conducted in real-world settings, little was done in Virtual Reality. Consumer-grade Eye Tracking devices constrain the user to always look ahead, and devices that enable the user to freely rotate the head are high-priced. With the actual VR technology, multiple consumer-grade VR devices are equipped with Eye Tracking technology like high-priced Eye Tracking devices. Moreover, Research indicated higher learning performance and engagement in VR compared to classic methods [7, 8], suggesting that current and new learning methods and platforms could target this technology soon including ITSs. These results suggest that VR could help to advance research on attention, by providing simulation environments that generalize to real-life conditions.

## 1.3 Eye Tracking

Lee et al. developed a system to monitor concentration level of learners in real-time by analyzing pupillary response and eye blinking patterns using a simple commercial eye tracker, and web camera [9]. A machine learning model was first trained to discriminate between the “concentrated” and “not-concentrated” states. Later, the predictions of the model were averaged over periods of one second and used to make a real-time concentration monitoring system. Although the system performed reasonably well with one second periods, other values were not investigated and could significantly improve the performance of the proposed system.

Hutt et al. studied the feasibility of integrating cheap eye trackers into ITSs to monitor the attention of learners using an extensive list of eye movements features [10]. The developed machine learning model achieved an F1 score of 59% in a classification problem with two states: attentive, and mind wandering (MW) in a participant-independent setting. They also used Cohen's d values to rank features by their contribution. Even though the model performance was higher than the chance level, there is still room for improvement and a distraction detection system with higher accuracy would be more efficient.

These results suggest Eye tracking could be used to effectively detect and monitor the distraction of individuals.

## 1.4 Relaxation and Attention Restoration

Attention Restoration Theory (ART) is a psychological framework suggesting that exposure to natural environments facilitates the restoration of voluntary attention capacity, reducing mental fatigue and enhancing cognitive functioning [11]. Gao et al. assessed ART with VR and EEG and showed that the experience had positive restorative effects on the individuals' attentional fatigue and negative mood [12].

In [13] authors used relaxation in Virtual Reality to study the possibility of decreasing negative emotions in the elderly including frustration, anxiety, and apathy. The preliminary results showed a decrease in anxiety and frustration, an increase in memorization performance, and an improvement in cognitive abilities, particularly in **attention** exercises.

Together, these results suggest it is possible to improve the performance of individuals on tasks, by restoring their attention and reducing their mental fatigue, using relaxation. In addition to monitoring and detecting distractions, providing ITS with a tool to improve the attention of learners would lead to more effective learning.

While Eye Tracking has previously seen extensive use in attention research, it was not the case for Virtual Reality. An easy way to study attention is to use Virtual Reality environments as they provide capabilities to easily manipulate the levels of attention experimentally.

The purpose of this study is to develop a means to detect and monitor **distraction**, and to develop methods to improve the **attention** of participants (**by reducing their distraction**) while completing cognitive tasks in a Virtual Reality environment using Eye Tracking technology. We put forward two hypotheses in this study: 1) it is possible to detect and monitor the visual distraction levels of participants using Eye Tracking in VR while completing cognitive tasks. 2) The **visual distraction** of participants will decrease after relaxation compared to before the relaxation period.

## 2 Experimentation

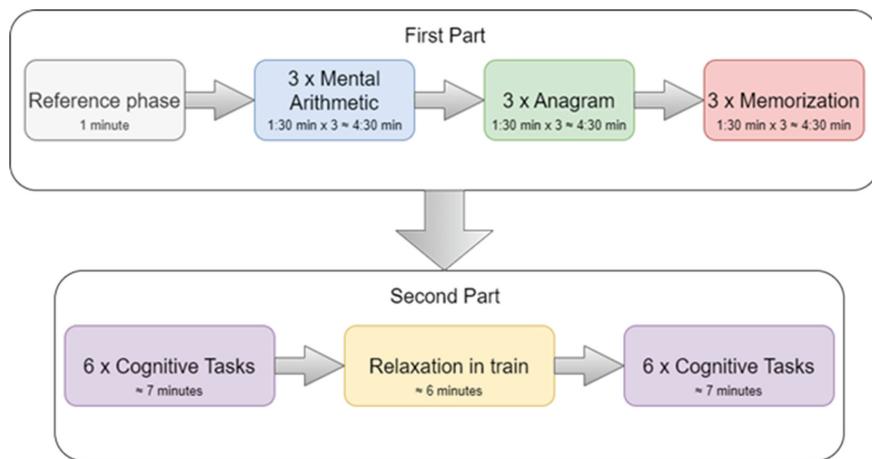
We built an experiment in VR to detect distraction and obtain data from Eye Tracking headset. The experiment was divided into two parts. The first part of the experiment was used to develop a means to detect and monitor the levels of visual distraction using Eye Tracking, while the second part of the experiment was used to investigate the effects of relaxation on the visual distraction levels of participants.

### 2.1 Hardware

For this experiment, we used an HTC VIVE Pro Eye VR headset with one right controller. The VR headset came with an integrated Eye Tracking system that could track eyes at a maximal frequency of 120Hz (90Hz in our experiment) and accuracy between 0.5° and 1.1°. All participants used the same hardware to complete the experiment.

## 2.2 Experiment

We implemented the experiment using Unity, a popular game development engine that uses the C Sharp programming language. The experiment took place in a room in Virtual Reality and consisted of two parts. Figure 1 shows an overview of the entire experiment and details of each part.



**Fig. 1.** Overview of the experiment

All participants completed the same tasks with the same values. The duration of the experiment was 40 min for each participant.

To detect the distraction, we used Eye Tracking data obtained from an HTC SRanipal SDK and included gaze origin as a three-dimensional point, gaze direction as a three-dimensional vector, pupil position as a two-dimensional vector indicating the position in sensor area, pupil size in millimeters as a real number, eye openness as a real number. Eye Tracking data were recorded manually in CSV file using obtained data and times in Unix timestamps format from Unity.

**In the first part** of the experiment, we used mental tasks to trigger the concentration of participants. However, to induce **visual distraction** state, we introduced **distractors** and **hints** during the experiment in order to divert participants' **attention** away from the main task towards other objects in the environment. Advice was given to participants to look left or right for a supposed hint. The hints had two objectives. The first objective was to keep the participants distracted away from the primary task. The second objective was to help participants find the correct answer in case they were not able to. The correct answer was always hidden randomly among hints, and we informed the participants of that, but we instructed them to try and solve the problems without help.

Participants performed three types of **cognitive tasks** a total of nine times. In mental arithmetic, they had to solve an arithmetic problem of addition and subtraction with five or six operands mentally. In the anagram task, participants had to rearrange a set of eight or nine letters to obtain a valid word in English. In the memorization task, participants

had to memorize a sequence of seven or eight digits in the reverse order, the digits were presented one by one with a one-second between-digit interval. Mental arithmetic task and anagram task were previously used to study attention [14, 15]. The memorization task was not used in attention research, but previous research suggests that attention is linked with memory [16], which makes it an interesting option.

In the **second part** of the experiment, we used the environment developed in [13, 17] to study the feasibility of improving concentration and decreasing distraction levels using **relaxation in Virtual Reality**. Participants were invited to travel in a virtual train which has proven to reduce negative emotions. Participants first completed a set of 6 cognitive tasks before going for relaxation in the virtual train, then completed another set of cognitive tasks like the first set. Relaxation was used to decrease the distraction of participants. The levels of visual distraction of the participants before and after relaxation were later calculated and compared.

Participants started the experiment by wearing the Virtual Reality headset. All participants did Eye Tracking calibration using the VIVE calibration tool. At the start of the experiment, a gray screen was displayed for one minute, and participants were instructed to not think about anything, data from this period were used as a reference for later analysis. The experiment consisted of two parts as described previously. Participants were seated and had to interact with the environment only using the VR controller.

Each task in the **first part** went as follows. The problem was first presented on the screen and participants were instructed to solve it. After 30 s, the problem was hidden, and a red window (distractor) was displayed over the screen instructing participants to look left to see a hint, at the same time two suggestions were presented on the left side of the room. After five seconds, the red window disappeared, and the problem reappeared (except for the memorization task where nothing appeared). After 15 s, the problem was hidden and a red window was displayed over the screen instructing the participant to look right to see a hint, at the same time two suggestions were presented on the right side of the room. After 15 s, a keyboard appeared where participants could enter and validate their answers. Participants did not have a time limit and could validate their answer until it was correct. Figure 2 shows an example of the red window and hints displayed for a given task to distract the participants.

In the **second part**, during relaxation, participants went for a virtual tour on a train that lasted six minutes approximately. The train was moving, and they could hear the rail wheel as well as relaxing music playing through the sound output system of the VR headset. They could also see other nonplayer characters on the train including a family that was seated next to the player. Participants visited three locations aboard the train, a forest, a frozen mountain, and a desert. A detailed description of the environment was reported in [13]. We later analyzed the data from the six cognitive tasks and compared the data before and after relaxation in order to assess the participants' attentional state during these two phases.

### 3 Analysis and Results

31 participants ( $M = 16$ ,  $F = 15$ ) aged between 17 and 44 (mean = 23, std = 5) undertook the experiment at BMU (Beam Me Up Labs Inc., Montréal, Quebec Canada).



**Fig. 2.** Example of the red window and hints for a given task. (Color figure online)

All the participants came from Canadian universities and were either current students or graduates, except one participant who was a CEGEP student.

The average duration of the experiment for each participant was 40 min. We discarded the data of one participant who did not finish the experiment. The data from 30 participants ( $M = 15$ ,  $F = 15$ ) were used for the rest of the study. Scikit-learn python library [18] was used for all the computational analyses, and matplotlib library for python [19] was used to create all the visual plots.

### 3.1 Feature Extraction

Collected Eye Tracking data were analyzed using a script written manually in Python language. When returning eye movement values at a given timestamp, SRanipal SDK also returns a value indicating the validity of data. Data are considered valid if the validity value is 31 and considered invalid otherwise. Thus, we dropped invalid eye movements for all participants.

Hutt et al. [10] and Benedek [20] used an extensive list of eye movements features in their study of attention using Eye Tracking. Here, we extracted and used the most relevant features.

Imaoka et al. [21] studied the feasibility of using an HTC Vive Virtual Reality headset to assess eye saccades, and results suggested that VIVE Pro Eye could function as an assessment tool for saccadic eye movement. Saccades and fixations were computed using a velocity-based identification algorithm [22], where we considered eye movements with a velocity higher than 300 degrees per second saccades and eye movements with a velocity lower than 100 degrees per second fixations, similarly to [22]. We then extracted the saccade count, average saccade amplitude, fixations count, and average fixation duration.

Blinks were also computed using a simple threshold method where eye movements with eye openness lower than 45% were considered blinks and were not considered blinks otherwise. Then, blinks count, and average blink duration were extracted.

Furthermore, the vergence angle of the eyes, which is the angle of the convergence point of the eyes was computed using gaze direction vectors. Then, the average angle and angle variance were computed. Finally, we extracted the average pupil diameter and pupil diameter variance.

To extract features for data samples, we used segments of data of length T. Moreover, to compute saccades, fixations, blinks, and pupil diameters, the average data of both eyes were used. In total, 10 features were extracted from Eye Tracking data.

### 3.2 Eye Tracking Model for Detecting Distraction

To study the feasibility of detecting **distraction using Eye Tracking**, we created a set of data with eye movement features. For each task, four samples with a time window T of five seconds length each were created. The first two samples included the data from second 10 to second 15, and from second 15 to second 20, and were labeled “not distracted”. The other two samples included data from the two five seconds following the appearance of the red window (onset of the distractor) and were labeled “distracted”. We justified the labels of the samples in Sect. 2.2. A total of 36 samples were created for each participant.

We trained multiple machine learning models to learn this classification task. The models were trained and tested using the leave-p-groups-out cross-validation (LPGOCV) method to evaluate the models on data from the groups it never observed during training. Each group corresponded to the data of a participant. Each time, the models were trained on a subset of n-p groups, then tested on a subset of p groups. This procedure was repeated until the model was tested once on all possible combinations of p groups, and the evaluation metrics were averaged. The maximal value  $p = 4$  (87% data for training and 13% for testing) was used, as larger values for p required significantly more time to complete. To increase the efficiency of the models, a scaler was set up using the training data. Average F1-score, Recall, and Precision over the two classes were used as metrics to evaluate the performance of the models.

Table 1 shows the average scores as well as standard deviations after evaluating the models using LPGOCV with  $p = 4$ . Logistic Regression, Random Forest, and Multilayer Perceptron achieved the highest F1-score of 86% for correctly detecting distraction, while K-Nearest Neighbors (KNN) achieved the lowest F1-score of 84%.

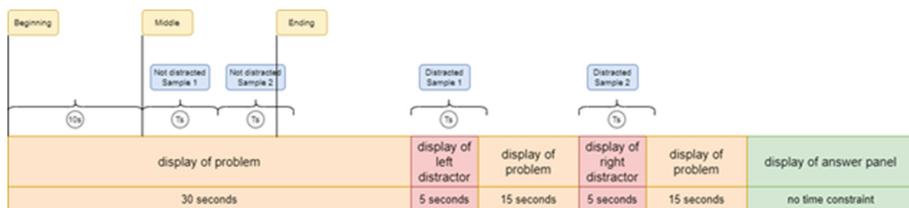
### 3.3 Investigation of Dataset Creation Parameters

**Time window T** represents the length in seconds of the data segments used to create the samples. Lee et al. [9] used a T value of one second to monitor attention but did not investigate other values. We compared the effects of different T values on the distraction models developed in Sect. 3.2. Moreover, due to the structure of the experiment used to collect data, we investigated another parameter, the moment which represents the time in seconds when the samples labeled “not distracted” were created.

**Table 1.** Results after model evaluation using LPGOCV with  $p = 4$ 

Model	F1 Distraction Mean, Std	Recall Distraction Mean, Std	Precision Distraction Mean, Std
KNN	84.84, 4.96	84.96, 9.08	85.79, 6.75
Logistic Regression	86.43, 4.66	86.50, 7.62	87.03, 5.99
RBF SVM	86.40, 5.21	86.48, 8.24	87.07, 6.39
Random Forest	86.74, 3.95	86.79, 6.72	87.23, 5.44
Naive Bayes	85.74, 4.84	85.83, 8.32	86.49, 6.37
Multilayer Perceptron	86.43, 4.87	86.51, 8.14	87.13, 6.24
Random Baseline	49.92, 4.66	50.00, 5.88	50.01, 4.2

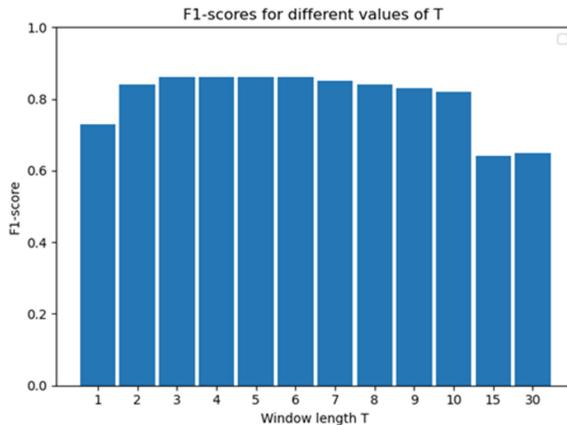
Figure 3 shows an overview of the different phases of a task in the first part of the experiment and illustrates the different parameters for the creation of samples.

**Fig. 3.** Overview of tasks from part one, and illustration of samples' creation parameters

For  $T$ , we tried values in the set  $S_t = \{1,2,3,4,5,6,7,8,9,10,15,30\}$ . For the moment parameter, we tried three moments. The first moment was at the beginning of the tasks (labeled “beginning”), and the second moment was ten seconds after the beginning of the tasks (labeled “middle”). The final moment was 20 s after the beginning of the tasks (labeled “ending”).  $T$  parameter values were investigated first, then the best  $T$  value was used to investigate the moment parameter.

We compared the different parameters values on the distraction detection task from Sect. 3.2 using a Logistic Regression model. Figure 4 shows results of F1-scores for different  $T$  values. The window length values 3, 4, 5, and 6 gave the highest F1-scores, 86%.

Window length  $T$  values from 3 to 6 inclusively had the highest average F1-score.  $T = 3$  was selected for the next analysis. Eye movements data extracted starting from the middle of the focus phase (moment = middle) resulted in the highest F1-score value of 86%, followed by the ending moment with 82%, and the beginning moment with 77%. We used the parameters  $T = 3$  and moment = middle in the rest of the study.



**Fig. 4.** F1-scores for different values of T.

### 3.4 Improvement of Participants' Attention

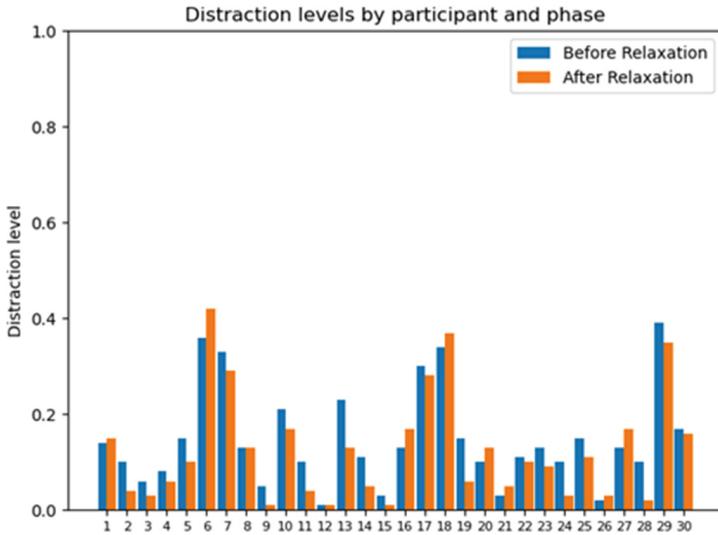
To investigate effects of the relaxation in VR on attention levels, Eye movements data of participants while completing cognitive tasks before relaxation and after relaxation were compared. This approach was used by Frasson and Ben Abdessalem [17] to investigate the effects of relaxation on the negative emotions in the elderly. We split the entirety of data from the second part into three-second segments and created eye movements samples. Then, the samples were classified using a Logistic Regression model trained on the dataset from Sect. 3.1 with parameters  $T = 3$  and moment = middle with either label “0” if not distracted or label “1” if distracted.

Figure 5 shows average visual distraction levels by participants and phase (before relaxation, and after relaxation) using the classification labels. Levels of distraction before relaxation exceeded levels of distraction after relaxation on average for 20 participants. Distraction decreased by 15% on average after relaxation compared to before relaxation.

A one-tailed paired t-test was performed to compare the levels of distraction of participants before relaxation and after relaxation. The results from pre-relaxation ( $M = 0.14$ ,  $SD = 0.10$ ) and post-relaxation ( $M = 0.12$ ,  $SD = 0.11$ ) indicate that the relaxation in Virtual Reality resulted in a decrease in visual distraction levels;  $t(29) = 3.03$ ,  $p = .002$ .

## 4 Discussion

The purpose of this study was to find a means to detect and monitor distraction, and to find methods to improve the attention of individuals while completing cognitive tasks in a Virtual Reality environment using Eye Tracking technology. To do that, we designed an experiment in two steps. In the first step, we manipulated the visual attention of participants experimentally using cognitive tasks and distractors. In the second step, participants completed a set of cognitive tasks before trying relaxation to improve their



**Fig. 5.** Distraction levels by participant and phase.

attention, then completed another set of cognitive tasks. We then compared the levels of attention of participants before and after relaxation to investigate the effects of our method.

We started by developing a tool to detect visual distraction by training machine learning models with samples containing eye movements features. The best model achieved an F1-score of 86% in a participant-independent setting with LPGOCV ( $p = 4$ ). These results suggest it is possible to effectively discriminate between the “distracted” state and “not distracted” state using Eye Tracking in Virtual Reality even for new participants never seen before.

Eye features were all computed manually using methods from literature, and were not the only focus of the study, suggesting that the performance of detection of distraction could be further improved. Saccades and fixations were computed using a simple velocity-based algorithm [22]. Blinks were computed using the same method as saccades and fixations, by collapsing successive blink points and non-blink points. We considered a high threshold for blinks (0.45 from Sect. 3.1) because of the temporal resolution of the Eye Tracking device (90 Hz). Blinks happen rapidly and the Eye Tracking device could not always capture the eyes openness during blinks, which resulted in the recorded eye openness value being high most of the time. The VIVE Pro Eye was previously validated as a tool to assess saccades and fixations [21]. In contrast, the computation of other eye movements features such as eye blinks using the VIVE Pro Eye was not validated.

We investigated the dataset creation parameters T and moment such as in [23]. The classification models were robust to the different parameters’ values. While moment = middle resulted by far in the highest F1-score, T values between 3 and 6 all performed similarly. We considered T = 3 for the subsequent analysis for it is a fine tradeoff between performance and window size, using a small value would permit to monitor attention levels closely and detect changes quickly.

To investigate the second hypothesis, we used the attention model from Sect. 3.2 to monitor the visual distraction levels before and after going through relaxation retrospectively using the monitoring method from [9]. Results revealed lower distraction levels after relaxation. To validate these results, a paired t-test was performed, and confirmed the results, suggesting that relaxation significantly decreased the visual distraction levels. These findings agree with the improvement in performance in attention exercises after relaxation [13].

Results from Sect. 3.4 confirmed that the levels of visual distraction of participants decreased after using relaxation. [13] also found that the performance of participants in attention tasks improved after relaxation compared to before relaxation.

## 5 Conclusion and Future Works

We developed an immersive environment in Virtual Reality in order to manipulate visual attention of individuals. We then used Eye Tracking data to create a model able to detect visual distraction levels of participants while they solved cognitive problems. The model was validated and later used to monitor retrospectively the distraction levels of participants while we tried to improve their concentration levels using relaxation in VR. Our findings suggest that it is possible to effectively detect the levels of visual distraction of learners using Eye Tracking and to decrease their levels of distraction by using relaxation in Virtual Reality. Modern learning systems can greatly benefit by integrating those methods to monitor the attentional state of learners and adapt the proposed content accordingly.

Our results suggest that relaxation in VR could help the visual distraction of individuals. These results are a good indicator that relaxation can improve the attention of individuals, but a study of the cognitive state of the participants using tools such as EEG is necessary to confirm this theory.

Many aspects of attention were studied, mind wandering in [10], and visual distraction in this work, both using Eye Tracking. While the work on mind wandering was promising, it also might have revealed the limitations of Eye Tracking when studying internal cognition [16]. Combining EEG and Eye Tracking may solve the existing problems in attention classification, for example by focusing the eye tracking on the visual (external) aspect of attention, and the EEG on the internal aspect of attention.

## 6 Limitations

As per our extensive research, we found no available tool to extract eye movement features from the HTC VIVE Pro Eye headset without incurring a cost. Therefore, we developed our own method to extract these features, taking into consideration budget constraints. It implies that utilizing specialized tools should result in even better-quality models.

**Acknowledgements.** We acknowledge NSERC-CRD (National Science and Engineering Research Council Cooperative Research Development), Prompt, and BMU (Beam Me Up) for funding this work.

## References

1. Ophir, E., Nass, C., Wagner, A.D.: Cognitive control in media multitaskers. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 15583–15587 (2009). <https://doi.org/10.1073/pnas.0903620106>
2. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard, R.: Affect-aware tutors: recognising and responding to student affect. *IJLT*. **4**, 129 (2009). <https://doi.org/10.1504/IJLT.2009.028804>
3. Sharot, T., Phelps, E.A.: How arousal modulates memory: Disentangling the effects of attention and retention. *Cogn. Affect. Behav. Neurosci.* **4**, 294–306 (2004). <https://doi.org/10.3758/CABN.4.3.294>
4. Sana, F., Weston, T., Cepeda, N.J.: Laptop multitasking hinders classroom learning for both users and nearby peers. *Comput. Educ.* **62**, 24–31 (2013). <https://doi.org/10.1016/j.compedu.2012.10.003>
5. Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R.: Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learn. Instr.* **21**, 267–280 (2011). <https://doi.org/10.1016/j.learninstruc.2010.07.004>
6. Zheng, J.M., Chan, K.W., Gibson, I.: Virtual reality. *IEEE Potentials* **17**, 20–23 (1998). <https://doi.org/10.1109/45.666641>
7. Allcoat, D., von Mühlenen, A.: Learning in virtual reality: effects on performance, emotion and engagement. *Research in Learning Technology*. **26**, (2018). <https://doi.org/10.25304/rlt.v26.2140>
8. Ghali, R., Abdessalem, H.B., Frasson, C.: Improving intuitive reasoning through assistance strategies in a virtual reality game. Presented at the The Thirtieth International Flairs Conference , Marco Island, Florida, USA (2017)
9. Lee, G., Ojha, A., Lee, M.: Concentration Monitoring for Intelligent Tutoring System Based on Pupil and Eye-blink. In: Proceedings of the 3rd International Conference on Human-Agent Interaction. pp. 291–294. ACM, Daegu Kyungpook Republic of Korea (2015). <https://doi.org/10.1145/2814940.2815000>
10. Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., D'Mello, S.: "Out of the Fr-Eye-ing Pan": Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 94–103. ACM, Bratislava Slovakia (2017). <https://doi.org/10.1145/3079628.3079669>
11. Kaplan, S.: The restorative benefits of nature: toward an integrative framework. *J. Environ. Psychol.* **15**, 169–182 (1995). [https://doi.org/10.1016/0272-4944\(95\)90001-2](https://doi.org/10.1016/0272-4944(95)90001-2)
12. Gao, T., Zhang, T., Zhu, L., Gao, Y., Qiu, L.: Exploring Psychophysiological Restoration and Individual Preference in the Different Environments Based on Virtual Reality. *IJERPH*. **16**, 3102 (2019). <https://doi.org/10.3390/ijerph16173102>
13. Ben Abdessalem, H., et al.: Immersion en train thérapeutique pour la relaxation de patients Alzheimer. *French Journal of Psychiatry*. **1**, S152 (2019). <https://doi.org/10.1016/j.fjpsy.2019.10.422>
14. Huang, M.X., Li, J., Ngai, G., Leong, H.V., Bulling, A.: Moment-to-Moment Detection of Internal Thought from Eye Vergence Behaviour. (2019). <https://doi.org/10.48550/ARXIV.1901.06572>
15. Myrden, A., Chau, T.: A Passive EEG-BCI for Single-Trial Detection of Changes in Mental State. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 345–356 (2017). <https://doi.org/10.1109/TNSRE.2016.2641956>
16. Chun, M.M., Golomb, J.D., Turk-Browne, N.B.: A Taxonomy of External and Internal Attention. *Annu. Rev. Psychol.* **62**, 73–101 (2011). <https://doi.org/10.1146/annurev.psych.093008.100427>

17. Frasson, C., Abdessalem, H.: Contribution of Virtual Reality Environments and Artificial Intelligence for Alzheimer. MRAJ. 10, (2022). <https://doi.org/10.18103/mra.v10i9.3054>
18. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. (2012). <https://doi.org/10.48550/ARXIV.1201.0490>
19. Hunter, J.D.: Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. **9**, 90–95 (2007). <https://doi.org/10.1109/MCSE.2007.55>
20. Benedek, M., Stoiser, R., Walcher, S., Körner, C.: Eye Behavior Associated with Internally versus Externally Directed Cognition. Front. Psychol. **8**, 1092 (2017). <https://doi.org/10.3389/fpsyg.2017.01092>
21. Imaoka, Y., Flury, A., de Bruin, E.D.: Assessing Saccadic Eye Movements With Head-Mounted Display Virtual Reality Technology. Front. Psychiatry. **11**, 572938 (2020). <https://doi.org/10.3389/fpsy.2020.572938>
22. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the symposium on Eye tracking research & applications - ETRA '00. pp. 71–78. ACM Press, Palm Beach Gardens, Florida, United States (2000). <https://doi.org/10.1145/355017.355028>
23. Atyabi, A., Fitzgibbon, S.P., Powers, D.M.W.: Multiplication of EEG Samples through Replicating, Biassing, and Overlapping. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) BI 2012. LNCS (LNAI), vol. 7670, pp. 209–219. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-35139-6\\_20](https://doi.org/10.1007/978-3-642-35139-6_20)

# **Neural Networks and Data Mining**



# Using Feature Interaction for Mining Learners' Hidden Information in MOOC Dropout Prediction

Tingfeng Pan<sup>1(✉)</sup>, Guang Feng<sup>2</sup>, Xin Liu<sup>2</sup>, and Wenyan Wu<sup>1(✉)</sup>

<sup>1</sup> School of Computer Science and Technology, Guangdong University of Technology,  
Guangzhou, China

2112005017@mail2.gdut.edu.cn, wuwy@gdut.edu.cn

<sup>2</sup> School of Automation, Guangdong University of Technology, Guangzhou, China

**Abstract.** Massive open online courses (MOOC) are increasingly prevalent as a result of the rise in internet usage in recent years. However, the current development of MOOC is being severely hampered by the high dropout rates. The primary research goal of this work is to develop prediction models to identify students who are likely to exhibit dropout behavior in advance. In this paper, we propose the Cross-TabNet, which efficiently learns feature-hidden information by explicit feature interaction and uses sequential attention-based TabNet for classification. The experimental results demonstrate that it outperforms existing machine learning and deep learning methods.

**Keywords:** MOOC dropout prediction · Feature interaction · Neural network · Attentional mechanism

## 1 Introduction

MOOC were created to allow students' learning to be unrestricted by time and space, making it easier for anyone who wants to study. The rise and development of MOOC is a great progress of the traditional education model, which differs from the traditional education model in that learners can study the courses on the MOOC platform at any place and any time, choose different courses according to their interests and hobbies, and become classmates with strangers from all over the world and use the same teaching resources together. MOOC allow learners to enjoy the quality and equal education through the Internet. Because of these advantages, MOOC are growing and becoming more and more popular.

However, despite the advantages of MOOC such as openness and fairness, the low pass rate and the high dropout rate of learners in MOOC courses have grown to be one of the key issues preventing MOOC progress [5]. According to a study [10] that analyzed the course completion rate by collecting MOOC-related literature, the completion rate of most courses was below 40%, and the

average completion rate of courses was around 15%, implying that nearly 85% of MOOC participants failed to complete the courses and dropped out ahead of time. It's critical to predicting learners' dropout behavior ahead of time so that the MOOC platform can provide useful interventions. By collecting data on students' learning behaviors, the MOOC platform allows researchers to figure out the relationship between their low completion rates and their learning behaviors. As a result, the new dropout prediction model based on machine learning and deep learning is being considered by a number of researchers and educational technologists [6].

In this study, we combine the TabNet and the Compressed Interaction Network to build a deep neural network called Cross-TabNet, which can effectively mine hidden information through explicit feature interaction from learners' learning activity features.

The following is the paper's structure: Sect. 2 describes the related work in MOOC dropout prediction; In Sect. 3, the datasets and the suggested prediction model are detailed; In Sect. 4, the experimental assessment and results are shown; and in Sect. 5, a summary and a suggestion for further research are provided.

## 2 Related Work

Since MOOC first appeared in 2012, the issue of a high dropout rate has been a significant factor impeding its growth. Researchers have started to experiment with identifying students who are likely to drop out of school by constructing predictive models as a result of the ongoing development of Internet technologies.

Lacave et al. [11] used Bayesian Network for the prediction of dropout behavior and proved through experimental results that the Bayesian network using K2 algorithm has the best performance. Considering that the performance improvement of a single model is always limited, Liu et al. [13] developed a hybrid method based on MOOC forum's data to mine learner roles for semantic feature extraction from MOOC forum posts and comments, and the testing findings revealed that the hybrid method worked better than other methods. Alamri et al. [2], on the other hand, expected to build simple models based on as few features as possible to improve the generality and applicability of the models across different MOOC platforms, so they derived that the quantity of times students access the course information and the time spent are the most important features in all courses based on the Gini index ranking. And he constructs machine learning models based on the above features, which include Random Forest, Gradient Boosting Machine (GBM), Adaboost, and Extreme Gradient Boosting Machine (XGBoost). The experimental findings demonstrated that the performance of different models did not vary much, and the prediction accuracy remained between 82% and 94% in all four selected courses, and the selected features were effective in predicting dropouts and continuing learners regardless of the model used.

In an experimental comparison, Basnet et al. [4] examined the effectiveness of the most popular machine learning models at the time, and the results pointed

out that among all machine learning models, the tree ensemble models (extreme gradient boosters, random forests, and Adaboost) performed the best. Mubarak et al. [16] predicted dropout behavior by Logistic Regression with the regularized item and Input-Output Hidden Markov Model (IOHMM) and studied that the more interaction students have with the course, the higher their probability of staying for further study.

However, machine learning-based models require high feature quality, and the model performance is mainly determined by the feature's information. Therefore, the machine learning model performance is limited by expert manual feature engineering and no effective learning of potential information among features. Some researchers have attempted to employ deep learning-related methods for prediction as deep learning research has advanced.

Wang et al. [17] constructed the prediction model CNN-RNN by combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), and the experimental outcomes showed that the model's performance is on par with feature engineering-based approaches and can significantly lessen the need on manual feature engineering. To solve the sample imbalance problem, Wu et al. [18] proposed CLMS-Net, which outputs CNN-LSTM into Support Vector Machine (SVM) and adjusted the sample imbalance problem by using the sample weight parameter in SVM and achieved better performance, where the AUC improved by about 0.45% and F1 by about 4.03%. Also addressing the sample imbalance problem, Mubarak et al. [15] optimized the cross-entropy loss function based on the CNN-LSTM model by introducing classes weights, thus reducing the impact of sample imbalance on model training optimization and prediction performance. Zhang et al. [20] applied SE-Net (Squeeze and Excitation Network) on top of CNN for local learning behavioral features extraction and attention weighting, and used Gate Recurrent Unit (GRU) to capture temporal correlations to improve the predictive accuracy. In addition to proposing optimizations for the model itself and the loss function, Jin et al. [9] considered a clustering algorithm to change the weights of student samples to address the sample imbalance and improve performance.

To increase the reliability of prediction, Feng et al. [7] integrated contextual information related to the learner into the input features as well and proposed the context-aware feature interaction network (CFIN). Fu et al. [8] proposed CLSA based on CNN-LSTM, which uses static attention for attention weighting of temporal feature sequences, and the improved model improved accuracy by 2.8% and F1 by 1.6% compared to CNN-LSTM. Alam et al. [1] developed the first multimodal spatiotemporal neural fusion model MSNF-MTCL for dropout prediction multitasking, also employing convolutional neural networks for higher-order feature extraction on structured data. Xu et al. [19] found that learners' learning behaviors between two adjacent days are similar, so based on this property, Lie Group regional covariance matrix was designed to represent the local correlation features of learning behaviors and used to extract higher-order local interaction features through CNN to improve the model performance.

Although deep learning models can perform automatic feature extraction by deep neural networks, deep neural networks are implicitly higher-order feature extraction, which is less interpretable and lacks learning of lower-order feature interaction.

### 3 Methodology

#### 3.1 Datasets

We compare the model's performance by conducting experiments on two large datasets: KDD Cup 2015 dataset and XuetangX dataset. The KDD Cup 2015 dataset records students' learning behavior records data on the XuetangX platform from 2013 to 2014. The dataset recorded a total of 79,186 learners' learning behavior records for 30 days on 39 catechism courses. And the XuetangX dataset records 42,110,402 log messages from 123,719 learners in 515 courses with 209,600 registrations on the XuetangX platform from 2015 to 2017 [7]. In both datasets, the label indicates whether the student will drop out in the future, with 1 indicating a dropout and 0 indicating no dropout.

#### 3.2 Feature Engineering

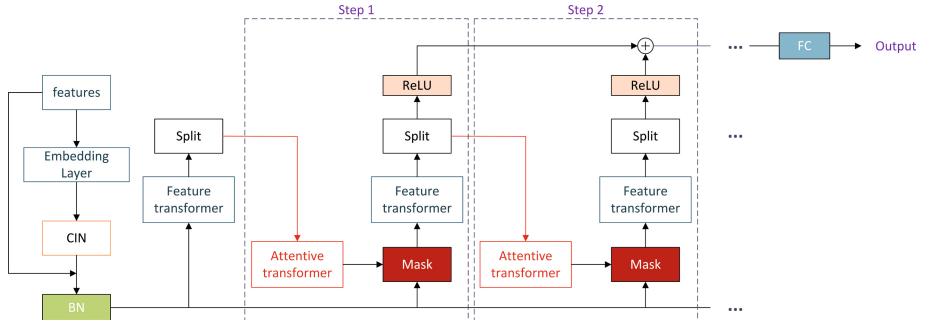
In KDD Cup 2015 dataset, the learning behavior features include the number of learning behavior records per day as well as the learning time records. Statistical features include the minimum, maximum, sum, mean, and standard deviation of the number of valid learning behaviors for the first ten days, the middle ten days, the last ten days, and the 30 days. Contextual features include course numbers and the number of selected courses for each enrollment id.

The learning behavior features and statistical features of the XuetangX dataset is constructed in the same way as the KDD Cup 2015 dataset. The contextual features include course number, course category, and number of course registrations, and the learner contextual features include learner age, education, education status, number of semesters of participation, and number of courses registered.

#### 3.3 Model

Cross-TabNet's structure is depicted in Fig. 1. The input features are transformed into low-dimensional dense vectors by embedding operation in the embedding layer, then the cross features are extracted by the Compressed Interaction Network (CIN) [12]. Then the extracted cross features are merged with the input features and passed into TabNet [3] for decision-making after being normalized by the Batch Normalization layer (BN).

Let  $X \in \mathbb{R}^m$  denotes the each student's learning activity feature, where  $m$  is the feature's dimension. Firstly, the input features  $X$  are transformed into the input matrix  $X^0 \in \mathbb{R}^{m \times D}$  through the embedding layer, where  $D$  is the embedding dimension. The output of the  $k$ -th layer  $X^k$  is calculated as Eq. 1:



**Fig. 1.** The architecture of the Cross-TabNet.

$$X_{h,*}^k = \sum_{i=1}^{H_{k-1}} \sum_{j=1}^m W_{ij}^{k,h} (X_{i,*}^{k-1} \circ X_{j,*}^0), \quad (1)$$

where  $H_k$  is the number of feature vector in the  $k$ -th layer,  $W^{k,h} \in \mathbb{R}^{H_{k-1} \times m}$  ( $1 \leq h \leq H_k$ ) is the parameter matrix for  $h$ -th feature vector, and  $\circ$  denotes the Hadamard product. According to the Eq. 1, the  $k$ -th interaction feature  $X^k$  is calculated from  $X^{k-1}$  and  $X^0$ . Therefore, as the cross layer depth increases, so do the interactions.

The output of each layer in Compressed Interaction Network is finally combined in dimension D by sum pooling (see Fig. 2). The  $k$ -th layer's sum pooling operation is calculated as Eq. 2:

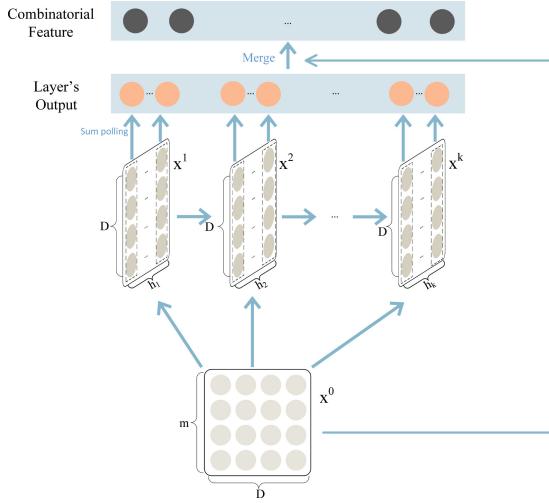
$$p_i^k = \sum_{j=1}^D X_{i,j}^k \quad (2)$$

where  $i \in [i, H_{k-1}]$ ,  $p_i$  denotes the value of the  $i$ -th feature by sum polling, then  $p^k = [p_1^k, p_2^k, \dots, p_{H_k}^k]$  denotes the  $k$ -th layer's output vector. The initial features are combined with each layer's output vector as given in Eq. 3. The new features  $G$  are obtained and passed to TabNet for feature selection after batch normalization.

$$G = [x, p^1, p^2, \dots, p^K] \quad (3)$$

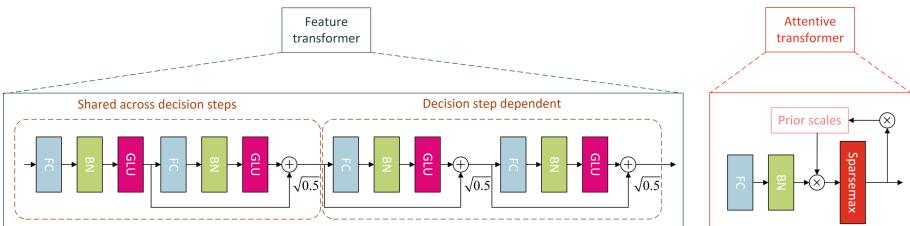
According to Fig. 1, TabNet is composed of several decision-making steps, and each of the steps includes a feature transformer, attentive transformer, mask layer, split layer, and Rectified Linear Unit (ReLU) activation. The attentive transformer and the feature transformer are two of the core components of TabNet, and the Fig. 3 depicts their architecture.

The feature transformer's primary duty is to perform feature transformation, and is made up of a series of GLU blocks, each of which is composed of a Fully Connected (FC) layer, a Batch Normalization (BN) layer, and a Gate Linear Unit (GLU) with a residual connection at a scale of  $\sqrt{0.5}$  (see Fig. 3). Compared with



**Fig. 2.** The architecture of the Compressed Interaction Network.

the simple FC layer, the variance transformation of the feature transformer fluctuates less, which can significantly increase the stability of the network learning process. Additionally, the feature transformer has two decision steps: a shared decision step that affects the entire network and a separate decision step that completes the independent feature transformation in each step.



**Fig. 3.** The architecture of the feature transformer and the attentive transformer proposed in TabNet [3].

After the feature transformer, the output is split by the split layer, and part of the features are used for the final decision and part of the features are used in the next step for a new round of feature selection, where the process of calculating the feature transformer and split layer in  $i$ -th step is calculated as Eq. 4:

$$[d[i], a[i]] = f_i(M[i] \cdot G), \quad (4)$$

where  $M[i] \in \mathbb{R}^{B \times m}$  denotes the attention weight matrix (Mask) learned by the attentive transformer,  $d[i] \in \mathbb{R}^{B \times N_d}$  is the decision step output,  $a[i] \in \mathbb{R}^{B \times N_a}$

is the next attentive transformer step input,  $N_d$  and  $N_a$  are split size hyper-parameters, and  $B$  is the batch size.  $f_i$  denotes the feature transformer's operator function. In particular, the Mask matrix in the calculation before entering the first decision step is set to  $1^{B \times m}$ .

The attentive transformer's main function is to calculate the Mask matrix based on the previous step's output, making the model capable of feature selection. The attentive transformer is composed of a FC layer, a BN layer, a Sparsemax layer, and a weighted scaling factor (Prior scales), is calculated as Eq. 5:

$$M[i] = \text{Sparsemax}(P[i-1] \cdot h_i(a[i-1])) \quad (5)$$

where  $a[i-1]$  denotes the segmentation output of the feature transformer in the previous step and  $h_i$  denotes a trainable function consisting of a FC layer and a BN layer. The prior scaling term,  $P[i] = \prod_{j=1}^i (\gamma - M[j])$ , indicates how many individual features were used in previous steps. If a feature is used in the previous steps, the relaxation parameter  $\gamma$  can reduce its weight in next feature selection.  $P[0]$  is initialized as a matrix:  $1^{B \times m}$ .

The Sparsemax layer is a sparse probabilistic activation function [14], and its output is more sparse than Softmax's. Therefore, Sparsemax is used in the attentive transformer to ensure that the output result Mask matrix distribution is sparse, thus serving the purpose of feature selection.

The loss function is calculated as Eq. 6:

$$L = \frac{1}{N} \sum_i -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda_{\text{sparse}} L_{\text{sparse}}, \quad (6)$$

with a sparse regularization term:

$$L_{\text{sparse}} = \sum_{i=1}^{N_{\text{steps}}} \sum_{b=1}^B \sum_{j=1}^D \frac{-M_{b,j}[i] \cdot \log(M_{b,j}[i] + \varepsilon)}{N_{\text{steps}} \cdot B} \quad (7)$$

to control the sparsity of Mask.

The output of the whole model is calculated as:

$$d_{\text{out}} = \sum_{i=1}^{N_{\text{steps}}} \text{ReLU}(d[i]). \quad (8)$$

## 4 Experiments and Result

### 4.1 Parameter Setting

For the KDD Cup 2015 dataset, the Cross-TabNet model use  $N_d = N_a = 8$ , batch size = 512,  $N_{\text{steps}} = 9$ ,  $\gamma = 1.3$ ,  $\lambda_{\text{sparse}} = 0.0001$ ,  $k = 2$ ,  $H_k = (128, 128)$  and embedding size D = 4. For the XuetangX dataset, the different parameter setting is  $H_k = (96, 96)$  and batch size = 256,  $\lambda_{\text{sparse}} = 0.0005$ .

## 4.2 Comparison Methods and Implementation Details

Comparison methods are posted as follow:

- LR: Logistic Regression Model.
- RF: Random Forest.
- GBDT Gradient Boosting Decision Tree.
- CNN-LSTM: CNN-based LSTM.
- CFIN: Context-aware Feature Interaction Network [7].
- CLSA: Attention mechanism based CNN-LSTM [8].

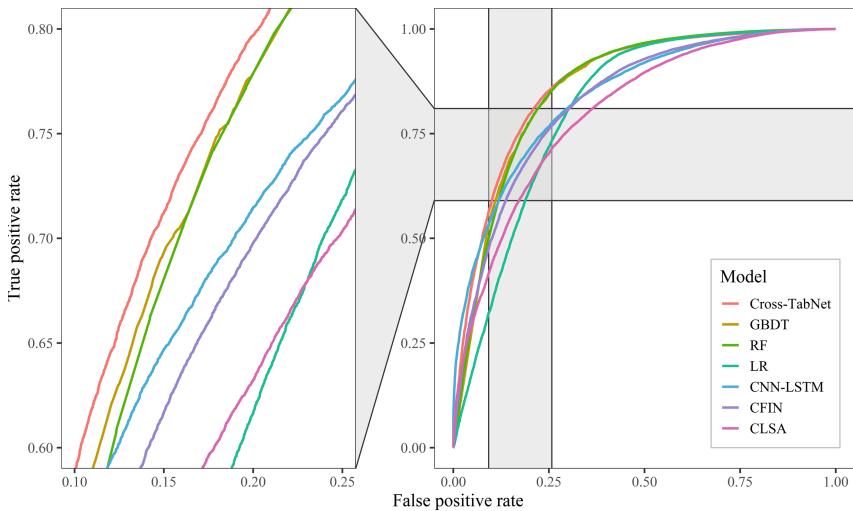
The above algorithms were implemented in Python using the scikit-learn package, PyTorch 1.8.1 and CUDA 11.2. Weighted precision (WP), weighted recall (WR), Area Under the ROC Curve (AUC), and weighted F1 Score are among the evaluation metrics.

## 4.3 Result

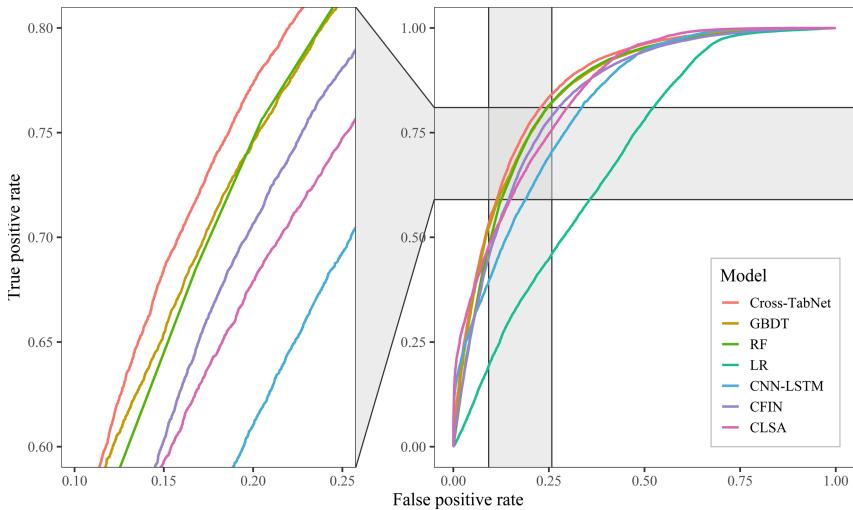
Figure 4 and Fig. 5 show the receiver operating characteristic curve (ROC) of the model on the KDD Cup 2015 dataset and the XuetangX dataset, respectively. According to the above figures, the performance of logistic regression and support vector machine is not very good on both datasets, and the performance of decision tree-based random forest and gradient boosting decision tree is closer, but both are slightly lower than Cross-TabNet. Cross-TabNet also performs better than deep learning-based models (CNN-LSTM, CFIN, and CLSA). To further compare the overall performance of all algorithms, Table 1 shows the performance of each method on KDD Cup 2015 dataset and XuetangX dataset, where the bolded data are the best results.

**Table 1.** Overall results on two datasets.

Model	KDD Cup 2015				XuetangX			
	AUC	WP	F1	WR	AUC	WP	F1	WR
LR	0.8043	0.8257	0.8212	0.8177	0.6900	0.8018	0.7407	0.7943
SVM	0.7846	0.8250	0.8001	0.8331	0.7845	0.8315	0.8207	0.8136
RF	0.8628	0.8846	0.8621	0.8496	0.8495	0.8511	0.8439	0.8393
GBDT	0.8629	<b>0.9270</b>	0.8661	0.8331	0.8548	<b>0.8587</b>	0.8461	0.8389
CNN-LSTM	0.8182	0.8258	0.7298	<b>0.9041</b>	0.8180	0.8361	0.8201	0.8391
CLSA	0.8031	0.8174	0.7739	0.8211	0.8468	0.8523	0.8311	0.8498
CFIN	0.8309	0.8271	0.8224	0.8407	0.8365	0.8242	0.8257	0.8276
Cross-TabNet	<b>0.8758</b>	0.8671	<b>0.8721</b>	0.8735	<b>0.8651</b>	0.8455	<b>0.8470</b>	<b>0.8510</b>



**Fig. 4.** Receiver operating characteristic curve for each model on the KDD Cup 2015 dataset.



**Fig. 5.** Receiver operating characteristic curve for each model on the XuetangX dataset.

As shown in Table 1, Cross-TabNet fully considers the information gained from feature interaction and learns the hidden information in features better than the other prediction model, thus achieving performance improvement. The main cause of the LR and SVM's poor performance is that both are better suited for data sets with fewer features, and as a result, underperform for large data sets with complex features. Both Random Forest and Gradient Boosted

Decision Tree are tree-based integrated models with faster training times, greater generalization capacity, and generally better performance. However, Random Forest and Gradient Boosted Decision Tree perform worse than Cross-TabNet because they don't learn about feature interaction. The weighted precision of GBDT is relatively better, but its AUC and F1 are both lower than Cross-TabNet. The neural network-based methods (CNN-LSTM, CLSA, CFIN) also have relatively poor performance because it only learns higher-order crossover features and lacks the learning of lower-order explicit features. CNN-LSTM has the highest weighted recall under the KDD Cup 2015 dataset, but other metrics are relatively weak, presumably because the model is more biased toward the classification of positive samples during the training process.

## 5 Conclusion

In this paper, the Cross-TabNet is proposed to fully utilize the potential information between features. The superiority of this algorithm lies in its mining of feature potential information through explicit feature interaction and achieve recognition of important features through attention mechanism. Its performance is experimentally demonstrated to exceed current mainstream algorithms with AUC values of 0.8758 and 0.8651, as well as weighted F1 values of 0.8721 and 0.8470.

Overall, the results show how well the proposed model works to predict students' likelihood of dropping out, and its future applicability will be evaluated based on its implementation on a MOOC platform. For the future research direction of dropout prediction, we can analyze student behavior from multiple perspectives based on more diversified data, such as video features, voice features, text features, etc. The second is to make timely interventions or curriculum adjustments based on the likelihood of students dropping out, thereby improving the quality of the course and helping learners sustain their learning.

**Acknowledgments.** The project received support from China's National Natural Science Foundation (71671048).

## References

1. Alam, M.A.U.: College student retention risk analysis from educational database using multi-task multi-modal neural fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12689–12697 (2022)
2. Alamri, A., et al.: Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In: Coy, A., Hayashi, Y., Chang, M. (eds.) ITS 2019. LNCS, vol. 11528, pp. 163–173. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-22244-4\\_20](https://doi.org/10.1007/978-3-030-22244-4_20)
3. Arik, S.Ö., Pfister, T.: TabNet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6679–6687 (2021)

4. Basnet, R.B., Johnson, C., Doleck, T.: Dropout prediction in MOOCs using deep learning and machine learning. *Educ. Inf. Technol.* **27**(8), 11499–11513 (2022)
5. Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T.: Studying learning in the worldwide classroom research into edX's first MOOC. *Res. Pract. Assess.* **8**, 13–25 (2013)
6. Dalipi, F., Imran, A.S., Kastrati, Z.: MOOC dropout prediction using machine learning techniques: review and research challenges. In: 2018 IEEE Global Engineering Education Conference (EDUCON), pp. 1007–1014. IEEE (2018)
7. Feng, W., Tang, J., Liu, T.X.: Understanding dropouts in MOOCs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 517–524 (2019)
8. Fu, Q., Gao, Z., Zhou, J., Zheng, Y.: CLSA: a novel deep learning model for MOOC dropout prediction. *Comput. Electr. Eng.* **94**, 107315 (2021)
9. Jin, C.: Dropout prediction model in MOOC based on clickstream data and student sample weight. *Soft. Comput.* **25**, 8971–8988 (2021)
10. Jordon, K.: MOOC completion rates: the data (2015). <http://www.katyjordan.com/MOOCproject.html/>
11. Lacave, C., Molina, A.I., Cruz-Lemus, J.A.: Learning analytics to identify dropout factors of computer science studies through Bayesian networks. *Behav. Inf. Technol.* **37**(10–11), 993–1007 (2018)
12. Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., Sun, G.: xDeepFM: combining explicit and implicit feature interactions for recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1754–1763 (2018)
13. Liu, Z., Zhang, Y.: A semantic role mining and learning performance prediction method in MOOCs. In: U, L.H., Xie, H. (eds.) APWeb-WAIM 2018. LNCS, vol. 11268, pp. 259–269. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01298-4\\_22](https://doi.org/10.1007/978-3-030-01298-4_22)
14. Martins, A., Astudillo, R.: From softmax to sparsemax: a sparse model of attention and multi-label classification. In: International Conference on Machine Learning, pp. 1614–1623. PMLR (2016)
15. Mubarak, A.A., Cao, H., Hezam, I.M.: Deep analytic model for student dropout prediction in massive open online courses. *Comput. Electr. Eng.* **93**, 107271 (2021)
16. Mubarak, A.A., Cao, H., Zhang, W.: Prediction of students' early dropout based on their interaction logs in online learning environment. *Interact. Learn. Environ.* **30**(8), 1414–1433 (2022)
17. Wang, W., Yu, H., Miao, C.: Deep model for dropout prediction in MOOCs. In: Proceedings of the 2nd International Conference on Crowd Science and Engineering, pp. 26–32 (2017)
18. Wu, N., Zhang, L., Gao, Y., Zhang, M., Sun, X., Feng, J.: CLMS-Net: dropout prediction in MOOCs with deep learning. In: Proceedings of the ACM Turing Celebration Conference-China, pp. 1–6 (2019)
19. Xu, C., Zhu, G., Ye, J., Shu, J.: Educational data mining: dropout prediction in XuetangX MOOCs. *Neural Process. Lett.* **54**(4), 2885–2900 (2022)
20. Zhang, Y., Chang, L., Liu, T.: MOOCs dropout prediction based on hybrid deep neural network. In: 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 197–203. IEEE (2020)



# On an Integrated Assessment for the Students Within an Academic Consortium

Popescu Doru-Anastasiu<sup>1(✉)</sup>, Cristea Daniela-Maria<sup>2,3</sup>, and Bold Nicolae<sup>1</sup>

<sup>1</sup> University of Pitești, Faculty of Sciences, Physics Education and Computer Sciences, Pitești, Romania

dopopan@gmail.com

<sup>2</sup> Babes-Bolyai University, Computer Science Department,  
Cluj-Napoca 43017-6221, Romania

<sup>3</sup> “1 Decembrie 1918” University of Alba Iulia, Alba-Iulia, Romania  
daniela.cristea@uab.ro

**Abstract.** The design of a good assessment is related to compliance with several requirements. A good design ensures the validity and correctness of the assessment process within education. This paper presents a modality of designing an assessment test in respect of the degree of difficulty of the items and the test in an integrated academic environment, meaning that the items used for assessment are collected at an academic consortium level and they can be used in an integrated way for all the similar faculties from the academic group of institutions. The presented model is composed of two main components: the item collector and the test generator. The generation of tests is made automatically, by using evolutionary algorithms (genetic-based) and machine learning (ML) methods. An implementation of the model for the genetic-based algorithm is also presented.

**Keywords:** genetic · assessment · consortium

## 1 Introduction

In universities, there are many joint courses, especially in university consortia. The degree of preparation of students can be best determined by tests with the same items. In order to generate assessment tests that can be used in several universities in the consortium, we propose a model web platform containing two components: Assessment - Item Collection (A-IC) and Assessment - Test Generation (A-TG). The novelty brought by the paper is related to the usage of an integrated assessment tool within an academic consortium, formed by several academic institutions. This characteristic can help with a standardization of the assessment within the academic environment.

In this matter, the paper presents several sections regarding the topic. The Literature review section presents state-of-the-art research regarding educational

assessment and the usage of genetic algorithms and machine-learning-based tools used for educational purposes. The Model description section presents the description of the model, its components and functionalities, and Results and discussions related to the implementation of the model for the genetic algorithm test generation with respect to the user requirements. A Conclusions section summarizes the entire research conducted for the paper.

## 2 Literature Review

In later years, the majority of the research papers regarding the assessment tests used in the educational process are being focusing on the automated generation of questions and learning analysis, due to the fast and great extent of the usage of online learning management systems. In this matter, research in the automated educational assessment area is mainly conducted in several directions, the main two of them being Question Generation (QG) and Answer Evaluation (AE), according to [5]. While the QG research is mainly conducted on the automated generation of questions from a corpus of text using specific language processing methods, the AE research is focused on the analysis of the assessment based on the responses given to the item used for assessment.

Regarding the generation of assessment tests in the context of the existence of several requirements, special attention can be given to the usage of genetic algorithms (GA) with optimal solutions for statistical issues [16], management [6,9] or healthcare [19].

The authors [14] discuss a type of optimization algorithm that is inspired by genetics, using notions such as chromosomes, genes, mutation, and crossover. The fitness of a sequence is determined by the number of keywords it shares with the user-defined set of keywords. The chromosomes are ordered by their fitness and the input data is made up of various parameters including the number of tests, the number of generations, and the user-defined keywords. The output data includes the first k solutions and the number of matching keywords for each sequence.

In [2], some systems use distributional similarity techniques or extract words from text content for multiple options generation. Other frameworks use deep reinforcement learning for automatic question generation from corpora.

The types of generated questions can be categorized by Bloom's taxonomy, which has six levels of increasing difficulty: Knowledge, Comprehension, Application, Analysis, Synthesis, and Creation. Mostow and Chen [12] developed an automated reading tutor that uses automatic question generation to improve students comprehension of text, while only 35.6% of generated questions were deemed acceptable, the accuracy of detecting counterfactual questions was high at 90.0%. Mitkov and colleagues [11] developed a system to generate multiple-choice closed questions with natural languages processing techniques such as shallow parsing, automatic term extraction, sentence transformation, computing of semantic distance, deploying hypernyms and coordinates in WordNet to retrieve concepts semantically close to the correct answer. For 1000 question

items, the development cost would require 30 h of human work using the system, while 115 h would be required without using the system.

In [13], there were used learning materials from a graduate-level introductory data science course at an R1 university in the northeastern United States. The course has six conceptual units and sixteen modules, each consisting of several data science topics. The modules include reading assignments, ungraded formative assessments, and weekly quizzes as graded summative assessments. Students also complete seven hands-on coding projects, which are evaluated by an automatic grading system. The authors focused on generating questions from the textual content of the six units in the course using a pipeline.

Robinson [17] distinguishes task complexity (cognitive factors), task difficulty (learner factors), and task conditions (interaction factors), meanwhile Campbell [4] contrasted multiple views of complexity, such as a psychological experience, a task-person interaction, and a function of objective characteristics, and created a typology of complex tasks.

In the context of language tasks, Mesmer et al. [10] explicitly distinguish text difficulty (based on the performance of readers) and text complexity (based on textual elements). Beckmann and Goode [1] distinguish between the concepts of complexity and difficulty and argue that complexity is a cognitive concept, while difficulty is a psychometric concept that reflects individual performance. They apply this distinction to the domain of complex problem-solving, such as controlling a dynamic system with feedback loops.

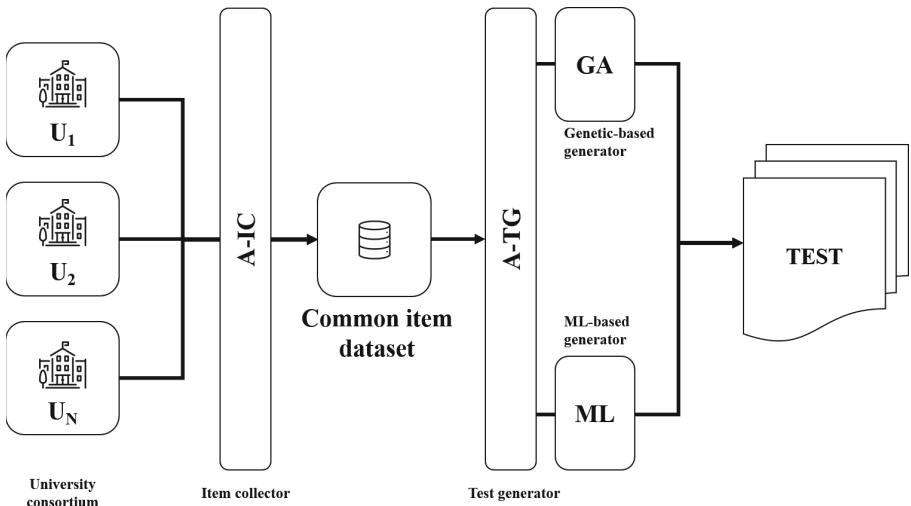
Gkatzia and Mahamood [7] conducted a study analyzing a corpus of 79 conference and journal papers on NLG published between 2005–2014, which revealed the increasing use of automatic evaluation over human evaluation and intrinsic evaluation over extrinsic evaluation.

Sancho-Asensio et al. [18] used a technique called “backtracking” to incrementally search for student profiles, while Zheng and Pinkwart [8] used the discrete particle swarm optimization algorithm to group learners according to their personality traits and gender.

### 3 Model Description

The model is built on the premise that the assessment must be standardized for an entire consortium, as the assessment would have a different degree of correctness and validity. The assessment is made by consolidating a common dataset of questions within the consortium, built using various modalities, such as a manual introduction or automated generation of questions from a corpus, using specific methods. This dataset has an organic dynamic, being extended by time. Also, the items that form the dataset have several characteristics that are permanently updated, one of the most important being the degree of difficulty, which will be pre-calculated using a specific method that will be described further, in the next subsection, and updated permanently after the item is starting to have answers. Also, other statistical characteristic of the questions will be taken into consideration, such as the standard deviation, the point-biserial coefficient and

the discrimination index, in order to establish the validity and the good design of the items and the tests. In this matter, a perfectible design of an item and a test will be ensured at all times, by using specific principles of Learning Analytics (LA) and Item Analysis (IA) (Figs. 1, 2, 3 and 4).



**Fig. 1.** Visual representation of the model

In order to perform all of these requirements, the assessment model is based on the existence of two major components: the item collector and the test generator. The two components will be presented briefly in the next subsections, in order to determine the main guidelines of the entire model.

### 3.1 Item Collection (A-IC)

**A-IC Components.** The item collection (A-IC) component is the one responsible for the creation of a dataset of items and the continuous improvement and update of the items present in the dataset. The model is based on the existence of the item and its characteristics. In this matter, the item, which will be denoted further by  $q$  ( $id_q; st_q; dd_q; v_q; t_q, ans_q$ ) is an object formed of the next components:

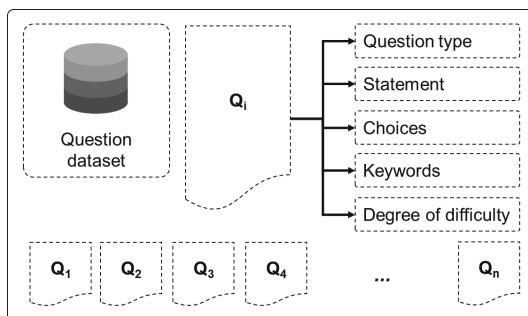
- the identification part of the item  $id_q$ , which consists of an identification particle of the item; in our model,  $id_q$  is considered to be a number and will play an important role in the implementation phase ( $id_q \geq 1, id_q \in \mathbb{N}$ );
- the statement  $st_q$ , formed of the phrase or a set of phrases that contain the requests of the item (either questions, statements, or exercises); the statement is an important part regarding the automated generation of tests;

- the set of keywords  $kw_q$ , of cardinal  $nr_{kw}$ , which contains the list of keywords that describe the item, either its category or the main topic of the item. A keyword  $kw_{qi}, i = 1, nr_{kw}$ , can be established by a human operator, or it can be generated using specific methods of language programming, such as Natural Language Programming (NLP);
- the degree of difficulty  $dd_q, dd_q \in [0, 1]$ ; which is firstly pre-established using the method presented in [15]. The degree of difficulty is then calculated as the ratio between the number of incorrect answers at the item and the total number of answers, being continuously updated by the answers given to the item;
- the item type  $t_q, t_q \in \{\text{'m'}, \text{'s'}, \text{'e'}\}$ , which shows the type of the item, whether it is multiple-choice based (m) or the answer is a textual one, given by the user, in case of short (s) and essay (e) types.
- choices set  $v_q$  (when  $t_q = \text{"m"}$ ), which can be formed of a list of two or more possible answers when the item type is multiple or is null when the item type is not ( $v_q \geq 1, v_q \in \mathbb{N}$ );
- the correct answer of the item  $ans_q$ , which contains the correct answer of the item; it has the form of:
  - a choice identifier, such as a letter (a, b, c, ...) or a number (1, 2, 3, ...), in case of multiple-choice items ( $t_q = \text{"m"}$ );
  - a real number, in case of numerical answers;
  - a text, in case of short or essay items ( $t_q \in \{\text{'s'}, \text{'e'}\}$ ).

Furthermore, an item q has additional characteristics related to the collection phase, such as:

- the university of collection,  $U_q$ , which contains the name of the university where the item has been collected;
- the course  $C_q$ , which contains the course from which the item was collected;
- the chapter or section  $Ch_q$ , which contains the chapter or the section from which the item was collected.

The three additional characteristics are useful in the context of the determination of the item source, as well as emphasizing the integration character of the model.



**Fig. 2.** Visual representation of an item  $q$

**A-IC Functionality.** The two important functionalities of the item collection component (A-IC) are the collection of the items and the improvement and update of the items.

The collection of items can be made by two main modalities: the human submission of items, which will be made by defining the characteristics of an item, possible in the implementation phase by the submission of a form within the interface of the implementation, or by automatic means, by extracting and formulating the questions and their characteristics using natural language programming (NLP) methods.

The improvement and update of the items are related to the Item Analysis integration within the model. This functionality is responsible for the performance of the item and collection and uses several item performance characteristics:

- the average score of the item  $m_q$ ;
- the number of correct answers  $cans_q$ ;
- the standard deviation  $sd_q$ , calculated as a normal standard deviation of the item using the specific formula described in the literature for a standard deviation for an item  $q$ .
- the item discrimination  $d_q$ ,  $d_q \in [-1, +1]$ , which determines for an item the amount of discrimination between the responses of the upper and the lower group and which is calculated using the specific formula described in the literature.

The item improvement and update functionality is useful for the good design of the items and the assessment tests, lower or higher values of the indicators showing a better or a poorer design of an item and permitting the different formulation or generation of the items or of the tests.

### 3.2 Test Generation (A-TG)

**A-TG Components.** An important component of the test generation model is a test T ( $S$ , DD), which is considered a set of items  $q_i$ ,  $i = 1, \|S\|$ , where  $S$  is the set of items that form the test,  $dd_i$  is the degree of difficulty of the item  $i$  and  $DD$  is the degree of difficulty of the test:

$$DD = \sum_{i=1}^S q_{dd_i} \quad (1)$$

**A-TG Functionality.** The A-TG component of the model is useful for the generation of assessment tests. Input data consists in:

- the desired course  $Co$ , which is the main topic of the desired assessment test topic;
- the set of universities  $U$ , which compiles the list of the  $nr_U$  universities in the consortium from which the items should be generated;
- the set of the chapters of the course  $ch$ , which comprises the list of the  $nr_{ch}$  chapters within the given course  $C$  from which the items should be generated;

- the number of items desired for each chapter within the  $ch$  list  $n$ , with the correspondence that the form the chapter  $ch_i$  is desired a  $n_i$  number of items for the assessment,  $i = 1, nr_{ch}$ .
- the minimum and maximum limits of the desired range for the degree of difficulty  $min$  and  $max$ .

The generation mechanism consists of the next steps:

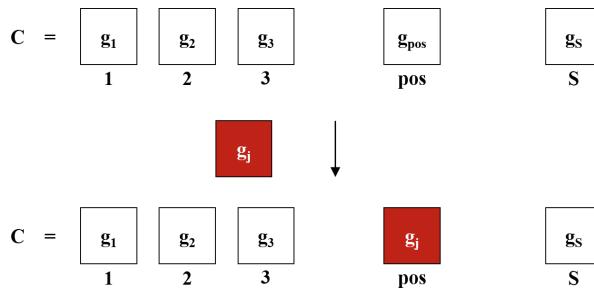
- Step 1. The items within the dataset are selected for each chapter from the set of chapters  $ch$ . Thus, for each chapter  $ch_i$  a list of  $n_i$  items  $L_i$  is pre-selected,  $i = 1, nr_{ch}$ .
- Step 2. Items are selected from the lists organized at Step 1 using a genetic algorithm or a method based on machine learning tools (ML). This is necessary, as the generated test must have a desired degree of difficulty. In this matter, the generated tests are optimized in reference to this requirement.
- Step 3 The final test is selected and output.

**A-TG Using Genetic Algorithms.** The genetic algorithm codifies items as genes within a chromosome, which will represent a test. In this matter, the next components will be needed:

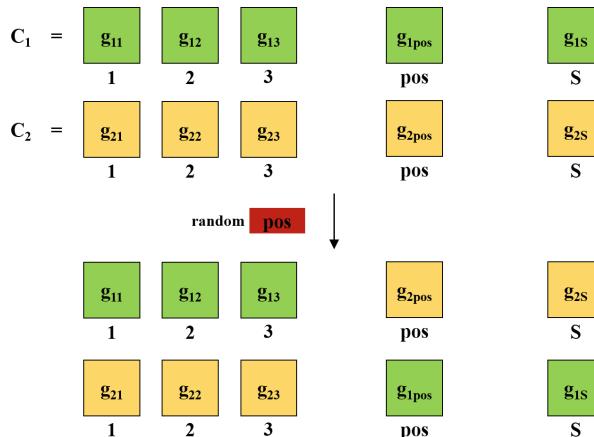
- a chromosome  $C$ , which codifies a test, formed of the next components:
  - order number  $id$ ,  $id \in \{0, \dots, NrI\}$ , where  $NrI$  is the total number of items within the database;
  - the gene set  $G_j = \{g_j | j \in \{1, \dots, S\}\}$ , where  $G = S$ ;  $j = 1, nr_T$ ,  $nr_T = n_1 + n_2 + \dots + n_i$ ,  $i = 1, n$ ;
  - the fitness function  $f$ , which calculates the closeness of the degree of difficulty of the chromosome  $C$  to the desired degree of difficulty  $DD$ . The fitness function calculation also gives the flexibility of choosing a lower or a higher level of difficulty, a higher value of  $min$  meaning a higher level of difficulty and a lower value of  $min$  combined with a lower value of  $max$  meaning a lower level of difficulty. The fitness function is defined as follows:

$$f(C) = \left| \sum_{i=1}^S q_{dd_i} - \frac{min + max}{2} \right| \quad (2)$$

- the genetic operators, the ones used in the presented model being:
  - the generation of a chromosome  $GenC$ , an operator that forms a new chromosome;
  - the mutation of a chromosome  $MutC$ , which can be defined as follows: for a given chromosome  $C$ , a random position within the chromosome  $pos$ ,  $pos = 1, S$ , and a randomly selected gene  $g_j$ ,  $j = 1, NrI$ , the mutation operation is defined as the shift of the gene  $g_{pos}$  with the gene  $g_j$ .

**Fig. 3.** Visual representation of  $MutC$  operator

- the crossover of a chromosome  $CsvC$ , which can be defined as follows: for two given chromosomes  $C_1$  and  $C_2$  and a randomly selected position within the two chromosomes  $pos$ , the first part of the chromosome  $C_1$  up to the gene  $g_{pos}$  is combined with the second part of the chromosome  $C_2$  from the gene  $g_{pos}$  to the end of  $C_2$  and the first part of the chromosome  $C_2$  up to the gene  $g_{pos}$  is combined with the second part of the chromosome  $C_1$  from the gene  $g_{pos}$  to the end of  $C_1$ , resulting in two new chromosomes  $C'_1$  and  $C'_2$ .

**Fig. 4.** Visual representation of  $CsvC$  operator

The methodology for the genetic algorithm comprises the next steps:

1. The input data is read (including the number of generations  $NG$ , which represents the number of generations of the genetic algorithm).
2. An initial population is generated using the repeated operator  $GenC$ . For each chromosome, the fitness value is calculated.

3. Genetic operators are applied:
  - (a) Crossover operator  $CsvC$  is applied for two chromosomes  $C_{k_1}$  and  $C_{k_2}$  chosen randomly,  $k = 1, Nrc$ .
  - (b) Mutation operator  $MutC$  is applied for a chromosome  $C_k$  chosen randomly.
  - (c) Fitness values are calculated for the newly-obtained chromosomes and the sort based on these values is made. The sorting operation is made ascending based on the fitness value because the optimal value of the fitness function is zero or tends to it for an optimal test that would respect the requirement related to the degree of difficulty.
  - (d) Substeps 3(a) and 3(b) are applied for  $NG$  times.
4. The first chromosome is selected as the best solution.

**A-TG Using ML-Based Tools.** In general, when faced with such challenges, you need to either get training data or find datasets that you can use for training, and then apply (transfer) the models learned on somewhat different data types.

This is sometimes referred to as transfer learning, although, this phrase also denotes instances where you take a trained model and retrain it on a new kind of data.

For Transfer learning, using a dataset meant for one use case, for training a model and using that model in a similar use case using a model trained on one task, and retraining it (fine tuning) on a different task.

## 4 Results and Discussions

In order to show the effectiveness of the automated generation of tests, we have established implementations of the two methods described in the previous section for the A-TG (Test Generation) component of the model. In this matter, two implementations, one for the genetic-based method and one for the ML-based method, were made starting from the same dataset of items. The dataset can be consulted online [3], in its CSV form. The implementation of the model presents the case of multiple-choice items ( $t_q = "m"$ ).

The implementation made for the genetic algorithm method was created using the PHP programming language, and the interface of the implementation was established as web-based. The initial data has the next values:

- the number of items in the dataset ( $N$ ) was 1000;
- the number of items desired in the test ( $S$ ) was 9;
- the desired course ( $Co$ ) was “ICT”;
- the set of chapters ( $ch$ ) was {“Microsoft Word”, “Microsoft Excel”, “Internet”};
- the number of questions desired for each chapter ( $n$ ) was {4, 3, 2};
- the minimum value of the degree of difficulty ( $min$ ) was 1 and the maximum value of the degree of difficulty ( $max$ ) was equal to  $S$ , meaning that the test is desired to have a medium degree of difficulty;

- the number of generations ( $NG$ ) was 100;
- the number of initial population was 10;
- the mutation rate (the proportion of the mutated chromosomes for each generation) was 0.1;
- the crossover rate (the proportion of the crossed-over chromosomes for each generation) was 0.9.

The results of a compiling situation of the genetic algorithm are presented in Table 1.

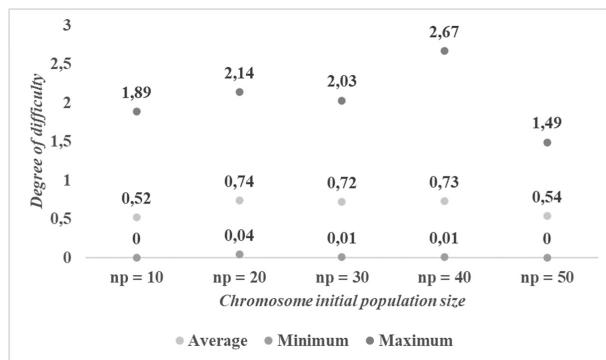
**Table 1.** Obtained values of the test for the input data

ID	Statement	University	Diff.
6	The Microsoft Word program is	UAI	0.34
10	What is the name of the direction of a page used for	UPIT	0.35
7	The extension of a file created in Word is:	UAI	0.82
9	The process of removing an unwanted part of an image	UAI	0.99
12	What is the term for unsolicited emails?	UPIT	0.59
66	What is a TCP/IP?	UPIT	0.95
14	URL means:	UPIT	0.16
23	The process of arranging the elements of a column....:	UPIT	0.65
23	Which function in Excel returns the average	UPIT	0.16

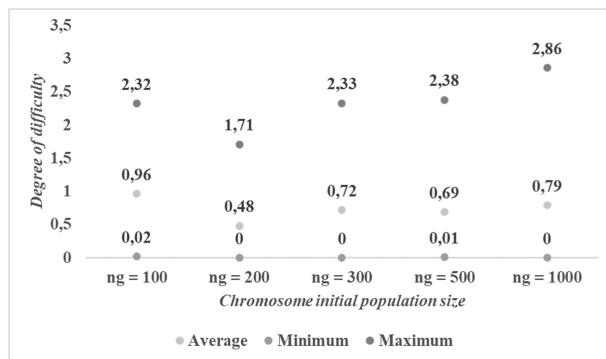
The final value of the fitness function for this test was situated at 0.01, which showed a degree of difficulty as close as possible to the desired degree of difficulty. The values presented in Table 1 show that the requirement related to the degree of difficulty is met by the presented modality of generating the test. We have also analyzed the performance of the genetic algorithm based on the requirement of the degree of difficulty related to the initial population size of the chromosomes and the number of generations. The results can be seen in the next figures.

Figure 5 shows that the best values for the closeness of the degree of difficulty can be found for a larger initial population size. In this matter, minimum and maximum values are the lowest for an initial population size of 50 from the ones taken into consideration.

Figure 6 shows that the best values for the closeness of the degree of difficulty can be found for a lower number of generations. In this matter, minimum and maximum values are the lowest for a number of 200 generations from the ones taken into consideration.



**Fig. 5.** The variation of the degree of difficulty based on the initial population size



**Fig. 6.** The variation of the degree of difficulty based on the number of generation NG

## 5 Conclusions

The presented model deals with the two considered aspects, specifically because it can be implemented in an integrated way within an academic consortium, and is related to the course, chapter, and degree of difficulty requirements of the user.

Future work consists of the development of the implementation of the model using ML-based methods, the automation of specific functionalities within the model (e.g., automated extraction of questions from a corpus of text) and the analysis of the results, which will lead to refinement and calibration of the assessment tool that was implemented and presented for the presented research.

## References

1. Beckmann, J., Birney, D., Goode, N.: Beyond psychometrics: the difference between difficult problem solving and complex problem solving. *Front. Psychol.* **8**, 1739 (2017). <https://doi.org/10.3389/fpsyg.2017.01739>

2. Blšták, M., Rozinajova, V.: Automatic question generation based on sentence structure analysis using machine learning approach. *Nat. Lang. Eng.* **28**, 1–31 (2021). <https://doi.org/10.1017/S1351324921000139>
3. Bold, N.: Item Dataset (2023). [https://github.com/nicolaebold/cim\\_get](https://github.com/nicolaebold/cim_get)
4. Campbell, D.: Task complexity: a review and analysis. *Acad. Manage. Rev.* **13**, 40–52 (1988). <https://doi.org/10.5465/AMR.1988.4306775>
5. Das, B., Majumder, M., Phadikar, S., Sekh, A.A.: Automatic question generation and answer assessment: a survey. *Res. Pract. Technol. Enhanced Learn.* **16**(1), 1–15 (2021). <https://doi.org/10.1186/s41039-021-00151-1>
6. Davis, L.: Job shop scheduling with genetic algorithms. In: Proceedings of the first International Conference on Genetic Algorithms and their Applications, pp. 136–140. Psychology Press (2014)
7. Gkatzia, D., Mahamood, S.: A snapshot of NLG evaluation practices 2005–2014 (2015). <https://doi.org/10.18653/v1/W15-4708>
8. 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, pp. 60–65. Association for Computational Linguistics, Doha, Qatar (2014). <https://doi.org/10.3115/v1/W14-4111>. <https://aclanthology.org/W14-4111>
9. Lee, C.K.H.: A review of applications of genetic algorithms in operations management. *Eng. Appl. Artif. Intell.* **76**, 1–12 (2018)
10. Mesmer, H., Cunningham, J., Hiebert, E.: Toward a theoretical model of text complexity for the early grades: learning from the past, anticipating the future. *Reading Res. Q.* **47**, 235–258 (2012). <https://doi.org/10.1002/rrq.019>
11. Mitkov, R., Ha, L., Varga, A., Rello, L.: Semantic similarity of distractors in multiple-choice tests, pp. 49–56 (2009). <https://doi.org/10.3115/1705415.1705422>
12. Mostow, J., Chen, W.: Generating instruction automatically for the reading strategy of self-questioning, vol. 200, pp. 465–472 (2009). <https://doi.org/10.3233/978-1-60750-028-5-465>
13. Nguyen, H.A., Bhat, S., Moore, S., Bier, N., Stamper, J.: Towards generalized methods for automatic question generation in educational domains. In: Hilliger, I., Munoz-Merino, P.J., De Laet, T., Ortega-Arranz, A., Farrell, T. (eds) Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption. ECTEL 2022. Lecture Notes in Computer Science, vol. 13450, pp. 272–284. Springer, Berlin (2022). [https://doi.org/10.1007/978-3-031-16290-9\\_20](https://doi.org/10.1007/978-3-031-16290-9_20)
14. Popescu, A.D., Bold, N., Nijloveanu, D.: A method based on genetic algorithms for generating assessment tests used for learning. *Polibits* **54**, 53–60 (2016). <https://doi.org/10.17562/PB-54-7>
15. Popescu, D.A., Bold, N.: The development of a web application for assessment by tests generated using genetic-based algorithms. *CEUR Workshop Proceedings* (2016)
16. Quirós, P., Lasheras, F.S.: Methodology for the projection of population pyramids based on monte Carlo simulation and genetic algorithms. *Appl. Intell.* 1–18 (2023)
17. Robinson, P.: Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Appl. Linguist.* **22**, 27–57 (2001). <https://doi.org/10.1093/applin/22.1.27>
18. Sancho-Asensio, A., et al.: Improving data partition schemes in smart grids via clustering data streams. *Expert Syst. Appl.* **41**(13), 5832–5842 (2014). <https://doi.org/10.1016/j.eswa.2014.03.035>
19. Sharma, S., Kumar, V.: Application of genetic algorithms in healthcare: a review. *Next Gener. Healthc. Inf.*, 75–86 (2022)



# An Approach to Automatic Flight Deviation Detection

Massimo Pietracupa, Hamdi Ben Abdessalem<sup>(✉)</sup>, and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal,  
Montréal H3C 3J7, Canada

{massimo.pietracupa, hamdi.ben.abdessalem}@umontreal.ca,  
frasson@iro.umontreal.ca

**Abstract.** Aircraft pilots are constantly undergoing situations where they must process significant amounts of data in very small periods of time, which may lead to mistakes. We have created a deviation detection system that is capable of auditing the cockpit in real time to detect actions that have been incorrectly added, omitted, or done out of sequence. This model assesses deviations based on hierarchical task data found in the Ontological Reference Model for Piloting Procedures, which contains knowledge-based reference procedures assembled by experts in the domain. Pilot actions are compared to ontology reference sequences using the Needleman-Wunsch algorithm for global alignment, as well as a Siamese LSTM network. An API that can be expanded to several Aerospace Simulators, as well as a Runner, was created to enable the Deviation Framework to connect to the XPlane simulator for real-time system monitoring. Synthetically created data containing sequence mutations were analyzed for testing. The results show that this framework is capable of detecting added, omitted, and out of sequence errors. Furthermore, the capabilities of Siamese networks are leveraged to understand the relation of certain sequence chain anomalies so that they can correctly be ignored (such as certain tasks that can be performed out of order from the reference sequence). These deviation assessments are capable of being run Real-Time (10 Hz) and have been clocked at 0.0179 s for one Takeoff sequence containing 23 actions. The evaluation results suggest that the proposed approach could be applied in aviation settings to help catch errors before harm is done.

**Keywords:** Virtual Reality · Simulation · Task control · Error Management · Human Computer Interaction · Neural Network

## 1 Introduction

Aircraft pilots are constantly undergoing situations in which they must process significant amounts of data in very small periods of time. The ability to detect whether tasks are being performed based on a set of guidelines can allow users the foresight to make adjustments early on in order to prevent failures, or even to increase their respective efficiency. An observational methodology study using a Line Operations Safety Audit (LOSA) (where expert observers are placed in the cockpit during normal flights to record threats) was

performed over a period of 15 years with confidential data collected on more than 3500 domestic and international flights. Supported by the Federal Aviation Administration and the International Civil Aviation Organization, their reports indicate that some of the most common types of errors by Aircraft Pilots include a conscious failure to adhere to procedures or regulations (such as performing checklists from memory) and following procedures with wrong executions (incorrectly entering data into the flight management computer) [1]: These are issues that can be significantly reduced with the help of a real-time deviation system, though these systems are challenging to create for complex systems such as an aircraft.

The complexity of these systems can be understood when we break down the process of performing a deviation assessment. The first steps of this analysis involve scrutinizing the tasks that are wrongly situated in the chain of activities, in comparison to a reference. Once those tasks have been identified, having an understanding of the relationship of these tasks with others can help us closer examine whether these activities are truly problematic when performed out of sequence. A basic illustration of this idea would be turning on the radio after starting a vehicle. This action would be recorded as an injection in a real time deviation system, even though we as users are aware that this particular action has little-to-no impact on the overall commencing process of the vehicle. Furthermore, tasks that are completed in tandem will not necessarily be recorded in the same order sequence at each capture. In such cases, these tasks should not be flagged incorrectly when out of place in relation to each other. For smaller systems, we can easily define a list of rules for each possible input with respect to others, though as the system becomes increasingly large and complex, this task grows exponentially. It becomes clear that a deviation detection method needs to understand the underlying relations of tasks in its system in order to be able to perform an audit.

As a result, the hypothesis that will be tested in this article are as follows:

**Hypothesis 1:** Is it possible to create a deviation model to determine problematic Pilot actions that are added, omitted, and out of sequence errors during flight that is capable of understanding the underlying relations of tasks?

**Hypothesis 2:** Will this model architecture be capable of performing its assessment in real-time (real time defined as simulation frequency of 10 Hz)?

In order to demonstrate these hypotheses, this paper will be organized as follows: First we will present a section for related works in the domain of automatic process deviation. Next, we will examine the methods implemented in this paper, and finally Sect. 4 will dive into the results.

## 2 Related Work

We can observe that deviation detection methods in process mining can be categorized as data-driven or knowledge-driven. As their names suggest, these can rely on developing process models from acquired data, or by having them created by domain experts. For data-driven methods, an average workflow is discovered and then used for comparison with individual activities. However, the average workflow is limited to a strictly

sequential view and does not account for concurrent activities or repetitive behaviors [2].

Process deviation approaches currently being used in aviation involve Flight Data Monitoring (FDM), where flight data is collected and analyzed to create rule-based models by human experts. Several machine learning techniques such as using ANN-based methodologies [3] and AutoEncoders [4] are trained on flight data and are able to predict anomalies by indicating a yes-or-no response. Unfortunately, they do not identify the issue at hand, even if it is process-specific.

Automatic process deviation has been in other domains as well. Using a model derived from actual behavior traces, Lu et al. performed deviation detection and identified common and uncommon behaviors in a business process [5]. Yang et al. approached a problem in deviation detection in complex medical processes, by combining both data-driven and knowledge-driven techniques [4]. Their approach was split up into different steps, dealing with hierarchical model structures. These model structures would have parent tasks containing a series of subtasks that would need to be performed. By manually recording the actions performed by medical staff through video footage, they were able to create action sequences. The first step in their process would be to annotate each action to its corresponding parent task. Once these new annotated sequences have been made, they would be aligned with a reference sequence in order to find the best possible split. From this point, they would process using a Conformance Checking algorithm readily available in the ProM software suite. This algorithm is capable of returning errors of commission, and omission, though this computation is very computationally demanding and would frequently cause computers to run out of memory (Dell, Windows 10 OS, Intel Xeon 3.7 GHz CPU, 48GB RAM), while also requiring all sequences to be in the form of Petri Nets (containing an activity trace of the same workflow). Further processing enabled Yang et al. to determine scheduling errors using a secondary algorithm, as the ProM conformance algorithm only dealt with commission and omission errors. Christov et al. performed deviation detections in processes for chemotherapy and blood transfusion through the creation of knowledge-driven workflows [6]. They simulated the error detections within the process by using synthetic activity traces (achieved by inserting artificial process errors).

Much like Yang et al., our approach tries to close the gaps of data- and knowledge-driven models by combining them together. Data-driven checking may better interpret observed data, at the cost of having sufficient samples. On the other hand, knowledge-driven methods have no dependence on data; though can become exponentially difficult to make as the complexity of the system increases. We can use machine learning aspects to understand nuanced relations between activities, while leveraging the workflow model to identify these problematic tasks.

### 3 Methods

#### 3.1 Ontological Reference Model for Piloting Procedures

When performing a deviation analysis, an action sequence needs to be compared to a baseline sequence for validation. The Ontological Reference Model for Piloting Procedures, which we will reference as the Aircraft Ontology, contains pilot tasks and actions

for procedures in various phases of flight [7]. This Aircraft Ontology focuses primarily on takeoff for the time being, and has been created by referencing standardized procedures from Airbus flight crew operating manual in circumstances including standard takeoff, loss of an engine, rejected takeoff, and other forms of takeoff. Current implementation of this ontology contains takeoff procedures as well as procedures for specific takeoff events. This ontology is provided as owl files, which can be loaded in python by using the owlready2 python package. Owlready2 is a library for working with ontology-oriented programming, which allows you to create, manipulate, and query ontologies in various formats such as OWL, RDF/XML, and RDFS. It provides a high-level API for ontology management and integrates with various reasoning engines for advanced querying and inference [8].

We can then use these ontologies to search for all Task Objects and Constraint objects. Once a Task object has been identified, we can query the hasNbConstraint and hasConstraint parameter to identify a task's execution conditions. A task's corresponding actions can also be queried by using the hasNbAction and hasAction parameter. These actions and constraints are utilized recursively and are compared with the simulation environment to validate whether a task can be executed. The ontology hierarchically organizes its objects into phases (e.g., takeoff), then tasks, and finally actions, which is a structure that can be effectively utilized in the deviation analysis.

### 3.2 Simulator API

The deviation analysis needs to be able to gather information from the simulated environment in order to perform its assessment. The simulator selected for this project is named XPlane, running on Windows, using the A320 Ultimate aircraft package. However, it is important to have the option of extending the usability of this deviation assessment beyond one single type of simulator. This design choice led us to creating an Application Programming Interface (API) that has the ability to accommodate different types of simulators. The API is developed in C++, to leverage its computational efficiency, and communicates with XPlane using a UDP connection. Access to the API is possible through any python application by simply using pip to install the *XPlaneApi* package. This architecture can then be used to retrieve data from XPlane, or the simulator of choice.

With the goal of having a multi-simulator capable system, the API is loaded with a *Subscriptions.yaml* file containing *labels* with their associated *tags*. Each tag contains the simulator specific data reference string, which represents a specific aircraft system within the simulator. As seen in Fig. 1, the *LeftThrustLever* label is the variable name that API clients will use to identify the system, while *sim/cockpit2/engine/actuators/throttle\_ration[0]* is the tag that XPlane uses to identify the thrust lever for engine 1. This system enables applications that communicate with the API to continuously reference the thrust lever for engine 1 as the label name, while the simulator specifications can be swapped out as needed in yaml files. The frequency in Hz at which a label can be polled from the simulation can also be selected. Increasing the frequency for highly used systems such as pedals, or sidesticks, while decreasing the frequency for infrequently used knobs and buttons will reduce the load on the system (if a large number of labels are being observed).

```
API > v1.0.6 > ! Subscriptions.yaml
129  LeftThrustLever:
130    tag: "sim/cockpit2/engine/actuators/throttle_ratio[0]"
131    frequency: 10
132    description: "Engine 1 Thrust value ratio between 0-1"
133  RightThrustLever:
134    tag: "sim/cockpit2/engine/actuators/throttle_ratio[1]"
135    frequency: 10
136    description: "Engine 2 Thrust value ratio between 0-1"
137  FMSDirection:
138    tag: "sim/cockpit/gyros/psi_ind_deg_m3"
139    frequency: 10
140    description: ""
141  Altitude:
142    tag: "sim/flightmodel/misc/h_ind"
143    frequency: 10
144    description: ""
```

**Fig. 1.** Labels, tags, and frequency in the Subscriptions.yaml for XPlane-API

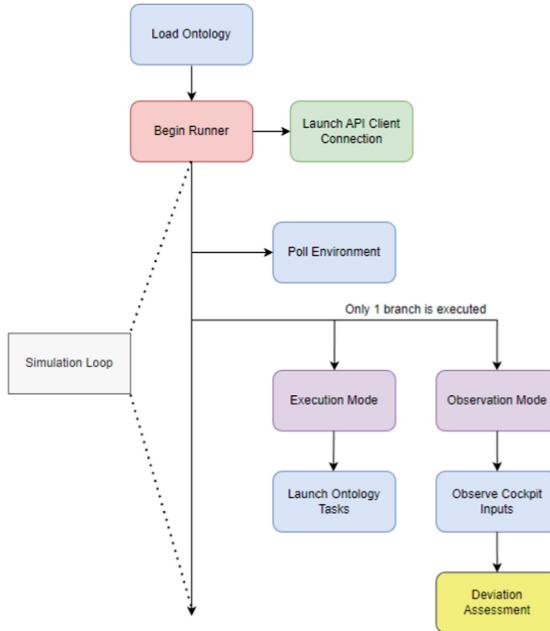
### 3.3 Runner

The Runner provides the main running utilities that the Deviation Model requires to gather information and remain in sync with the simulator. As illustrated in Fig. 3, upon initialization, the Runner will load the Aircraft Ontology dictionary (using the owlready2 ontology package) into a data object, that unpacks all required data, while also creating additional structures to facilitate data access during simulation time (such as querying for next tasks and previous tasks in sequences). Once all libraries are loaded, the Runner initializes the simulator API with its corresponding subscriptions file, and connects to it as a client. The successful connection to the API allows the Runner to commence the simulation loop, which runs at a frequency of 10 Hz.

Once the simulation loop begins, the Runner can run in **Execution** Mode or in **Observation** Mode. As the names suggest, execution mode will follow the sequence of takeoff tasks present in the Aircraft Ontology and execute them in the XPlane cockpit. Observation mode, on the other hand, will observe the cockpit and record all performed actions. Labels are assigned to the action sequences according to their predicted task, encoding in the form taskNumber+actionNumber. A completion flag is also appended onto these actions to ensure that they have been completed according to the task specification. For example, if the expected next task in the sequence requires both throttles to be engaged to 50% and the cockpit polls a 40% placement, then this task will be recorded as started – but not completed. These actions will then be passed onto the Deviation Model for sequence assessment, which can be seen as the yellow block in Fig. 2.

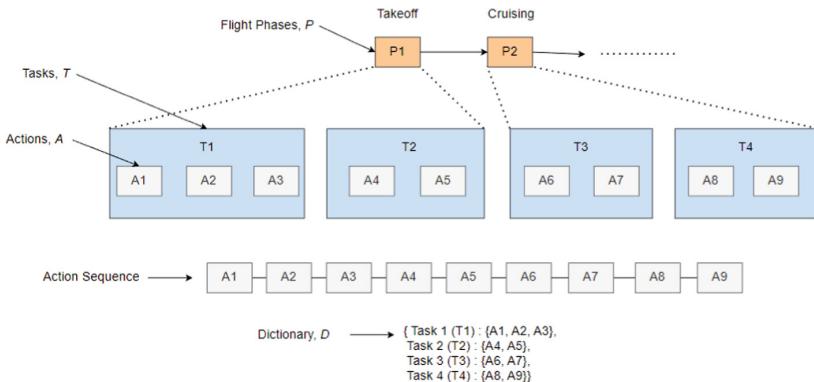
### 3.4 Deviation Model

The Deviation Model receives the action sequences from the cockpit, which we label as our Pilot Actions, and those from the Aircraft Ontology, referred to as our Reference Actions. We have divided the deviation assessment into four main steps as follows:



**Fig. 2.** XPlane Runner execution diagram.

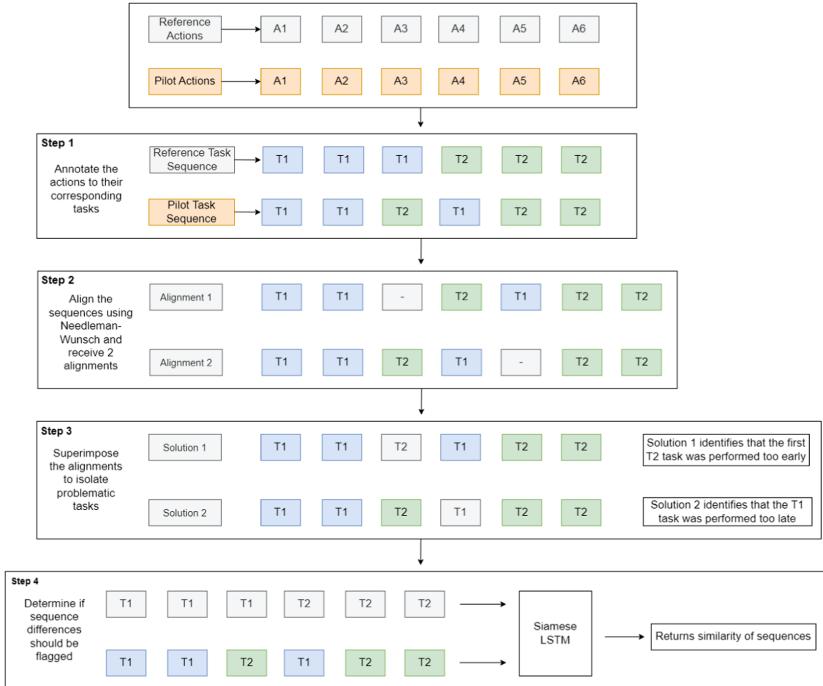
**Step 1:** As seen in Fig. 3, the tasks are hierarchically organized into phases, then tasks, and finally actions. Phases, such as takeoff, may contain several tasks, each with their associated action.



**Fig. 3.** Hierarchical organization of phases, tasks, and actions.

An example of a task includes setting the throttles, which contains two separate actions of setting the throttle of engine 1 and engine 2. The action sequence that is received by the Deviation Model contains actions that have been polled from the cockpit by the Runner. In this first step, actions are annotated using the Task Dictionary, which

contains a dictionary of tasks with their associated actions. This process can be seen in Fig. 5, as actions are annotated to their corresponding task (Pilot actions 1, 2 and 4 belonging to task 1, and actions 3, 5, and 6 belonging to task 2). Task actions should always be adjacent to each other, as they should be done in sequence with each other. After annotating the Reference and Pilot actions in Fig. 4 step 1, we can already start seeing where sequences are not fully lined up, which leads us to identifying problematic tasks.



**Fig. 4.** Deviation Model Overview

**Step 2:** These newly annotated sequences now need to be aligned together before being able to validate which tasks are out of place. The Needleman-Wunsch global alignment algorithm is a commonly used global alignment algorithm, which has been adopted in many sequence alignment software, and will be the choice of algorithm for this deviation assessment [9]. This algorithm is frequently employed in computational biology and bioinformatics for DNA or protein sequencing [10]. It utilizes dynamic programming to determine the ideal global alignment between two sequences.

**Step 3:** Once the alignments have been received from step 2, they are superimposed, and processed so we can extract problematic tasks along with their actions. Once sequence gaps are identified, they can easily be classified into added, omitted, and out of sequence. The algorithm for this process is described in Fig. 5.

**Step 4:** According to the Aircraft Ontology, certain tasks may be executed simultaneously, such as tasks which describe checking the takeoff N1, gradually releasing the

**Algorithm 1:** Deterministic Deviation Assessment

---

**Input:** Reference Sequence  $S_R$ , Pilot Sequence  $S_P$ , Action-Task dictionary  $D$

**Output:** Labeled Problematic Actions  $A_{Labeled}$

1. Annotate actions with Task labels:  $S_R^P = \text{Annotate}(S_R, D)$ ,  $S_P^P = \text{Annotate}(S_P, D)$
2. Insert anchor action at start of both sequences  $S_R^{P'}, S_P^{P'} = \text{Insert}(A_{Anchor}, \text{index} = 0)$
3. Perform Needleman-Wunsch alignment to receive 2 alignment sub-traces  $t_1, t_2$
4. Generate Gapped Sequence  $S_G$  (containing gaps wherever problems found) **for**  $i = 1$  **to**  $\text{len}(t_1)$  **do**
  - if**  $t_1[i] == t_2[i]$  **then**
    - store action into  $S_G$
  - else**
    - store a gap ‘ ’ into  $S_G$
5. Extract indices  $I_G$  of gapped entries in  $S_G$
6. Gather added actions  $A_{Added}$  from subtraction of unique entries from  $S_P$  and  $S_R$
7. Gather omitted actions  $A_{Omitted}$  from subtraction of unique entries from  $S_R$  and  $S_P$
8. Loop through each index in  $I_G$  to find out of place actions  $A_{OutOfPlace}$ 
**for**  $i \in I_G$  **do**
  - if**  $S_R[i + shift]$  in  $A_{Omitted}$  and  $S_P[i + shift] == S_R[i + shift]$  **then**
    - shift -= 1
    - continue
  - else**
    - if**  $S_P[i + shift]$  in  $S_R$  **then**
      - append action into  $A_{OutOfPlace}$
    - else**
      - continue
9. Label and combine  $A_{Added}$ ,  $A_{Omitted}$ ,  $A_{OutOfPlace}$  to receive  $A_{Labeled}$
10. return  $A_{Labeled}$

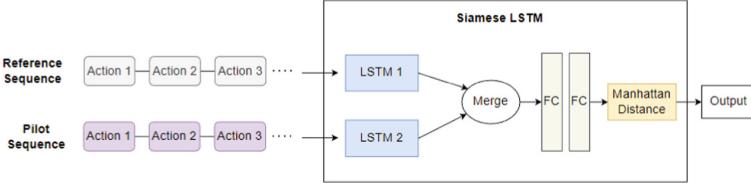
---

**Fig. 5.** Algorithmic description of deterministic deviation assessment

sidestick, and monitoring the primary flight display for the Takeoff phase. These tasks can be recorded in different orders continuously by the Runner due to their tandem nature. Siamese Long Short-Term Memory (LSTM) networks are a type of recurrent neural network architecture that is often used for tasks that involve measuring similarity or distance between two input sequences. It uses identical LSTM networks with shared weights to process each sequence, and then combines the resulting hidden states to make a prediction. Siamese LSTM networks have been applied to a wide range of tasks, and had a lot of success in identifying these hidden relations between sequences in several domains including text similarity, image matching, and speaker verification [11–13]. As illustrated in Fig. 7, the pilot and reference sequence are passed through a Siamese LSTM that we have trained on 6607 generated piloting sequences (correct and incorrect pairings). If the Siamese LSTM indicates that the sequence is the same as the reference sequence, then the flagged tasks will be recorded as acceptable deviations and will be placed in an ignored table for future iterations. The Siamese LSTM was trained using a learning rate of 0.001, LSTM input size of 300, hidden size of 100, and 2 layers (Fig. 6).

## 4 Results and Discussion

In order to assess a preliminary evaluation of the proposed deviation detection method, this model was applied to the takeoff procedure defined in the Aircraft Ontology. Sequences of steps were generated based on the Ontology’s specifications and were mutated to represent problematic process executions. This process is referred to as **Mutation Testing**, and is a software testing technique used to evaluate its effectiveness by



**Fig. 6.** Siamese LSTM structural overview

measuring its ability to detect changes or mutations. The types of mutations that have been introduced to the system are deletions (removal of an action), insertions (adding an action), substitution (replacing an action with another), and swap (switching the places of two actions).

We first performed a random walk through the Aircraft Ontology and retrieve a take-off sequence. Each individual action is then encoded with a unique identification number. The deviational model then consumed this sequence and outputs the problematic tasks with a label that indicates if the action is out of place (1), has been added (2), or omitted (3). An example of each type of mutation is performed on the action sequence that is input into the model and the model's respective outputs can be seen in Fig. 7. The outputs for these cases demonstrate that the deterministic assessment of the deviation model performs as expected.

		Deviation Model Outputs	Expected Outputs
1	Reference	[10011, 10022, 100323, 10033, 1005101, 100724, 10075, 10097, 101110, ...]	[] None
2	Insertion	[10011, 10086, 10022, 100323, 10033, 1005101, 100724, 10075, 10097, ...]	[• 10086 - Added]
3	Deletion	[10011, 10022, 100323, 10033, [empty], 100724, 10075, 10097, 101110, ...]	[• 1005101 - Removed]
4	Substitution	[10011, 10022, 100323, 10097, 1005101, 100724, 10075, 10097, 101110, ...]	[• 10097 - Moved • 10033 - Removed]
5	Swap	[10097, 10022, 100323, 10033, 1005101, 100724, 10075, 10011, 101110, ...]	[• 10097 - Moved • 10011 - Moved]
6	Layered Mutations	[100323, 10022, 10011, 10033, [empty], 100724, 10086, 10075, 102112, ...]	[• 100323 - Moved • 10011 - Moved • 102112 - Moved • 10086 - Added • 10097 - Removed • 1005101 - Removed]

**Fig. 7.** Mutation testing on synthetic action sequences.

The actions *102112*, *102011*, and *102314* in Fig. 8 represent respectively the actions of checking the takeoff N1, gradually releasing the sidestick, and monitoring the primary flight display mentioned previously. These are tasks that are meant to be completed at the same time, and should not be returned as problematic if they would be completed in an order that is not represented in the reference sequence. In order to replicate a scenario where these three actions are executed out of order, we perform a swap mutation on these actions alone. This swap mutation would trigger the deterministic deviation assessment and would return these tasks as problematic due to the fact that they no longer match the reference sequence. The Siamese LSTM, however, successfully captures the relationship between these tasks through trained examples and overrides the problematic labels by indicating that there are no deviations – because these tasks can be performed

in any order with respect to each other. We can therefore validate our first hypothesis by confirming that the deviation assessment is capable of detecting several types of problematic actions, while also understanding the relation between certain tasks and ignoring them if necessary.



**Fig. 8.** Siamese LSTM interpretation of tandem tasks.

Furthermore, deviation assessments using this method are extremely fast at 0.0179 s for an analysis sequence with length of 23 actions (for one takeoff sequence walk). When increasing the analysis length to 70 actions, the model increased in time to 0.0499 s (an average increase of 0.001 s per additional action), which is still well within the bound of a 10 Hz (0.1 s) simulation time requirement. Times were taken on Windows 10 computer using i7-5930K 3.50 GHz processor. If we compare these values to the 70.5 average sequence length at 5.98 s by Christov et al. for their real time medical deviation, we have achieved a substantial increase in assessment speed. However, computer specifications were not defined in their work and could affect cross comparisons. We can then validate our second hypothesis by confirming that this assessment can be completed at frequency of 10 Hz for a maximum-estimated 120 actions per phase.

#### 4.1 Limitations

Despite promising results, there are, however, still some limitations to this technique when it comes to training the Siamese LSTM. A sufficient amount of distributed data sequences is required to train this model, along with clever data engineering for known relational tasks.

Finally, this experimental evaluation is synthetic and has not been tested in real-time on pilot data. Even though the sequence chains are taken directly from the Aircraft Ontology, results might change if pilot cockpit behavior varies drastically, which may add complexity. These behavior changes could also introduce mutation types other than the ones that have been synthetically tested (deletions, insertions, substitutions, and swaps).

### 5 Conclusion

Based on the results, we conclude that it is certainly possible to perform real time deviation assessments of piloting tasks. We can see that it is also possible to combine both data- and knowledge-driven process-mining methods to capture an accurate representation of a correct process execution. The knowledge-driven models provided by the Aircraft Ontology serve as a structural backbone for this deviation model, though with

certain tasks required to be performed in tandem, a strict assessment cannot be directly performed. By utilizing the abilities of the Siamese LSTM, we can perform better assessments of correctly executed task sequences by leveraging previously accepted takeoff sequences. Through the synthetic generation of mutations on extracted Aircraft Ontology sequences, we have validated our first hypothesis by assessing deviations of different varieties, including added, omitted, and out of sequence errors, while also understanding relationships between tasks. These relational tasks will not be flagged if deemed in conformance by the Siamese LSTM.

**Acknowledgments.** We acknowledge the support from NSERC-Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

## References

1. Helmreich, R.L.: On error management: lessons from aviation. *BMJ* **320**(7237), 781–785 (2000). <https://doi.org/10.1136/bmj.320.7237.781>. PMID: 10720367; PMCID: PMC1117774
2. Mehmood, A., Maqsood, M., Bashir, M., Shuyuan, Y.: Deep Siamese convolution neural network for multi-class classification of Alzheimer disease. *Brain Sci.* **10**(2), 84 (2020)
3. Yang, et al.: Asymmetric Siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–18 (2022)
4. Yang, S., Sarcevic, A., Farneth, R.A., Chen, S., Marsic, O.Z.A.I., Burd, R.S.: An approach to automatic process deviation detection in a time-critical clinical process. *J. Biomed. Inform.* **85**, 155–167 (2018)
5. Frische, F., Mistrzyk, T., Lüdtke, A.: Detection of pilot errors in data by combining task modeling and model checking. In: Gross, T., et al. (eds.) *INTERACT 2009*. LNCS, vol. 5726, pp. 528–531. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-03655-2\\_59](https://doi.org/10.1007/978-3-642-03655-2_59)
6. Christov, S., Avrunin, G., Clarke, L.: Online deviation detection for medical. In: *AMIA Annual Symposium Proceedings* (2004)
7. Courtemanche, M.-A., Tato, A., Nkambou, R.: *Ontological Reference Model for Piloting Procedures*. Bucharest, Romania (2022)
8. W3C. Web Ontology Language (OWL) (2012)
9. Chao, J., Tang, F., Xu, L.: Developments in algorithms for sequence alignment: a review. *Biomolecules* **546** (2022)
10. Rose, J., Eisenmenger, F.: A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *J. Mol. Evol.* 340–354 (1991)
11. An, N.: Anomalies detection and tracking using Siamese neural networks. Master dissertation, Computer and Information Science, Auckland University of Technology, New Zealand (2019). <https://openrepository.aut.ac.nz/handle/10292/13525>. Accessed 17 Dec 2021
12. He, A., Luo, C., Tian, X., Zeng, W.: A twofold Siamese network for real-time object tracking. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018)
13. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. Department of Computer Science, University of Toronto, Canada (2015). <https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf>. Accessed 19 Dec 2021



# Automatic Learning of Piloting Behavior from Flight Data

Ange Tato<sup>(✉)</sup>, Roger Nkambou, and Gabrielle Tato

École de Technologie supérieure, Université du Québec à Montréal, Montreal, Canada  
[ange-adrienne.nyamen-tato@etsmtl.ca](mailto:ange-adrienne.nyamen-tato@etsmtl.ca)

**Abstract.** This paper proposes a model that will learn piloting actions using Recurrent Neural Networks. Neural networks, a sub-branch of machine learning, have been widely used for generalization problems. They can learn hidden patterns from data. They have been widely used to learn generalized human behaviors from data representing many behaviors. This research aims to extract meaningful behavior patterns from log data related to the piloting of an aircraft. The proposed learned model will predict the piloting actions to be performed over time (time series data) and during a given flight segment according to the takeoff conditions (inputs) such as environmental variables, contextual variables, and aircraft state. The resulting model aims to mimic the skills of an average human pilot in the same conditions for a specific takeoff task. The current model version has been evaluated using standard metrics, and experts' evaluation and results are promising.

**Keywords:** Piloting actions · Learned model · Neural Networks

## 1 Introduction

The airline industry is a crucial driver of global socioeconomic growth and is critically important as a catalyst for economic development, creating direct and indirect jobs, supporting tourism and local businesses, and stimulating foreign investment and international trade. Human pilots are trained to handle flight uncertainties or emergencies such as extreme weather conditions or system failure. In contrast, automatic flight control systems (AFCS/autopilot) are highly limited, and capable of performing minimal piloting tasks [1]. This work aims to develop a learned model of piloting behavior from log data. In many fields, such as data science and machine learning, expressing a relationship directly as an explicit and deterministic function is often impossible. From this perspective, the relationship is expressed by a dynamically varying implicit model learned from the available data or information. The model must be updated through training or learning to reflect the new information when the data or information changes. Such models are called learned models [16]. Developing a learned model is equivalent to extracting knowledge (patterns) from usage data. This model can predict the control action to be performed over time and during a

given flight segment based on takeoff conditions such as environmental variables, contextual variables, and aircraft state. The resulting model aims to mimic the skills of a human pilot in the same conditions for a specific takeoff task.

This paper presents a neural network model powered by flight data provided by one of the leading civil aviation actors (CAE<sup>1</sup>), which aims to automatically perform a takeoff as close as a human will do. The problem we want to resolve was raised by the industrial partners themselves (including Bombardier). They needed a framework to facilitate testing and simulations without involving human pilots all the time, hence the need for a learned model. The learned model will not replace a human pilot; it will be used in test situations where a specific type of human pilot will be needed. The proposed model can be seen as a synthetic pilot. An automatic pilot (autopilot) will act like a machine that differs from a synthetic pilot, closer to a human when performing actions. The synthetic pilot will help in simulation, where we want to evaluate how a specific type of pilot will behave on a specific task. It can also be an extension of the capabilities of modern autopilots and enable them to autonomously adapt their piloting to suit multiple scenarios ranging from standard to emergencies.

## 2 Related Work

Machine learning has been widely used in the air traffic domain, mainly for detecting abnormal events and operations during flights. In this sense, Li et al. [10], instead of using predefined criteria to identify risks, used a cluster-based anomaly detection to detect abnormal flights from flight data (365 B777 flights and 25,519 A320 flights), which can support domain experts in detecting anomalies and associated risks from routine airline operations. Sheridan et al. [14] did the same by using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect anomalous flights. On the contrary, Fernandez et al. [4] performed a descriptive analysis that employed clustering techniques to extract patterns and correlations within data and identify clusters of similar observations. Moreover, these clusters might reveal specific points as rare events isolated from the rest of the observations. They also conducted predictive analytics, which employed deep learning ANN and AutoEncoders, to detect these abnormal events. Their methodology relies on learning what “normal” observations look like since they usually are the majority of the cases. Afterward, if they process an abnormal flight, the model will return a high reconstruction error because of the deviation from the training data.

None of the above-mentioned work presented focused on the study of the pilot itself. The ability to learn and forecast pilot profile/behavior within a cockpit could become a powerful tool for aviation as well [5]. This research presents an approach to learning pilot behaviors from flight data. Few studies have been conducted to extract pilot behaviour from flight data. Camille Bodin, in her

---

<sup>1</sup> CAE is a major training partner of aviation professionals, airlines, large fleet operators, and aircraft manufacturers. It has the most prominent civil aviation network in the world.

work, identifies using machine learning techniques (logistic regression, support vector machines and neural networks, classification) flight maneuvers from real flight test data. These extracted maneuvers are labeled for actions performed (rolling roll, 360° rolls, etc.), not behavioral profiles [2]. “Numerical measurement of pilot gain has always been a dark art.” [6]. A study summarizes an approach for validating several potential pilot gain measurements based on pilot models and associated results in the time and frequency domain. The validation is based on data from a simulator study with 12 experimental test pilots and 12 operational pilots who varied their pilot gain/aggressiveness on command during a closed-loop tracking task [12]. This study presents potential measures of pilot gain, which are grouped into three classes: time domain-based measurements, frequency domain-based measurements, and pilot model-based measurements. These methods may require specially designed tasks, while in flight testing, this additional load may not be accepted if its sole purpose is to determine the pilot’s gain. Garcia Lorca et al. [5] proposed a characterization of pilot profiles through non-parametric classification of flight data using the K-means clustering technique. They used data from takeoffs and landings for feature identification and then classified various clustering centroids representing different profiles in pairs. The extracted flight profiles separated data from each flight based on specific score similarities.

Current studies focus on the automation of procedures in air traffic. Decision support systems, computer-based information acquisition, trajectory planning systems, high-level graphic display systems, and all advisory systems are considered automation components related to Next Generation airspace [17]. Several solutions have been developed for modeling and making predictions in airspace, especially predictions of pilot behavior. Yildiz et al. [17] proposed an approach that mathematically defines pilot goals in a complex system. These goals are, for example, “staying on the trajectory”, “not getting close to other aircraft,” or “having a smooth landing.” They then use game theory and machine learning to model the outcomes of the overall system based on these pilot goals together with other automation and environment variables. Christofer et al. [11] proposed a probabilistic method to predict pilot behavior in uncontrolled airspace. This prediction consists of trajectory prediction in the absence of intent pilot behavior. They use a navigation model representing the aircraft’s motion as a Markov chain of manoeuvre mode change points. Baomar et al. [1] proposed a system that observes and imitates a human pilot. They proposed a learning-by-imitation approach, which uses human pilots to demonstrate the task to be learned in a flight simulator while training datasets are captured from these demonstrations. The captured dataset is then fed to Artificial Neural Networks that automatically generate control models.

In this paper, behavior refers to how (the amplitude, the time, and the duration) a pilot performs actions. This perspective of the pilot behavior differs from what is done in the state of the art. Pilot dynamics research has shown that a pilot can change behavioral strategy when faced with a change in aircraft dynamics, resulting in different values of the pilot’s “gain” [8]. It is, therefore,

essential to consider the parameters external to the piloting actions (environmental parameters, parameters related to the aircraft, etc.) to best predict piloting behavior. As far as we know, none of the existing works proposed a learned model (that will mimic an average human pilot in the same condition) of piloting behavior based on flight data divided into flight segments to address the change in behavioral strategy.

### 3 Methodology

This work focused only on the takeoff phase of an Airbus 320. However, it can be extended to other flight phases. The takeoff task has been proposed by experts given (1) its critical nature in a flight and (2) the magnitude of the data. The idea is to eventually arrive at a model capable of learning and reproducing the behavior of a pilot for an entire flight. A concerted effort with experts in the field makes it possible to highlight ten segments of a takeoff phase: verification of thrust symmetry, takeoff thrust, takeoff for the climb phase, flap retraction, etc. The data (telemetry) used in our work is collected from CAE Level D full flight simulators and are takeoffs (normal and abnormal) of approximately 90 pilots. The data included plane orientation, speed, pilot inputs, and contextual data. The flight data (which are sequential) is associated with a temporal index (the tenth of a second) representing the variation of the parameters over time. We only considered data from flights performed under normal conditions for the training of our model.

#### 3.1 Step 1: Preprocessing

The collected data can have many irrelevant and missing parts. Cleaning is essential in machine learning, aiming to prepare data for learning. The cleaning process for this work consisted of data scaling and removing biased data, such as those whose values are out of a normal range. Variables measured at different scales do not contribute equally to model fit and model learning function. Thus, to deal with this potential problem, the standardization step is generally used before fitting the model. The raw data has been standardized using the scikit-learn python library. After the data preprocessing step is completed, Recurrent Neural Networks generate different learning models from the captured piloting data through offline training.

#### 3.2 Step 2: Training with Recurrent Neural Networks (RNNs)

RNNs have an inner architecture that allows them to process temporal information compared to other neural network architectures. They process an input sequence one element at a time, maintaining a “state vector” in their hidden units that implicitly contains information about the history of all past elements in the sequence [9]. Recently, the deep neural network (DNN) architecture and its long short-term memory (LSTM) variant are more accurate than traditional

statistical methods in modeling time series data” [13]. Given that the data are sequential, the choice was made towards recurrent neural networks, specifically LSTM. During the training phase, example data are provided to the LSTM, which produces one or more outputs:

- Inputs: contextual and environmental variables and the state of the aircraft
- Outputs: actions performed by the pilot (pitch stick, rudder pedal, rolls stick).

The model learns to recognize patterns and characteristics of different pilots when they perform the takeoff task. For each new data given as input, the model generates outputs representing piloting actions according to the takeoff segments. The system learns from its mistakes throughout trials, and its success rate improves until its predictions become accurate.

## 4 Experiments

Our proposed learned model consists of seven different control models learned from each of the seven out of ten segments learnable. In this paper, we will only focus on the two most important segments of the take-off phase:

- Segment 6: Rotation. When the indicated airspeed reaches VR (rotation speed), the pilot will pull back on the side stick to achieve a pitch rate of  $3^{\circ}/s$  towards a pitch angle of  $15^{\circ}$ .
- Segment 7: Lift off to Climb Phase. This segment begins as soon as the aircraft leaves the ground.

### 4.1 Segment 6: Rotation

Segment 4 continues to be applied. When the indicated airspeed reaches VR, the pilot will pull back on the side stick to achieve a pitch rate of  $3^{\circ}/s$  towards a pitch angle of  $15^{\circ}$ . The next segment begins as soon as the aircraft leaves the ground. For this segment, the pilot input is the Pitch Stick. The aircraft outputs are Pitch Angle, Pitch Rate, Airspeed Indicated, and VR.

### 4.2 Segment 7: Lift Off to Climb

Once airborne, the pilot will gradually release pressure on the rudder pedals (from segment 4). Wings level will be maintained via roll stick commands. Upon reaching  $15^{\circ}$  (Pitch Angle), the pilot will pitch to match the commanded flight director’s pitch. The flight director’s pitch command will change to assist the pilot in keeping an indicated airspeed of  $V2+10$ . The pilot will also roll to match the flight director’s command roll. For this segment, pilot inputs are Pitch Stick, Roll Stick, and rudder pedals. The aircraft outputs are Pitch Angle, Pitch Rate, Bank Angle, Heading, Airspeed, and Flight director pitch command. The Table 1 presents the reference scale of those specific values, which will later be used for validation.

**Table 1.** Reference intervals for segment 7 parameters (inputs and outputs)

URI	Range	Comments
ControlLateralPosition (Roll Stick)	0 to 100	Full left = 0 (bank left), Full right = 100 (bank right), Neutral (center released by pilot)=50.
ControlLongitudinalPosition (Pitch Stick)	0 to 100	Full Forward = 0, full aft = 100, Neutral (center released by pilot) = 50
ControlDirectionalPosition (Rudder Pedals)	0 to 100	Right Fwd = 0, Left Fwd = 100, Neutral (center released by pilot) = 50
StandardAircraft/BankAngle	-45 to 45	bank left = negative, bank right = positive
StandardAircraft/PitchAngle	-13 to 25	pitch up = positive, pitch down=negative

### 4.3 Implementation

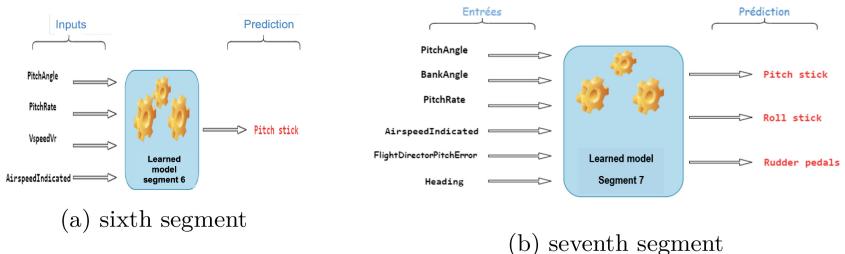
For each segment, we implemented a model that learns how a group of pilots behave in that specific segment. This learned model can be seen as a median of multiple observed behaviour. The model learns piloting behaviours and can predict the actions based on specific observations (inputs) as a human pilot will do. The implementation is divided into 3 phases:

**Data Collection.** Data were collected from the aircraft leaving the ground until the takeoff ended. The data are limited to the takeoff; the end of this phase is therefore determined. Data for the sixth segment are taken from 3s before the indicated speed (Airspeed Indicated) reaches the rotational speed (Vspeed Vr) until the altitude (Altitude Pressure) increases (beginning of segment 7: the aircraft leaves the floor). The duration for this sixth segment ranges from 17 to 107 tenths of a second, depending on the pilots. The seventh segment lasts from 105 to 908 tenths of seconds.

**Resampling the Data.** Pilots do the seventh segment with different duration. Also, the duration of each segment is not equal between all pilots. However, working with variable-length inputs and outputs is not possible with batch training, so a preprocessing step is needed. Since batch training is done due to the large size of the dataset, computational intensity of LSTM and efficiency purposes, LSTM requires that vectors passed as inputs have the same length or equal time series (the data must be recorded over an equal time interval). Inputs and outputs vectors are sequences of observed pilots' inputs and aircraft's outputs) We need to adjust the data so that all the pilots have a fixed duration for each segment of the takeoff phase. We padded the data with a specific value, but the resulting learned model could not generalize well. Instead of using padding techniques, we opt for a resampling technique. Resampling involves a change in the time-frequency of the time series data. The module used is *scipy.signal* function, which is related to waveform processing in Scipy library. *Scipy.signal.resample* uses the Fourier transform, which is based on the assumption that the signal is

periodic. After the sampling (oversampling and undersampling), the variation curve and the amplitude remain the same, except it is over a shorter or longer time (the average length of all duration, which is 535 tenths of a second for the seventh segment). Thus all segment 7 with durations below the average duration (535 ms) were oversampled to 535, and those below the average duration were undersampled. 13 out of 67 pilots were significantly resampled since their durations are far from the average duration, which is 535 ms. The rest of the data (54 pilots) were lightly resampled since they were close to the average duration.

**Machine Learning.** The learned model built for the sixth segment predicts the value of the pitch stick based on specific aircraft outputs. Figure 1a illustrates the inputs and outputs of the learned model for that segment. The inputs correspond to the aircraft outputs, which must be observed during the rotation step. The actions to execute or output the model correspond to the pilot input: pitch stick. The learned model for the seventh segment is a set of three machine-learning models predicting the Pitch stick, roll stick, and rudder pedals based on the inputs. Figure 1b illustrates the inputs and outputs of the learned model for that segment. The inputs correspond to the aircraft outputs, which must be observed during the lift-off to climb phase. The actions to execute or outputs of the models correspond to the three pilot inputs: pitch stick, roll stick, and rudder pedals. In other words, the learned model based on the current state of the aircraft must predict the value of the three pilot inputs, which are percentage, so that the action to perform (lift-off to climb) is well executed and correspond to the average behaviour of all pilots used for the training.

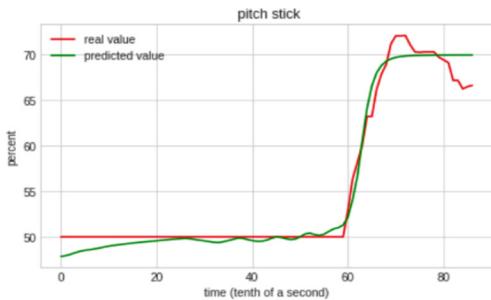


**Fig. 1.** Illustration of the sixth and seventh segments' learned model.

Before training, data are preprocessed and partitioned into training and test set. Next, the data are fed to the LSTM. The number of layers and other hyperparameters are chosen after a grid search to find the best combination. The method for choosing all models' parameters in this work is based on a rule-of-thumb [7], which indicates that neural networks process information more quickly with fewer layers, among other rules. This approach attempts to avoid overfitting caused by too many neurons and layers by having the number of hidden neurons less than or equal to twice the size of the input layer.

## 5 Results

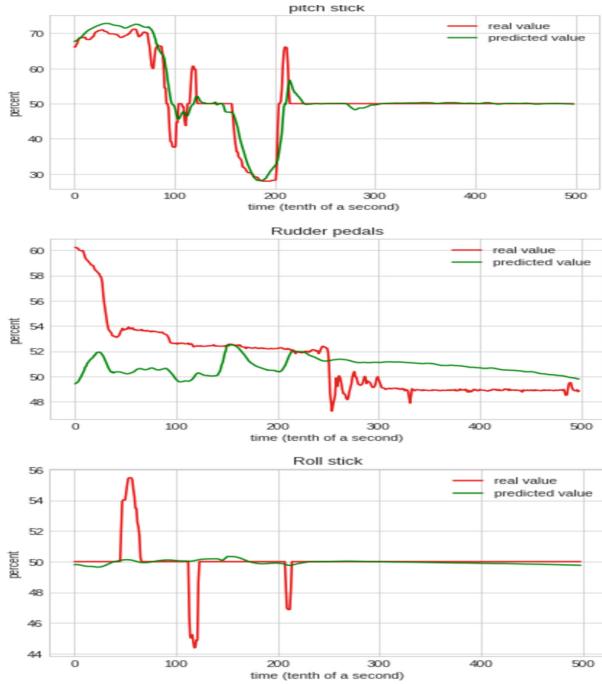
This section describes the results of the conducted tests on the learned model of segments six and seven. We recall that the learned model for the lift-off to climb phase is a concatenation of 3 LSTM where each predicts one of the outputs presented in Fig. 1b. The expected models should behave as close to a human pilot in the same conditions. Thus for our model to be acceptable, it should pass two tests: (1) predict actions that are feasible in the simulation environment (proper actions), and (2) predict actions that are close as possible to what a human pilot (belonging to the set of pilots used during the learning step) will do. The feasibility of the predicted actions is evaluated by comparing their predicted values to those in the reference model presented in Table 1. The Mean Squared Error (MSE) and Mean absolute error (MAE) metrics evaluate the difference between predicted and real data. The MSE and MAE are common ways to measure a regression model's prediction. The lower the MSE/MAE value, the better a model can predict values accurately. The MSE was used during the training, and the MAE was used only on the testing step. These metrics will be close to zero once the learned model is close as possible to a human pilot (belonging to the set of pilots used during the learning step).



**Fig. 2.** The learned model (predicted values) versus a real pilot on segment 6.

Figures 2 and 3 show concrete examples of prediction for segment 6 (prediction of pitch stick) and segment 7 (prediction of pitch stick, roll stick, and rudder pedals values), respectively for a given human pilot who did not participate in the training. Figure 4 show the inputs (aircraft outputs) that the same human pilot and learned models take into consideration before taking any actions. The x-axis of all four figures depicts the time in ms. The learned model does not need to be identical to what that specific human pilot does in the same conditions. Instead, the underlying behaviour learned by the learned model must be the same as the human pilot (does the learned model well capture the hidden pattern in the training data?).

We can see in Fig. 2 that the predictions obtained correspond to what the real pilot does. The MAE for the learned model of segment six is 2. An MAE of

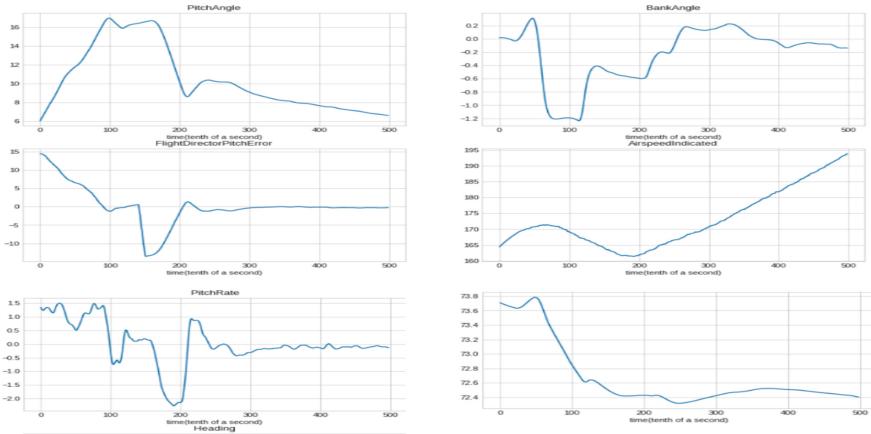


**Fig. 3.** The learned model (predicted values) versus a real pilot on segment 7.

2 corresponds to: if we take a real pitch stick equal to 50, the prediction will be  $50 \pm 2$ . The goal of the sixth segment is that when the indicated airspeed reaches VR, the pilot will pull back on the side stick (pitch stick) to achieve a pitch rate of  $3^\circ/\text{s}$  towards a pitch angle of  $15^\circ$ . We see that this is what the real pilot and the learned model are doing, as shown in Fig. 2, based on observed parameters. They both try to do the rotation.

Figure 3 shows predictions for segment seven. The predictions are almost consistent with the human pilot. We can see that the real pilot gradually releases the pressure on the rudder pedals. On the other hand, this relaxation is not observed at the level of the predictions. Instead, the learned model reflects the consistency of all pilots in the training set. The pilot tries to maintain the level of the wings via the roll stick. This consistency is observed more in the predicted data than in the actual action of the pilot. After discussion with the experts, they concluded that the real pilot taken for the example did not behave like the majority, hence the big difference in the roll stick. When the aircraft reaches  $15^\circ$  of pitch angle, the pilot adjusts the aircraft bank to follow the flight director's commands to help the pilot maintain an indicated airspeed, so pitch stick variations are observed at this effect. The pitches predicted here are very close to the actual pitch. As specified above, the goal is not to imitate the real pilot because the learned model is a kind of representation (at the level of the behavior) of

all the pilots used during the training. We used the MAE metric to confirm the observed results. All the real and predicted values were used to compute the MAE. We obtained an MAE equal to 3 for the pitch stick, 2.5 for the roll stick, and 2.5 for the rudder pedals. An MAE equals 2.5 for the pitch stick means there is an absolute difference of 2.5 between the real and the predicted pitch stick values for all the 67 pilots, knowing that the pitch stick value varies between 0 and 100. This is the best-obtained result considering that the predicted actions must be executable in the simulation environment. Since we did not yet test our learned model in the X-Plane environment, we focused on a preliminary test, verifying if the predicted values were within the expected reference values (tolerance intervals). We make the hypothesis that predicted actions with values that are within reference intervals are considered executable. Future work will consist of verifying our hypothesis.



**Fig. 4.** Segment 7 learned model inputs for the same pilot shown in Fig. 3

## 6 Conclusion and Future Work

We presented a learned model able to perform a takeoff task as a human pilot should. This learned model differs from an autopilot since the latter will behave as a robot/machine and has a predefined behavior. The model is developed based on takeoff segments identified by experts and flight data provided by one of our partners (CAE). The data consisted of 67 A320 takeoffs taking place under normal conditions. Because the acceptability of a solution is very difficult in the aviation domain, a plausible solution could have been a model predictive control (MPC) [3]. However, MPC is a process control method used to control a process while satisfying constraints. The goal of our research is not the automation of a task; instead, it is to control the plane as a particular type of human pilot

would do (ex: fly as a low gain or a high gain). The learned model aims to test how a human pilot would act (without involving a human pilot) in a different environment and under certain predefined test conditions.

The experiments showed the ability of the learned model to capture high-level tasks. The predictions obtained from the learned model pass the primary evaluation, which is: (1) prediction of actions (values of pitch stick, roll stick, rudder pedals) in segments six and seven that are feasible and (2) prediction of actions that are close to actions a human pilot will do. We hypothesized that action is feasible if its values are within reference intervals specified by experts. The learned models are intended to be used uniquely for simulation purposes, for example, in cases where we want to test how a specific type of pilot will behave in a specific situation. They will also be used for prediction purposes where we want to predict what a specific type of pilot will do in a specific situation, to modify a parameter in the environment accordingly, for example. The learned model will also allow the detection of possible errors and prevent them in a certain way at the cockpit level.

The next step of this study will consist of validating our hypothesis by performing simulations in the X-Plane simulator. The proposed model learns hidden behaviour patterns from all the pilots; thus, the obtained learned model behaves like the average pilot. One of our previous works extracted different classes of pilot profiles [15]. The second step in this work will be developing a learned model specific to pilots' different classes of behavior/profiles. For example, we will have a learned model for low gain pilots and a learned model for high gain pilots.

Finally, a future effort will focus on giving the learned model a more human-like behaviour by considering using a cognitive architecture such as ACT-R. In that sense, cognition is another important thing to consider, especially the workload (mental demand). We could think of a learned model that considers the mental state before predicting actions. Some of our team members are working on collecting a complete dataset, including mental states.

**Acknowledgement.** We acknowledge the support of CRIAQ, the Natural Sciences and Engineering Research Council of Canada (NSERC), CAE, Bombardier, and BMU.

## References

1. Baomar, H., Bentley, P.J.: An intelligent autopilot system that learns piloting skills from human pilots by imitation. In: 2016 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 1023–1031. IEEE (2016)
2. Bodin, C.: Automatic flight maneuver identification using machine learning methods (2020)
3. Camacho, E.F., Alba, C.B.: Model Predictive Control. Springer, Cham (2013)
4. Fernández, A., et al.: Flight data monitoring (FDM) unknown hazards detection during approach phase using clustering techniques and autoencoders. In: Proceedings of the Ninth SESAR Innovation Days, Athens, Greece, pp. 2–5 (2019)

5. Garcia Lorca, F., Gururajan, S., Belt, S.: Characterization of pilot profiles through non-parametric classification of flight data. In: AIAA Information Systems-AIAA Infotech@ Aerospace, p. 0914 (2017)
6. Gray, W.: A boundary avoidance tracking flight test technique for performance and workload assessment. In: Proceedings of the 38th Symposium of Society of Experimental Test Pilots, San Diego (2007)
7. Heaton, J.: Introduction to neural networks with Java. Heaton Research, Inc. (2008)
8. Klyde, D., Brenner, M., Thompson, P.: Wavelet-based time-varying human operator models. In: AIAA Atmospheric Flight Mechanics Conference and Exhibit, p. 4009 (2001)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
10. Li, L., Das, S., John Hansman, R., Palacios, R., Srivastava, A.N.: Analysis of flight data using clustering techniques for detecting abnormal operations. *J. Aerosp. inf. Syst.* **12**(9), 587–598 (2015)
11. Lowe, C., How, J.P.: Learning and predicting pilot behavior in uncontrolled airspace. In: AIAA Infotech@ Aerospace, p. 1199 (2015)
12. Niewind, I.: A new approach for the validation of potential pilot gain measures. In: EuroGNC 2013, 2nd CEAS Specialist Conference on Guidance, Navigation and Control (2013)
13. Sagheer, A., Kotb, M.: Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. *Sci. Rep.* **9**(1), 1–16 (2019)
14. Sheridan, K., Puranik, T.G., Mangortey, E., Pinon-Fischer, O.J., Kirby, M., Mavris, D.N.: An application of DBSCAN clustering for flight anomaly detection during the approach phase. In: AIAA Scitech 2020 Forum, p. 1851 (2020)
15. Tato, A., Nkambou, R., Nana Tato, G.J.: Towards adaptive coaching in piloting tasks: Learning pilots' behavioral profiles from flight data. In: Crossley, S., Popescu, E. (eds.) ITS 2022. LNCS, pp. 105–114. Springer, Cham (2022)
16. Yang, X.S.: Engineering Mathematics with Examples and Applications. Academic Press, Cambridge (2016)
17. Yildiz, Y., Agogino, A., Brat, G.: Predicting pilot behavior in medium-scale scenarios using game theory and reinforcement learning. *J. Guid. Control. Dyn.* **37**(4), 1335–1343 (2014)



# Towards Student Behaviour Simulation: A Decision Transformer Based Approach

Zhaoxing Li<sup>1</sup> , Lei Shi<sup>2</sup> , Yunzhan Zhou<sup>1</sup> , and Jindi Wang<sup>1</sup>

<sup>1</sup> Department of Computer Science, Durham University, Durham, UK  
`{zhaoxing.li2,yunzhan.zhou,jindi.wang}@durham.ac.uk`

<sup>2</sup> Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK  
`lei.shi@newcastle.ac.uk`

**Abstract.** With the rapid development of Artificial Intelligence (AI), an increasing number of Machine Learning (ML) technologies have been widely applied in many aspects of life. In the field of education, Intelligent Tutoring Systems (ITS) have also made significant advancements using these technologies. Developing different teaching strategies automatically, according to mined student characteristics and learning styles, could significantly enhance students' learning efficiency and performance. This requires the ITS to recommend different learning strategies and trajectories for different individual students. However, one of the greatest challenges is the scarcity of data sets providing interactions between students and ITS, for training such ITS. One promising solution to this challenge is to train "sim students", which imitate real students' behaviour while using the ITS. The simulated interactions between these *sim students* and the ITS can then be generated and used to train the ITS to provide personalised learning strategies and trajectories to *real students*. In this paper, we thus propose SimStu, built upon a Decision Transformer, to generate learning behavioural data to improve the performance of the trained ITS models. The experimental results suggest that our SimStu could model real students well in terms of action frequency distribution. Moreover, we evaluate SimStu in an emerging ITS technology, Knowledge Tracing. The results indicate that SimStu could improve the efficiency of ITS training.

**Keywords:** Student Modelling · Decision Transformer · Intelligent Tutoring Systems · Behavioural Patterns

## 1 Introduction

The recent COVID-19 has significantly impacted people's educational activities, which promoted the Intelligent Tutoring System (ITS) to achieve outstanding development. Data-intensive approaches have been proposed for ITS to improve the quality of education services [23]. However, these need to be powered by data-hungry machine learning models, whose performance relies heavily on the size of training data available [22]. Moreover, similar to the scarcity of labelled data in

many Artificial Intelligence (AI) fields, the shortage of student behavioural data has become one of the greatest challenges for ITS advancements [24]. Our work thus aims to tackle this challenge, by answering the following research question:

***How to create adequate high-fidelity and diverse simulated student behavioural data for training ITS?***

In this paper, we propose a Transformer-based approach based on our previous work [11]. The intuition of SimStu (shown in Fig. 1) is that after an ITS collects *a small amount of real* student behavioural data in the early stage development, it feeds the data into a generator, which produces *a large amount of simulated* student behavioural data. These simulated data can then be combined with the real student behavioural data, to train the ITS, thus improving ITS training. The generator, which we call “SimStu”, is built upon the Decision Transformer [3]. In the subsequent research, to train and evaluate our SimStu model, we used the EdNet dataset<sup>1</sup>, which is the largest student-ITS interaction benchmark dataset so far. Moreover, we improve the model’s performance by modifying the input and hyperparameters.

In this work, we proposed an upgrade version of the SimStu, which obtains better performance. The results suggest that our method could simulate real students well on the metrics of action distribution. In addition, we applied our method in real educational scenarios, Knowledge Tracing models. Knowledge Tracing (KT) is a method that predicts the student’s next action based on their previous ones. Many ITS use KT models’ prediction results to improve the student learning experience, e.g., giving recommendations for the next learning materials. Therefore, we applied our method’s generated data in the state-of-the-art KT models, i.e., SAINT, SSAKT and LTMTL, to evaluate the performance of our model. The experimental results show that our method could improve the KT model’s performance.

The main contributions of this paper lie in the following three aspects:

1. We propose a student learning behaviour simulation approach (SimStu) based on the Decision Transformer, aiming to provide adequate training data for ITS.
2. Our experiments demonstrate that a trained SimStu model can simulate real student behaviour well and outperform imitation learning based models.
3. We evaluate SimStu in a real ITS education scenario - applying SimStu in three state-of-the-art KT models (SAINT, SSAKT, LTMTL), and the results show that our approach could improve the performance of each KT model.

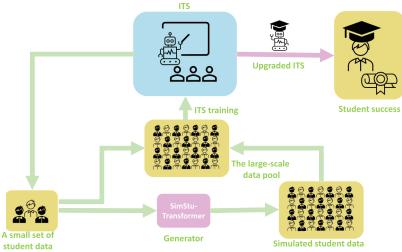
## 2 Related Work

### 2.1 Student Modelling

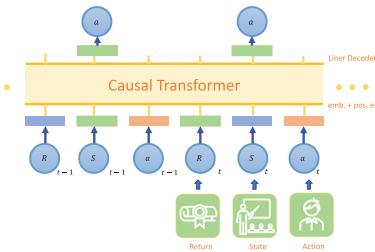
With increased attention to personalised learning, the traditional one-size-fits-all method can no longer satisfy user needs [2]. In offline scenarios, personalised

---

<sup>1</sup> <http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com>.



**Fig. 1.** The intuition for the proposed SimStu pipeline in ITS.



**Fig. 2.** SimStu architecture.

learning can be supported by teachers in various ways. For example, a teacher can gain valuable information about their students, by observing their learning process and interactions and then design the most suitable and beneficial learning strategy for them [7]. However, the lack of teacher-student interactions in online learning environments makes the personalisation process extremely difficult [1, 12]. In such online scenarios, student modelling can and has been applied, as a powerful tool to combat this issue [5]. Thereby, in the current study, we take advantage of the benefits of group-level student modelling and train our system using the learning data from a large number of individual students to learn the patterns of student learning in the system. This can then enable the system to recognise “optimal” learning behavioural patterns, which lead to the better student experience, performance, and learning results, as well as “poor” learning behavioural patterns, which may result in failure, thus recommending not only personalised but also optimal learning trajectories to the students, or providing a reminder of progressing to potential failure. To achieve this objective, it is crucial to have a decent quality and quantity of training data to feed to ‘data-hungry’ machine learning models.

## 2.2 Knowledge Tracing

Knowledge Tracing (KT) is a common method of personalising learning strategies for individual students. It predicts whether a student has the capability to master a new piece of knowledge, by tracing the student’s current knowledge state, which depends on past learning behaviour. The two major KT approaches are Bayesian Knowledge Tracing (BKT) and Deep Knowledge Tracing (DKT) [15].

BKT is a probabilistic method for student model generalisation [9]. It uses the Hidden Markov Model (HMM), to model their knowledge state as a set of binary parameters, each of which indicates whether a single Knowledge Concept (KC) has been understood or not [6]. DKT considers knowledge tracing as a sequence prediction problem. It uses Recurrent Neural Network (RNN) to model a student’s knowledge state in one summarised hidden vector [15]. DKT is powerful for capturing a complicated depiction of human learning. However, the

parameters of the DKT model are non-interpretable [10], which may result in students distrusting the system and teachers being unable to understand student behaviour. Additionally, when dealing with sparse data, DKT may encounter the problem of not generalising well [8]. The main limitation of BKT and DKT is that they both rely on a huge amount of students' historical learning data [13]. Different from BKT and DKT, our approach generates *simulated* student learning data, thus not relying on a huge amount of *historical* data, and more importantly, the simulated data can be visualised in statistical charts, showing student's learning behavioural patterns and thus being able to mitigate the KT model's limitation of non-interpretability.

### 2.3 Transformers

Transformers have risen to prominence in the field of deep learning in recent years, particularly in natural language processing and image generation tasks [14, 21]. A Transformer is an encoder-decoder Sequence2Sequence architecture to model sequential data, which consists of stacked self-attention layers.

Before self-attention was introduced, the best-in-class architecture was the seq2seq model [18], with an attention component from the decoder to align weights to input positions in the encoder, deciding how much information to retrieve from each position of inputs. Based on the Transformer architecture, Chen *et al.* [3] proposed the Decision Transformer, which abstracts the reinforcement learning problem, as a sequence modelling objective. The key in this algorithm is to generate actions based on *desired returns in the future*, rather than rewards in the past, and they proposed feeding a sequence of returns-to-go (sum of future rewards)  $\hat{R}_t = \sum_{t'=t}^T r_{t'}$  into the model. This model first learns a linear layer for each in returns-to-go, state, and action, to project them to the embedding dimension, followed by a layer normalisation. A time-step embedding is also learned and added to the tokens, which are then fed into a GPT [17] architecture, with the goal of generating future actions. Our proposed Sim is built upon this model. We feed the sequence of interactive data between students and the ITS into the Decision Transformer, to generate simulated student behaviour data.

## 3 Method

### 3.1 Architecture

The proposed SimStu is built upon the Decision Transformer [3] originally proposed by Chen *et al.* It consists of an encoder and a decoder and models the joint distribution of the sequence of student returns-to-go, states, and actions. Figure 2 illustrates the architecture. It separates student interactive trajectory sequences into two parts: one is used as the input embedding of the encoder, and the other is used as the output embedding of the decoder [21]. Then, the encoder takes the first part of the trajectory sequence embeddings as input and passes an output trajectory to the decoder. The decoder accepts a shifted embedding trajectory as input to produce the final output trajectory.

### 3.2 Dataset

The dataset used in our experiment is EdNet [4] - the largest student-ITS interaction benchmark dataset in the field of ITS. It contains more than 780K students' data extracted in South Korea over two years by a multi-platform ITS called SANTA<sup>2</sup>. EdNet consists of four hierarchical datasets, classified according to the number of interactions. We conducted our experiments based on EdNet-KT4, which includes problem-solving logs. Compared to KT-1 to KT-3, KT-4 provides the finest detailed interaction data, allowing access to specific features and tasks.

### 3.3 Trajectory Representation

The key desiderata of selecting the model features are to provide the algorithm with meaningful information to generate *the most likely trajectories*. We replaced the timestamps with the difference between the individual timestamps, i.e., the time between switching actions. The single timestamp could contain little information, and the time values in the UNIX system that generated them are large. We thus reduced the large UNIX time integers to small values, which also are more suitable for training. Furthermore, we removed from the modelling data types with very sparse data, where it is difficult for the Decision Transformer model to learn anything from the small number of values actually presented in the data. For instance, as *cursor\_time* is sparse, with a usual value of *NaN*, we removed *cursor\_time* from the data. *action\_type* is used to imitate students' behaviour, denoted by  $a$  in the Decision Transformer Trajectory  $\tau$ . *user\_answer*, denoted by  $R$ , is used for evaluating student performance, thus partitioning them into groups. We examined whether the student's answers (options of a, b, c, and d) matched with the correct answers: if yes, they received a positive reward of 1; and if no, they received a reward of 0. *item\_id* is used for evaluating the feasibility of the learning paths, which takes as the state of the student and is denoted by  $s$ . Due to the fact that *user\_id* does not affect or represent student behaviour, we chose to generate it randomly, after the SimStu generation procedure ended.

### 3.4 Experiments

The SimStu was implemented using the Pytorch framework and trained on an Nvidia RTX 3090 GPU. We used the Adam optimiser with a batch size of 64. We set Adam betas as (0.9, 0.95). The initial learning rate was 0.0006, and the dropout rate was 0.1. To evaluate the proposed SimStu, we conducted three experiments.

In the first experiment, we compared the simulated data generated by the SimStu model with the original data. More specifically, we examined the average number of actions for the generated and original data amongst the five student groups. Furthermore, we compared the similarity of the generated data

---

<sup>2</sup> <https://www.riiid.co.kr>.

and the original data using the Pearson product-moment correlation coefficient (PPMCC). PPMCC is a measure of the linear correlation between two variables [16]. A high PPMCC value in the experiment means a high correlation between the original and the generated data and thus indicates our SimStu can simulate student behaviours well.

In the second experiment, we compared our SimStu with Behaviour Cloning, an imitation learning based method proposed by Torabi [20]. We used RELU as the nonlinearity function, with a standard batch size of 64. We set the initial learning rate as 0.0001 and the dropout rate as 0.1. In this experiment, we compared the similarity of students’ “elapsed time” between generated and original data using PPMCC. As in the first experiment, a high PPMCC indicates a high simulation performance.

In the third experiment, we evaluated SimStu using three top-performance KT models selected from the Riiid Answer Correctness Prediction Competition on Kaggle<sup>3</sup>, which include SAINT, SSAKT, and LTMTI<sup>4</sup>. In the competition, Kaggle provides a dataset containing 2,500 student records to test models. Each student record contains the student’s sequence of discrete learning actions. We thus assume that 2,500 student record is sufficient for KT model training. Therefore, We selected five datasets that contained 500, 1,500, 2,000, and 2,500 student records, respectively. We fed these five datasets into the SimStu models, which then generated another five simulated datasets respectively. The generated data size was equal to the original data size. Lastly, we fed these five mixed datasets into the three KT models respectively to compare whether using SimStu could affect the performance of KT models. The metric we used here is AUC (Area Under Curve).

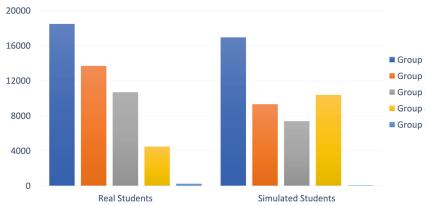
## 4 Result and Discussions

Figure 3 shows the results from the first experiment: the average number of actions performed by the real students (on the left) and by the simulated students (on the right), across all those five groups. This suggests that the distributions between the real student data and the simulated student data share some similar statistical characteristics, i.e., in both real and simulated scenarios: 1) the “very good student” group (Group 1) is the largest group, whilst the “very poor student” group (Group 5) is the smallest group; 2) the “good student” group (Group 2) and the “average student” group (Group 3) have similar sizes; and 3) both the “good student” group and the “average student” group are much smaller than the largest “very good student” group (Group 1), and 4) both the “good student” group (Group 2) and the “average student” group (Group 3) are much larger than the smallest “very poor student” group (Group 5). However, the only difference is that in the Real Students scenario (on the left), the “poor student” group (Group 4) is the second smallest group and smaller than both the “good student” group (Group 2) and the “average student” group (Group

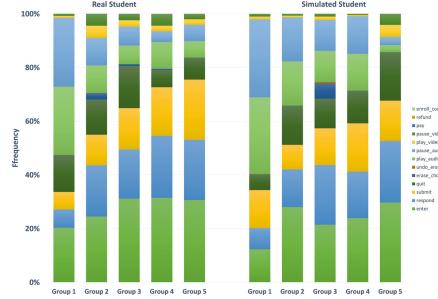
<sup>3</sup> <https://www.kaggle.com/code/datakite/riiid-answer-correctness>.

<sup>4</sup> <http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com>.

3), whilst in the Simulated Students scenario (on the right), the “poor student” group (Group 4) is the second largest group and larger than the “good student” group (Group 2) and the “average student” group (Group 3). Nevertheless, this result suggests that our SimStu model can generate student data similar to real student data.



**Fig. 3.** Action statistics of real student data (left), and simulated student data (right).



**Fig. 4.** Action frequency distribution of real student data (left), and simulated student data (right).

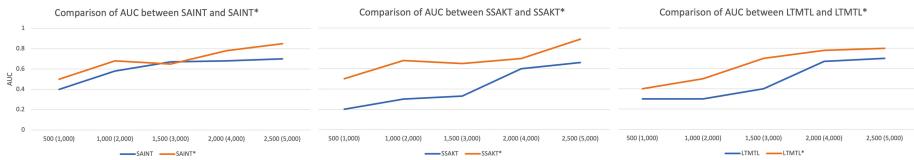
As Fig. 3 shows, the SimStu performed better in simulating the behaviour of students with higher grades (i.e. groups 1 (“very good”) to 3 (“average”)) than for lower grades students (i.e. groups 4 (“poor”) and 5 (“very poor”)). This is in line with the difference in the amount and the frequency of actions. The reason may be that students who study better generally spend a longer time interacting with the ITS, compared to students with relatively poor learning performance. This pattern makes many actions sparse and the causal relationship between actions weak, so the model cannot understand students’ behaviours well. To paraphrase Tolstoy’s words, “All good students may behave alike, but all poor performance students have their own reasons” [19].

Figure 4 shows the action frequency distribution of the real student data (on the left) and the simulated student data (on the right). This result shows that the simulated data generated by our SimStu is similar to the real data in major action frequencies. For example, the main actions of the generated data, such as *respond*, *enter*, *play\_audio*, and *submit*, have similar frequencies in each group. However, there are some differences in the actions that occur less frequently, such as *pay* and *undo\_erase\_choice*. The resulting PPMCC value of all actions is equal to 0.714, which suggests that the simulated student data and the real student data are 71.4% similar in the average distribution of actions. The result suggests that simulated data is statistically similar to real data.

In the second experiment, we fed the same training data and test data to the Behaviour Cloning model, which generated 600 students’ trajectories data (a total of 4,413,561 actions). The PPMCC value of the SimStu simulated data versus the real data is 0.762, while the PPMCC value of the Behaviour Cloning model simulated data versus the real data is 0.683. This indicates that the SimStu

simulated data is more similar to the real data, which suggests that our SimStu model outperforms the Behaviour Cloning model. This result may be due to the fact that when processing sequential student behavioural data, the student actions sequence context allows the SimStu to identify which policy can result in an action that promotes better learning states and improve training dynamics.

In the third experiment, we evaluated SimStu using the three state-of-the-art KT models. Figure 5 shows the pairwise AUC comparisons of these three KT models trained on the original datasets (SAINT, SSAKT and LTMTL, in blue) and trained on the mixed dataset (SAINT\*, SSAKT\* and LTMTL\*, in orange). In particular, the curves of SSAKT\* and LTMTL\* are constantly higher than those of SSAKT and LTMTL. The curve of SAINT\* is higher than that of SAINT in every dataset, except for the dataset size of 3,000. The results suggest that our method could improve the performance of KT models (AUC, in particular).



**Fig. 5.** Pairwise AUC comparisons of the three KT models trained on only original students' data (SAINT, SSAKT, LTMTL, in blue) and trained on the mixed dataset (SAINT\*, SSAKT\*, LTMTL\*, in orange). On the horizontal axis, 500, 1,000,...,2,500 indicate that the grey curve model uses the original dataset, and (1,000),(2,000),...,(5,000) indicate that the red curve model uses the mixed dataset. (Color figure online)

## 5 Conclusion

In this paper, we have proposed SimStu, a Transformer-based approach to simulating student behaviour, aiming to tackle the challenge of the scarcity of datasets for training ITS. We used the EdNet data to train the SimStu model, which generated learning behaviour data that could simulate the learning trajectories of different students. This method could be implemented in an ITS, such that ITS starts with collecting a small amount of student data, then uses our method to generate a large amount of simulated student data, mixes the original data and the generated data to build a new dataset, and finally uses the new dataset to train the ITS and improve its performance. The experimental results showed that SimStu could simulate the students' behaviour data well in terms of *action distribution*. Moreover, we evaluated SimStu by using three state-of-the-art KT models. The results indicated that our method could improve the performance of KT models.

## References

1. Bouhnik, D., Marcus, T.: Interaction in distance-learning courses. *J. Am. Soc. Inform. Sci. Technol.* **57**(3), 299–305 (2006)
2. Brusilovsky, P.: Adaptive hypermedia for education and training. *Adapt. Technol. Training Educ.* **46**, 46–68 (2012)
3. Chen, L., et al.: Decision transformer: reinforcement learning via sequence modeling. arXiv preprint [arXiv:2106.01345](https://arxiv.org/abs/2106.01345) (2021)
4. Choi, Y., et al.: EdNet: a large-scale hierarchical dataset in education. In: Bitten-court, I.I., Cukurova, M., Mulder, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 69–73. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-52240-7\\_13](https://doi.org/10.1007/978-3-030-52240-7_13)
5. Chrysafiadi, K., Virvou, M.: Student modeling approaches: a literature review for the last decade. *Expert Syst. Appl.* **40**(11), 4715–4729 (2013)
6. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**(4), 253–278 (1994)
7. Horntvedt, M.E.T., Nordsteien, A., Fermann, T., Severinsson, E.: Strategies for teaching evidence-based practice in nursing education: a thematic literature review. *BMC Med. Educ.* **18**(1), 1–11 (2018)
8. Kang, W.C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 197–206. IEEE (2018)
9. Kasurinen, J., Nikula, U.: Estimating programming knowledge with Bayesian knowledge tracing. *ACM SIGCSE Bull.* **41**(3), 313–317 (2009)
10. Khajah, M., Lindsey, R.V., Mozer, M.C.: How deep is knowledge tracing? arXiv preprint [arXiv:1604.02416](https://arxiv.org/abs/1604.02416) (2016)
11. Li, Z., Shi, L., Cristea, A., Zhou, Y., Xiao, C., Pan, Z.: Simstu-transformer: A transformer-based approach to simulating student behaviour. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Imitrova, V. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. AIED 2022. Lecture Notes in Computer Science, vol. 13356, pp. 348–351. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-11647-6\\_67](https://doi.org/10.1007/978-3-031-11647-6_67)
12. Li, Z., Shi, L., Cristea, A.I., Zhou, Y.: A survey of collaborative reinforcement learning: interactive methods and design patterns. In: Designing Interactive Systems Conference 2021, pp. 1579–1590 (2021)
13. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint [arXiv:1907.06837](https://arxiv.org/abs/1907.06837) (2019)
14. Parmar, N., et al.: Image transformer. In: International Conference on Machine Learning, pp. 4055–4064. PMLR (2018)
15. Piech, C., et al.: Deep knowledge tracing. arXiv preprint [arXiv:1506.05908](https://arxiv.org/abs/1506.05908) (2015)
16. Puth, M.T., Neuhäuser, M., Ruxton, G.D.: Effective use of Pearson's product-moment correlation coefficient. *Anim. Behav.* **93**, 183–189 (2014)
17. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
18. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
19. Tolstoj, L.N., Gerasimov, V.: Anna Karenina. BHB (1969)
20. Torabi, F., Warnell, G., Stone, P.: Behavioral cloning from observation. arXiv preprint [arXiv:1805.01954](https://arxiv.org/abs/1805.01954) (2018)

21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
22. Vincent-Lancrin, S., Van der Vlies, R.: Trustworthy artificial intelligence (AI) in education: promises and challenges (2020)
23. Weitekamp, D., Harpstead, E., Koedinger, K.R.: An interaction design for machine teaching to develop AI tutors. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2020)
24. Yang, S.J.: Guest editorial: precision education-a new challenge for AI in education. *J. Educ. Technol. Soc.* **24**(1), 105–108 (2021)

# **Augmented Intelligence and Metaverse**



# Mixed Reality Agents as Language Learning Tutors

Benedikt Hensen<sup>(✉)</sup> and Danylo Bekhter

Advanced Community Information Systems, Chair of Computer Science 5,  
RWTH Aachen University, Aachen, Germany  
[hensen@dbis.rwth-aachen.de](mailto:hensen@dbis.rwth-aachen.de)

**Abstract.** A challenge of language learning is to support its content with visual elements to add visual memories to textual and auditory recall. Whereas the depiction of nouns that portray objects is straightforward, verbs which describe actions require actors to clearly communicate the content. Here, mixed reality agents can guide the learner by demonstrating the action. In this paper, we investigate how such mixed reality agents can be utilized for language learning by combining them with the language learning method of total physical response. We compare the resulting open-source application to traditional flashcards in a between-subjects user study. There is no significant difference between the learner's performance with mixed reality and the flashcards but students preferred the mixed reality application with regard to engagement and technology acceptance. The study highlights the potential of accessible mixed reality agents which can be expanded into general tutoring systems with personalized tasks and individual feedback.

**Keywords:** Mixed Reality · Mixed Reality Agents · Tutoring · Language Learning

## 1 Introduction

The potential of mixed reality lies in the visual display of content in a mixed virtual and real environment. In learning, it has, therefore, mainly been applied for practical use cases where learners work with specific objects that can be augmented. Although more challenging to integrate, mixed reality also offers advantages for language learning as it allows students to experience multimodal content. Textual descriptions and auditory output can be supported by visual enactments. Here, objects like an apple are straightforward to visualize, e.g., as images or 3D models, but verbs which describe actions such as “peeling an apple” require moving imagery to understand them. A static image depicting a person with a knife next to an apple has the ambiguous meaning of either cutting or peeling the apple. Videos are one option to clarify this but they are usually created for a general audience and cannot be adjusted to the skill level

of individual learners due to their production effort. Therefore, mixed reality agents provide a suitable and flexible alternative to demonstrate actions. These computer-controlled entities which act in a mixed environment with real and virtual elements simulate real humans and their behavior [5]. The language content that they convey and their movement animations can be adjusted dynamically and they can be updated according to the student's learning progress. Moreover, mixed reality agents are interactive as they can react to input by the student and can be integrated into the student's familiar surroundings. Hence, in this paper, we investigate how the interactions between mixed reality agents and objects can be utilized to provide a visual and clear addition to language learning.

The remainder of this paper is structured as follows: In Sect. 2, we present related work about mixed reality language learning. Section 3 highlights the implemented resulting mixed reality agents learning application. In Sect. 4, we describe the user study which was conducted with this application and in Sect. 5, we discuss the results. The paper closes with a conclusion in Sect. 6.

## 2 Related Work

On a conceptual level, our work is related to methodologies of language learning. Here, one possible methodology is the Total Physical Response (TPR) as proposed by Asher [1]. The learner works with a tutor who combines the learning of phrases with physical movements. Hence, a pronounced sentence is acted out so that the learner is confronted with multisensory input. This leaves not just an auditory memory but also a visual element which the learner can associate with the word which can lead to a more reliable recall. Once the students are proficient with the vocabulary, they can also perform the action while repeating the phrase so that the memory also gets connected with motoric experiences.

One previous approach which combines TPR with augmented reality was presented by Martínez et al. [10]. It proposes multiple units which are supported by augmented reality to form the TPR language learning activities for children. In their design, augmented reality (AR) adds a layer on top of markers to tell a story using images and audio clips. The project highlights promising results regarding the learning performance, user acceptance and student motivation by combining TPR with augmented reality. Agents have previously been applied by Jia and Liu for language learning [8]. In their virtual reality application, they applied situated learning by combining the vocabulary with a virtual world. The agent mainly acts as a guide for students which provides tasks, offers help and gives feedback both using expressions and detailed explanations. In ARbis Pictus, vocabulary learning in AR was explored on the Microsoft HoloLens [7]. In the research project, objects in the real environment of the learner are labeled. Virtual annotations on the object display the word for the object in a given language. Their study showed the potential for visual language learning where no base language is required but instead, learners directly associate a visual impression with a word. Moreover, the results led to a better performance in AR than learning the vocabulary with flashcards. Similarly, Huynh et al. [6] also labeled

the learner's environment and used artificial intelligence for automatic object recognition on the Microsoft HoloLens. This way, vocabulary can be explored in the student's environment by studying day-to-day objects in the real surroundings. To realize interactions between agents and objects, the definitions of these interactions needs to be formulated. One approach here is to encode the interactions in the objects instead of central data in the agent's behavior module. This smart object approach was, e.g., proposed by Kallmann and Thalmann [9].

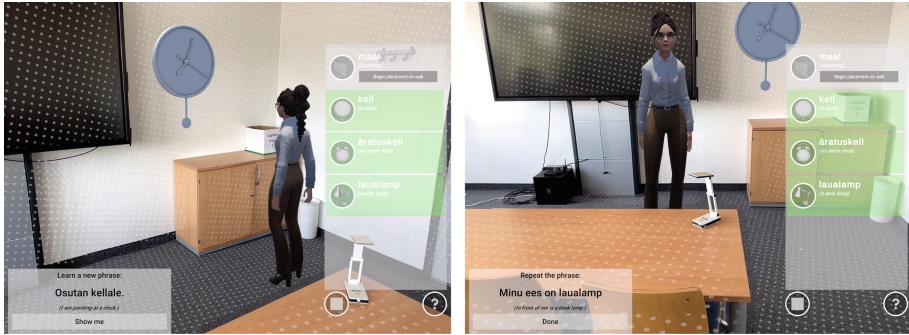
### 3 Mixed Reality Language Tutor

We applied the mixed reality agents in a language learning application. Here, the agent helps convey the meaning of vocabulary by performing the corresponding action on objects. The application runs on smartphones and tablets and is built using the Unity 3D engine. The agents are created with our open-source mixed reality agent framework which provides a foundation for building agents in 3D environments [4]. The mixed reality tutor project is also open-source and available on GitHub<sup>1</sup>. The learning application consists of two modes: The user starts in a setup mode where the device creates a scan of the room. The application loads a set of pre-defined vocabularies. In the camera feed on the screen, the user is asked to set up given objects. The list of objects reflects the set of vocabularies and the user can either choose to place down a virtual object or can mark a position in the real world as the present object. For instance, a virtual clock can be hung on the wall or the user can select a real clock in the room. After that, a series of relevant areas are set up which tell the agent how to move and interact with a given object. After this authoring phase, the user can switch to the learning mode. Here, the user places an agent on the floor. Then, the application displays vocabulary or a sentence on screen with a translation. After that, the user can switch to the mixed reality view where the agent will demonstrate the sentence according to the TPR methodology as shown in Fig. 1. For instance, the sentence "I am pointing at a clock" requires the user to set up a clock object in the authoring mode. In the learning mode, the sentence and its translation in a given language is displayed and once the user confirmed it, the agent starts walking over to the clock object, displays the sentence and points at the object. This way, the student gets a more intensive learning experience where translations are not just read and skipped over but where they are set into a context. When designing the learning content, it is possible to create a story by learning individual but connected sentences one after another. This story is then acted out by the agent which can help learners remember the content better and raises their interest in the learning activity.

The main technical challenge to realize this learning study is to provide a description language for the mixed reality agents so that they can interact with real and virtual objects in their environment. This description includes information about how the agent interacts with the object like instructions about its pose, animations to play and, if applicable, procedural adjustments to these

---

<sup>1</sup> <https://github.com/rwth-acis/ar-tutor>.



**Fig. 1.** Mixed Reality Tutor Language Learning Application

animations. For instance, if an agent should point at an object, the animation transforms the arm and hand into a pointing gesture. The procedural adjustment then applies post processing to align the arm with a target position so that the agent actually points at a given part of the object. Another major requirement is the extensibility of the system. To allow for this without changing the core definition of the agent, we decided to encode the interaction information in the object itself, instead of bundling it in the agent definition.

Apart from the usual properties which define 3D objects in space such as its position, rotation and scale, we added three properties to the smart objects which are relevant to the agents. These are its physical manifestation, the affected area and the interaction area. The physical manifestation is a value which defines whether the object is virtual or physical. This is relevant, e.g., regarding the abilities of the agents to manipulate the object. We are mainly concerned with virtual agents who can only change or pick up virtual items whereas physical objects have to be treated as being largely static. Physical manifestations of virtual objects also indicate that a 3D model needs to be rendered so that the object shows up for the user. The affected area is a region on the floor in which the agent can stand during the interaction. It contains a starting point where the agent is located when it starts the interaction and it defines a blueprint for areas where the agent will move to during the interaction. With this affected area, the application can ensure that there is enough space in the environment if the interaction, e.g., requires rather outgoing gestures or a large movement radius. In contrast to this, the interactive area is on the object itself and defines where the agent actually interacts with an object. For pointing actions, this area defines parts of the model where the agent should point to or if the agent is supposed to turn on a virtual computer, its on-off button becomes the interactive area.

By combining these three elements, each smart object contains at least one affordance. The affordance is an offered interaction which the agent can decide to perform. It also lists a series of requirements which the agent needs to fulfill in order to be able to perform the interaction. For instance, picking up an object requires the agent to possess at least one arm. Because of this additional

check, the system can also be expanded by arbitrary agents models who can then automatically determine which actions are possible for them.

To streamline the creation process of a smart environment by the user, the definition of the smart objects is divided into two phases: Developers of the application can create blueprints of new smart objects in which the affordances and their requirements are set up. With the blueprints, users can create instances in the environment by deciding whether a virtual or real physical manifestation is applied. For virtual objects, the object is created by placing its 3D model in the environment. For real objects, the point of the real object is marked. After that, the user can add the interactive and affected areas for both virtual and physical smart objects as prompted by the editing mode. This separation implies that the more technical tasks such as defining the actual interaction with the underlying animation is done at design time whereas the interaction affordances can be configured flexibly at runtime so that they fit into the real surroundings.

## 4 Evaluation in a Language Learning Use Case

We evaluated the impact that studying with mixed reality agents based on TPR has on vocabulary learning, motivation and effectiveness. This intervention was compared to a more traditional learning method of learning with flashcards.

### 4.1 Study Setup

The two learning methods are compared in a between-subjects design with a sample size of 20 ( $n = 20$ ) and ten persons per group. In both groups, subjects were provided with an iPad. In the experiment group, participants started the mixed reality agents applications. The control group was provided with a separate application without mixed reality and agents. Here, the same vocabulary was provided as flashcards on the screen in a Keynote presentation. As learning content, we chose vocabulary from an English-Estonian course. We opted for this combination in order to keep the number of possible participants high who know one of the languages but have had no prior contact to the other. Estonian is only spoken by 1.1 million native speakers and 7 million people speak a similar language from the Balto-Finnic language family. This language branch is also very distinct and shares few similarities to more commonly spoken languages. Before the evaluation, we also explicitly asked whether they know Estonian or another Balto-Finnic language and would have excluded participants who were familiar with it. This eliminates external influences on the learning outcomes due to previous knowledge of the language itself or a similar language from the same language branch. When separating the participants into the two groups, we balanced them so that their distribution of age groups, prior knowledge of mixed reality and interest in language learning were roughly similar. Both groups were given the same sentences in the same order so that learning effects from early vocabulary for the later sentences stay comparable. Before the evaluation, participants were given a brief introduction on how to use the applications using

learning content from an English-Spanish course. After that, the participants could start learning the vocabularies with actions around given objects. Both groups got 15 min to learn the sentences as they were shown in the respective applications. The experimental group was also granted five more minutes in order to account for them switching between the displayed sentences and watching the agent perform the corresponding actions. After this phase, the participants filled out a demographic questionnaire, the System Usability Scale (SUS) [3], and questions according to the Technology Acceptance Model (TAM) [11]. Moreover, the experiment group was asked more specific questions about the user experience of the mixed reality agent application and about the agent itself. After the user experience questionnaire, participants filled out a retention test which required them to translate ten English sentences to Estonian and vice versa. A second retention test was conducted on the next day in order to assess how much content was remembered in the short-term and intermediate-term memory compared to the long-term memory. During this 24 h period between the retention tests, we asked participants to not revise the content again.

## 4.2 Results

Out of the 20 participants, 11 are women and 9 are men. The average age in the experiment group is 33.3 with participants aged between 12 and 67. In the control group, the age mean is 32.8 and the ages are in the range between 22 and 60. All but one participant is fluent in English. For the remaining participant, a translation of the English parts into their native language was provided. For the mixed reality group, four participants reported no previous experience with mixed reality, four participants stated that they had rarely used mixed reality technology and the remaining two persons reported extensive developer experience with the technology. In the flash card group, two persons had never learnt with flash cards, five rarely use them and one participant had each rarely created new flash cards, regularly learnt with them and regularly created them. Hence, both groups have a similar previous experience with their assigned learning technology. None of the participants stated prior knowledge about the Estonian language.

In the SUS questionnaire, participants state their level of agreement with ten statements on a five-point Likert scale. We analyzed the trend of the responses by calculating an overall score between 0 and 100 for each questionnaire according to Brooke [3]. The average score in the experiment group is 78.25. The control group scored a mean of 77.5. According to Bangor et al. [2], this indicates a good usability which is above average. Despite the fact that the experiment group had little experience with mixed reality, using the mixed reality application seems to give a similar usability as using the flashcard app on the screen. The averages and standard deviations for the individual statements are listed in Table 1. Since the SUS formulates the statements alternatingly positive and negative, for odd numbered questions, a higher value indicates a better result and for questions with an even number, better results have lower values. The largest difference between averages can be seen in the statement SUS 1 about the intention to use the system frequently where the experiment group gave a 3.8 and the control

group a 2.9, indicating that the mixed reality system is preferred to be used often. We performed a two-tailed t-test with a significance level of 0.05 to check for differences between the two groups. We used the null hypothesis whether the means in the agreements to the statements are equal and searched for evidence to reject this null hypothesis. A significant difference with  $p = 0.048$  was found in the question SUS 4 whether support by a technical person is required. The experiment group stated more need for technical support which is to be expected since the technology was quite new for eight of the participants. For the TAM questionnaire, the results are listed in Table 2. A two-tailed t-test with the same parameters showed no significant differences between the means. In the custom questions listed in Table 3, there is a statistical difference regarding the means of the statements about the fun factor of the application and the visual enjoyment. In both instances, the mixed reality application was rated higher with  $p = 0.042$  and  $p = 0.027$  which supports the idea that visual elements of mixed reality create a pleasant and visually appealing learning experience.

**Table 1.** Results of the SUS Questionnaire for the Experiment Group (E.) and the Control Group (C.)

SUS Question According to Brooke [3]	E. mean	E. SD	C. mean	C. SD
SUS 1: Frequent use	3.8	0.87	2.9	1.37
SUS 2: Unnecessary complexity	1.2	0.4	1.4	0.49
SUS 3: Easy to use	3.7	1.27	4.3	1.19
SUS 4: Support of a technical person necessary	1.8	0.6	1.2	0.6
SUS 5: Well integrated functions	4.1	0.7	3.5	1.36
SUS 6: Inconsistency	1.4	0.49	1.5	0.92
SUS 7: Quickly to learn	4	1.1	4.5	0.67
SUS 8: Cumbersome usage	1.7	1.19	2.3	1.19
SUS 9: Confidence	3.7	1	3.8	1.08
SUS 10: Steep learning curve	1.9	0.7	1.6	1.2

The retention tests were rated according to their level of correctness with ten points for a completely correct translation, eight points for correct answers with spelling mistakes, five points for answers that still were close to the solution and two points for incomplete answers. The closeness of the answer was rated based on the same number of words, same starting letters and similar length of words. We subjected the means of the separate retention tests for both groups to t-tests to find statistical significant differences. However, the mean scores on both retention tests were statistically insignificant and no difference was found. For the retention test immediately after the experiment, a p-value of 0.424 and 0.541 for the two translation directions were determined. For the retention test on the second day, the p-values were 0.69 and 0.627 for the two translation directions.

**Table 2.** Results of the TAM Questionnaire for the Experiment Group (E.) and the Control Group (C.)

TAM Question According based on Park [11]	E. mean	E. SD	C. mean	C. SD
TAM 1: Perceived ease of use	4	1.18	3.6	1.2
TAM 2: Perceived usefulness	3.9	1.13	3.8	0.98
TAM 3: Attitude	4.2	1.17	3.3	1.62
TAM 4: Behavioral intention	3.4	1.02	2.7	1.27
TAM 5: Self-efficacy	4	1.55	4.4	1.02
TAM 6: System accessibility	3.7	1.27	3.9	1.37

**Table 3.** Results of the Custom Questionnaire for the Experiment Group (E.) and the Control Group (C.)

Custom Question for Learning Factors	E. mean	E. SD	C. mean	C. SD
C 1: Fun factor	4.3	1.27	2.9	1.45
C 2: Recommendation	3.9	1.14	3	1.48
C 3: Visual enjoyment	4.3	1	3	1.26
C 4: Engagement	4.3	1.27	3	1.61
C 5: Motivation	3.6	1.11	2.8	1.54
C 6: Mobility	4.2	1.25	3.2	1.4
C 7: Automation	2.9	1.64	2.2	0.75

**Table 4.** Results of the Questionnaire about the Agents

Question	Mean	SD
A 1: Clear and understandable	4.3	1
A 2: Controllable	3.9	0.94
A 3: Clear actions	3.9	1.14
A 4: Animations helpful	3.6	1.11
A 5: Agent distracting	1.89	1.2
A 6: Presence of the agent	3.8	1.17
A 7: Friendliness	3.75	0.97
A 8: Uncanny appearance	2	1.26
A 9: Robotic movements	3	0.77

The experiment group's results about the agent questionnaire in Table 4 show that the acceptance and perceived performance of the agent is high. The participants felt in control of the agent and regarded the demonstrations as fully understandable and helpful. They were not of the opinion that the agent was distracting as they averaged an agreement value of 1.89 out of 5 on the Likert scale. In addition, the agent appeared not to be uncanny for the learners.

## 5 Discussion

The results of the user study show a promising direction for using mixed reality agents in language learning. An agent adds a beneficial visual component to language learning to enhance the experience. We see potential to elaborate the agent into a more versatile mentor which can give expressive and individual feedback about the learning activities. In addition, the mentor can also be adapted to other learning scenarios. During the development of the learning application, the virtual agents framework proved to be a useful tool as it provides a configurable foundation for the creation of the agent. The agent can quickly be personalized by switching out the model of the agent and by authoring the underlying behavior. Another positive aspect is that the mixed reality approach using smartphones and tablets is accessible for a broad number of students.

We decided to author the learning unit with the Estonian language so that the group of potential participants stays large since it is unlikely that they have previous knowledge. One interesting aspect for future research would be to repeat the user study with other languages which have more learners worldwide. One factor here that can influence the performance is whether the learner has a personal motivation to study this specific language, e.g., for work, vacations or personal relationships. Here, an adjustable difficulty can account for previous knowledge and keeps the student interested in the learning experience. Another possible factor which may affect the outcome are possible influences between the short-term retention test directly after using the application and the retention test on the next day. Testing the learnt content can cause a separate learning effect which could affect the result of the second retention test [12]. One possibility to quantify this effect is to conduct another study with separate and larger groups that only participate in the immediate test or the long-term test.

## 6 Conclusion

In this paper, we presented an open-source mixed reality agent system for language learning using the total physical response methodology. The agents act out the content of sentences to visually underline descriptions of actions in a given language. To realize this, we added a smart object definition to our agents framework which allows them to interact with both virtual and real objects in the learner's surroundings. The resulting mixed reality learning application was evaluated with 20 participants in an English-Estonian language learning use case. The between-subjects design compares the mixed reality agents to a traditional

flashcards method. The results indicate a good usability of the developed system and there are no significant differences in the learning performances. However, participants showed a preference for the mixed reality agents with regard to motivation and technology acceptance.

**Acknowledgements.** We thank the German Federal Ministry of Education and Research for their support within the project “Personalisierte Kompetenzentwicklung und hybrides KI-Mentoring” (tech4compKI; id: 16DHB2213).

## References

1. Asher, J.J.: The total physical response approach to second language learning. *Mod. Lang. J.* **53**(1), 3 (1969). <https://doi.org/10.2307/322091>
2. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Stud.* **4**(3), 114–123 (2009). <http://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/>
3. Brooke, J.: SUS: A Quick and Dirty Usability Scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) *Usability Evaluation in Industry*, pp. 189–194. Taylor & Francis (1996), <https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke>
4. Hensen, B., Bekhter, D., Blehm, D., Meinberger, S., Klamma, R.: Mixed reality agents for automated mentoring processes. In: de Paolis, L.T., Arpaia, P., Sacco, M. (eds.) *XR Salento 2022*. LNCS, vol. 13446, pp. 3–16. Springer Nature Switzerland, Cham (2022). [https://doi.org/10.1007/978-3-031-15553-6\\_1](https://doi.org/10.1007/978-3-031-15553-6_1)
5. Holz, T., Campbell, A.G., O'Hare, G.M., Stafford, J.W., Martin, A., Dragone, M.: MiRA—mixed reality agents. *Int. J. Hum.-Comput. Stud.* **69**(4), 251–268 (2011). <https://doi.org/10.1016/j.ijhcs.2010.10.001>
6. Huynh, B., Orlosky, J., Hollerer, T.: In-situ labeling for augmented reality language learning. In: Teather, R., Itoh, Y., Gabbard, J. (eds.) *Proceedings, 26th IEEE Conference on Virtual Reality and 3D User Interfaces*, pp. 1606–1611. IEEE, Piscataway, NJ (2019). <https://doi.org/10.1109/VR.2019.8798358>
7. Ibrahim, A., Huynh, B., Downey, J., Hollerer, T., Chun, D., O'Donovan, J.: ARbis pictus: a study of vocabulary learning with augmented reality. *IEEE Trans. Visual Comput. Graphics* **24**(11), 2867–2874 (2018). <https://doi.org/10.1109/TVCG.2018.2868568>
8. Jia, T., Liu, Y.: Words in kitchen: an instance of leveraging virtual reality technology to learn vocabulary. In: *Adjunct proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality*, pp. 150–155. IEEE Computer Society, Conference Publishing Services, Los Alamitos, California (2019). <https://doi.org/10.1109/ISMAR-Adjunct.2019.00-59>
9. Kallmann, M., Thalmann, D.: Modeling behaviors of interactive objects for real-time virtual environments. *J. Vis. Lang. Comput.* **13**(2), 177–195 (2002). <https://doi.org/10.1006/jvlc.2001.0229>
10. Martinez, A.A., Benito, J.R.L., Gonzalez, E.A., Ajuria, E.B.: An experience of the application of Augmented Reality to learn English in Infant Education. In: *2017 International Symposium on Computers in Education (SIIE)*, pp. 1–6. IEEE (2017). <https://doi.org/10.1109/SIIE.2017.8259645>

11. Park, S.Y.: An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *J. Educ. Technol. Soc.* **12**(3), 150–162 (2009). <http://www.jstor.org/stable/jeductechsoci.12.3.150>
12. Roediger, H.L., Karpicke, J.D.: Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* **17**(3), 249–255 (2006). <https://doi.org/10.1111/j.1467-9280.2006.01693.x>



# Metaverse and Virtual Environment to Improve Attention Deficit Hyperactivity Disorder (ADHD) Students' Learning

Azza Mohamed<sup>(✉)</sup> Mai Zohiar, and Ibrahim Ismail

Department of Information Technology, Khawarizmi International College, Al Ain, UAE  
azza.mohammed@khawarizmi.com

**Abstract.** In recent years, educators in information technology departments and teachers caring for students with disabilities, such as autism and attention deficit hyperactivity disorder, have worked together to create a more inclusive environment for students with disabilities, and have focused their attention on the Metaverse. Mark Zuckerberg, a notable business magnate, recently stated that Facebook will be rebranded as Meta and will have its own virtual world, also known as the Metaverse. As a result, interest in the Metaverse has grown considerably. Although educators may be aware of the Metaverse as a concept, its potential cannot be dismissed once it has been deployed. As the existence of the Metaverse is still in its infancy, it is crucial to develop a comprehensive understanding of its benefits, applications, and effects on the education system once it becomes an official part of the education experience in the current world. This article aims to shed light on how the Metaverse will transform education and help students suffering from diseases such as ADHD and an autism spectrum disorder.

**Keywords:** Metaverse in education · ADHD learner · Metaverse

## 1 Introduction

As the world inches closer to digital takeover, its undeniable that innovations in computer science will inevitably shift the current perspective on technology used in educational sectors, with the Metaverse being the lead force for the change in how education system will function in the future due to major funding and recent advertisement campaigns. The concept of the Metaverse has been around for a long time, first mentioned in science fiction novels such as Snow Crash. Crash [1, 2] is a two-dimensional or three-dimensional virtual environment in which users can interact with and manipulate objects, as well as communicate in real time with users in remote physical locations for work or play [3]. There are numerous well-established and well-known Metaverses and virtual worlds, such as Second Life, World of Warcraft, and Roblox., however, when the Metaverse project was announced by Mark Zuckerberg the interest in what the Metaverse could become, especially in education, has increased significantly.

Although the Metaverse has existed for plenty of time, its implementation for schools and overall teaching experience has been quite minimal. If executed properly, the possibilities that the Metaverse has in the education sector are nearly endless. Applying a two-dimensional virtual world will not be effective for studies as it may seem as a distraction for students, it may not be accessible for impaired students and sometimes it doesn't follow the study plan organized by the tutors. Whereas a Metaverse with augmented and virtual reality technologies could fulfil the students' and instructors' needs, by customizing and manipulating the teaching environment. Teachers could have a platform that will meet all their requirements that would help students with disorders such as ADHD and autism as they find it difficult to focus and pay attention, by gamifying the class. With the help of the Metaverse, students alike will have less trouble paying attention to what is being taught.

Gamification has long been used in educational settings to engage students, promote learning, motivate action, and solve problems using, game thinking, aesthetics, and game mechanics [4]. By making studies seem like a game it entices students to give their attention to what is happening around them, participate and perform their best. Inclusion of the Metaverse platform that uses Virtual Reality (VR) will take gamification to an entirely different level, considering that users will be able to interact with objects and create their own environments in real-time with minimal delay, ensuring maximum cooperation and paid attention.

The following is how the research is structured. Section 2 contains a literature review on ADHD and the Metaverse. Section 3 of the paper includes the research methodology, study hypothesis, and objectives. Section 4 includes the results. Section 5 summarizes the study's findings.

## 2 Related Work

This section is divided into four subsections and is based on literary research discussing the topic of the Metaverse. Section 2.1 reviews previous literature on Attention Deficit and Hyperactivity Disorder (ADHD), Sect. 2.2 looks over the Metaverse in education, Sect. 2.3 discusses the Metaverse's features in education and Sect. 2.4 involves the Metaverse's Educational Challenges.

### 2.1 Attention Deficit and Hyperactivity Disorder

Deficit Hyperactivity Disorder (ADHD), a chronic neurodevelopmental disorder that affects a person's ability to pay attention, control impulsivity and regulate motor activity, prevents affected children from follow directions at home, being organized and taking turns in conversations, because of the difficulty in regulating the frontal lobe, which controls executive functions [9, 10–14].

Children with ADHD also have difficulties with inhibitory control, working memory, and cognitive flexibility, leading to poor educational performance [15, 16]. Treatments such as parent training [17, 18], pharmacotherapy, and behavioral therapy are commonly used for rehabilitation [19–21].

First-line treatment options for ADHD have several drawbacks, including drug side effects, and high cost [22, 23]. Recently, researchers have explored the use of virtual reality (VR) applications and augmented class simulators as adjunctive treatments for ADHD [24]. These advanced systems simulate 3D environments and provide individuals with a fully immersive and realistic experience, which is more effective and safer than conventional treatments, allowing users to learn new skills and better understand real world problems [25–28].

Furthermore, the mechanism for creating an “embodied simulation” is shared by VR and the brain [29]. Virtual reality (VR) is an embodied technology because it provides a sense of presence and immersion while also allowing for interaction. This contributes to the development of context-aware movement, behavior, and emotion [30].

According to recent research, virtual reality (VR) aids in the transition of learning into everyday life while maintaining high relevance and motivation [31]. Both virtual and augmented reality are viewed positively by children because they improve the quality of experience, motivation, and learning [32].

A VR-based tool has been developed to aid in the diagnosis and treatment of children with ADHD. Children appear more open to the engaging environments VR can create and patient compliance improves, implying that the Metaverse can be a helpful educational tool for students with ADHD [33].

## 2.2 The Metaverse in Education

The Metaverse has numerous potential applications in education, such as in teacher training, and science. Along with allowing students to perceive something that requires long-term practice and encouraging students to try to create or explore something that they usually cannot afford to do or is too risky in the real world [34].

## 2.3 Metaverse Features in Education

Table 1 compares classroom learning and learning in the Metaverse. Metaverse learning has the potential to replace classroom learning. Below, the following properties and their meanings are explained:

- The time and place for students to participate in class

Teachers and students traditionally meet in physical classrooms at set times [35]. On the one hand, the Metaverse has the potential to become a near-universal educational space because smart wearable devices keep teachers and students consistently immersed in educational environments [36, 37]. Consider a teacher who has been invited to a conference out of town, he/she can instruct the students to participate in classes in the Metaverse world using wearable devices. High-capacity networks enable us to move seamlessly between the real world and the Metaverse, bridging the gap between formal and informal learning environments, allowing teachers to experiment with various methods of learning, such as asynchronous and synchronous learning [38].

- Interaction in learning

In physical classrooms, learners interact primarily through face-to-face communication [39]. Interactions in the Metaverse typically involve embodied multisensory participation using interaction technologies such as sensors, BCI, VR, AR, and XR, which can highly stimulate and motivate senses, improving learners' interests and performance outcomes [40, 41].

- Educational resources

Traditional learning resources such as printed textbooks tend to be in relatively static formats [42]. Metaverse-based education visualizes and decentralizes learning resources so that learners can interact with them [43]. Traditional classroom sessions may include lectures about the Earth using maps. However, with the help of augmented reality, students can see the virtual earth from all angles. Learners can also collaborate with peers and teachers to create or edit learning resources.

- Evaluation of learning

Because it is difficult to collect learning data in traditional learning environments (through testing, for example), teachers often evaluate learners collectively based on their learning outcomes, creating educational inequalities [44]. Using learning logs and analytics based on both formative and aggregate data, teachers can more thoroughly assess learner performance in the Metaverse.

- The metaverse facilitates learning through virtual experiments.

Learning through virtual experiments is an integral part of science (physics and life sciences) curricula [45]. In the Metaverse, learners can participate in a variety of virtual experiments that include real-time interaction with interaction technology.

However, this is limited by limited funding for materials and infrastructure, and by physical limitations due to COVID-19, prioritizing hands-on experimentation over theoretical learning [46, 47]. Given the Metaverse's educational properties, the following potential applications are listed [48]:

1. Supporting experiments that may be dangerous, irreversible, or toxic in the real world, especially explosive experiments.
2. Conditions for experiments and scientific phenomena that cannot be replicated in the real world, for example experiments that need to be done in a vacuum.
3. Real-world experiments with relatively high costs are supported.
4. Experiments requiring long-term observation and recording in the real world, such as an experiment that requires students to observe and record the entire life cycle of an insect.

## 2.4 Metaverse's Educational Challenges

Although the Metaverse opens new avenues for education, we should be aware of its challenges; this section focuses on two educational challenges faced by the Metaverse.

- Technology and tools

To enable both teachers and students to teleport into the Metaverse, well-designed, low-cost smart wearable devices are required [49]. According to [50], advances in

technology have accelerated advances in hardware (e.g. HMDs), but they still need to improve. Users may experience symptoms (such as cyber sickness) after wearing the wearable for an extended period, posing a potential health risk [51]. Additionally, the current equipment is prohibitively expensive for most people [18]. The Metaverse has the potential to evolve if more sophisticated solutions to integrate the Metaverse into education are developed, through for example 5G networks and digital twins.

- Identity and interpersonal communication

In the Metaverse, the digital identities can directly reflect a user's ego when engaging in various activities, perhaps blurring the line between virtual and real identities. As a result, students in communities, schools, and families must distinguish between the physical and virtual worlds and rationally approach the Metaverse.

### 3 Study Hypothesis and Objectives

The study aimed to find effective ways to help students with ADHD learn. Traditional classroom and learning resources in static formats are not suitable for ADHD students [50]. The study will investigate the effectiveness of Metaverse-based education, as well as other virtual reality platforms. Comparative research is essential when implementing new technology in education. The discussion will focus on two questions.

- Can learners outperform traditional classrooms in the Metaverse/virtual world?
- What aspects of learning performance have improved significantly compared to other environments?

We believe the above issues deserve further investigation.

#### 3.1 Gamification and Education

Gamification involves using game mechanics and elements in non-gaming contexts to increase learning motivation. It includes features such as users, scores, challenges, leader boards, and levels. Incorporating gamification technology into education can increase engagement and productivity [8]. The text suggests using gamification for children with ADHD and lists six game parts such as:

1. Brain

The brain is very important in games because stimulating the brain can affect a child's her IQ and her EQ of children, particularly children with Attention (ADHD), and the logic game method is expected to help them become better.

2. Think

Thinking helps children develop their learning skills and learning through games, which will greatly aid in the learning process.

3. body

The body is the most important part of the game; with body movement on the body, the bass body exercises without even realizing it, so with one game, the brain and body move at the same time.

#### 4. Move

Along with movement, thinking is one of the most important parts of the game, and its diversity allows players to practice how their mind and body work, so players are not passively silent during the game. Gamification is one of the learning media that uses media games, and it is expected that users can learn quickly and easily handle it. With the ease of education, it is hoped that knowledge will be effortless and enjoyable, especially for hyperactive children.

### 3.2 Learning 3D Application Software and Virtual Environment

The use of 3D software for creating computer-generated images in various applications. While traditional education methods use printed materials and physical models, they do not address 3D visual space applications [7]. However, using 3D software for teaching on desktop computers can pose some issues, and spatial ability has no effect on learning outcomes in a desktop VR learning environment. Learning 3D animation software is a visual experience [6]. We pointed out that spatial awareness of 3D objects is important when using visual learning. Learners with high spatial awareness outperform learners with low spatial awareness. Learning is improving, but the gap is shrinking. Spatial awareness is closely related to the development of intelligence and reasoning ability and is a prerequisite for the development and use of 3D animation software. "Spatial visualization ability refers to an individual's ability to replay, rotate, rotate, or flip images to reflect the effects of mental stimuli. Figures 1 and 2 are examples of 3D art that have been modelled and enhanced using technology using 3D software (Mid-Journey, AI Image Builder). ADHD learners can view art in 360° by zooming in and out, interacting and becoming part of the art.



**Fig. 1.** (a) The Spirits of the Pumpkins Descended into the Heavens. (b) Learners are inspecting the artwork for themselves. Yayoi Kusama National Gallery of Australia, Canberra, Australia.

"VR technology can also help users improve their spatial awareness [5]. Google Tilt Brush was used as a test tool in this study to create 3D paintings in a virtual reality environment (Fig. 3). Google's Tilt Brush is a room-scale 3D painting virtual reality



**Fig. 2.** Dall-E AI image creator, Yayoi Kusama artworks in the Metaverse.

painting application. Learners can freely select different brushes and painting tools with the VR controller while viewing the virtual space through the HTC Vive HMD display. The entire virtual space becomes a canvas, allowing learners to draw various pictures in the air. Learners can also move freely within the virtual space, become immersed in the process of creating a 3D painting. Tilt Brush is used in the following ways:

- The ADHD learners will first put on his VR headset, such as (Meta quest 2), to set up the drawing/play area, then use the controller to touch the floor and confirm the floor level.
- Second, the learners will define the play area virtually in the headset by drawing boundaries with the controller.
- Third, they will launch the “tilt brush” app and begin a new design in the 3D journey.
- At the end, they have all the tools for Photoshop and more, such as rotation and animation, and they can rotate the entire design, enlarge, or reduce it, create the sky/ceiling and ground, and all the surroundings. Day or night, the ADHD learners can create designs with dynamic brushes and some animated 2d or 3d drawings.

Literally, anything they can think of. Figure 3 shows an example of a learner designing in the Metaverse.

ADHD learners hold the key to countless ways to test their understanding and create their own examples. As a result, you will be able to allow ADHD learners to review and withdraw their actions within the learning procedures, and it is an excellent learning and assessment tool.

## 4 Results/Discussion

In this study, we employed 3D animation learning software and VR-generated 3D models to aid in the education of students with ADHD. These tools were specifically selected to cater to the learning needs of a smaller group of ADHD students and we discovered that teaching ADHD students is difficult in a traditional classroom/lecture setting, especially



**Fig. 3.** An example of the Tilt Brush output (Li-Hsing Ho, Hung Sun, and Tsun-Hung Tsai, 2019)

when it comes to art and design, because they may not follow along or recognize the differences, so we tried to use a logic game (e.g., gamification) and obtained the following results:

- This method emphasizes logic and uses pictures to attract the attention of hyperactive children. This approach appears to be appropriate for engaging children with ADHD and may be an effective method for improving their learning outcomes.
- The game content is designed to be enjoyable and not boring for the user. This method is also appropriate for engaging children with ADHD and has the potential to improve their learning outcomes.
- Additionally, this method emphasizes the game's ease of use by employing several buttons or arrows to make it simple and enjoyable to play. This method can be beneficial for children with ADHD who may struggle with complex game controls.

We propose that learners with ADHD perform better in a Metaverse environment compared to traditional in-person learning in a physical classroom. This is because a metaverse environment can provide a more flexible and customizable learning experience that can be tailored to the specific needs of learners with ADHD. Additionally, a Metaverse environment can offer various interactive and immersive learning activities that can engage learners with ADHD and help them retain information better than traditional teaching methods.

The findings of this study are consistent with those of other researchers [51] who found that VR has a profound effect on emotions. According to a comparison in Table 1 of ADHD learners in classroom and Metaverse learning, learners were more engaged in learning. In terms of test scores, VR learning outperformed instructional videos. Improved student learning performance may have been facilitated by 3D immersion and interaction with the VR environment.

**Table 1.** Comparison of classroom learning, and Metaverse learning.

Factors	ADHD learner in Classroom learning	ADHD learner in Metaverse learning
Time / Location	Specific time and location	Any time and location
Interaction	Real teachers and peers	Avatar and VR environment
Resource	Usual resources such as books, presentations, videos, and physical activities	3D immersion and 3D VR content
Evaluation	Concentrate on learning outcomes	Incorporate formative and summative data
Facilitates	Mostly based on the teacher's lectures	VR activities such as VR games and videos

## 5 Conclusion

Overall, gamification is what people want, especially those who want to learn through play, so it will continue to expand. The goal of gamification is to transform a hyperactive student into a calm and directed student.

Immersion VR is engaging for both educators and students. Participants, on the other hand, needed help fitting and wearing the device. In addition, operating procedures had to be defined so that the software could be used successfully. Also, only one person can use it at a time. If spatial limitations and equipment issues are addressed, virtual reality can be used in 3D-enhanced education to allow learners to experience and develop their spatial skills.

To promote engagement of ADHD learners with 3D animation learning software, 3D model building using VR or manipulation of animation production software combined with immersive experiences to generate and sustain attention and interest and motivate exploration. We need to strengthen the instruction to match the user's learning and learning needs.

## References

- Stephenson, N.: *Snow crash: a novel*. Spectra (1992)
- Tarouco, L., Gorziza, B., Corrêa, Y., Amaral, É.M., Müller, T.: Virtual laboratory for teaching Calculus: an immersive experience. In: 2013 IEEE Global Engineering Education Conference (EDUCON), pp. 774–781. IEEE (2013). <https://doi.org/10.1109/educon.2013.6530195>
- Dionisio, J.D.N., Burns, W.G., III., Gilbert, R.: 3D virtual worlds and the metaverse: current status and future possibilities. ACM Comput. Surv. **45**(3), 38, Article no. 34 (2013). <https://doi.org/10.1145/2480741.2480751>
- Kapp, K.M.: *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Wiley, Hoboken (2012)
- Usmani, S.S., Sharath, M., Mehendale, M.: Future of mental health in the metaverse. Gen. Psychiatry **35**, e100825 (2022). <https://doi.org/10.1136/gpsych-2022-100825>

6. Polanczyk, G.V., Willcutt, E.G., Salum, G.A., Kieling, C., Rohde, L.A.: ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int. J. Epidemiol.* **43**(2), 434–442 (2014). <https://doi.org/10.1093/ije/dyt261>. PMID: 24464188; PMCID: PMC4817588
7. Hwang, G.-J., Chien, S.-Y.: Definition, roles, and potential research issues of the metaverse in education: an artificial intelligence perspective. In: Elsevier Computers and Education: Artificial Intelligence Conference (2022)
8. Angelova, N.: Gamification in Education (2018). <https://doi.org/10.4018/978-1-5225-5198-0>
9. Alqithami, S., Alzahrani, M., Alzahrani, A., et al.: AR-therapist: design and simulation of an AR-game environment as a CBT for patients with ADHD. *Healthcare* **7**(4), 146 (2019)
10. Areces, D., Dockrell, J., García, T., et al.: Analysis of cognitive and attentional profiles in children with and without ADHD using an innovative virtual reality tool. *PloS One* **13**(8), e0201039 (2018)
11. Baragash, R.S., Al-Samarraie, H., Moody, L., et al.: Augmented reality and functional skills acquisition among individuals with special needs: a meta-analysis of group design studies. *J. Spec. Educ. Technol.* **35**(3), 382–389 (2020)
12. Barbaresi, W.J., Katusic, S.K., Colligan, R.C., et al.: Long-term school outcomes for children with attention-deficit/hyperactivity disorder: a population-based perspective. *J. Dev. Behav. Pediatr.* **28**(4), 265–273 (2007)
13. Bashirim, A., Ghazisaeedi, M., Shahmoradi, L.: The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: a literature review. *Korean J. Pediatr.* **60**(11), 337–343 (2017)
14. Coleman, B., Marion, S., Rizzo, A., et al.: Virtual reality assessment of classroom-related attention: an ecologically relevant approach to evaluating the effectiveness of working memory training. *Front. Psychol.* **10**, 1851 (2019)
15. Cubillo, A., Halari, R., Ecker, C., Giampietro, V., Taylor, E., Rubia, K.: Reduced activation and inter-regional functional connectivity of fronto-striatal networks in adults with childhood Attention-Deficit Hyperactivity Disorder (ADHD) and persisting symptoms during tasks of motor inhibition and cognitive switching. *J. Psychiatric Res.* **44**(10), 629–639 (2010). <https://doi.org/10.1016/j.jpsychires.2009.11.016>
16. Daley, D., Birchwood, J.: ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom?: ADHD and academic performance. *Child: Care Health Dev.* **36**(4), 455–464 (2010). <https://doi.org/10.1111/j.1365-2214.2009.01046.x>
17. Dalsgaard, S., Nielsen, H.S., Simonsen, M.: Consequences of ADHD medication use for children's outcomes. *J. Health Econ.* **37**, 137–151 (2014)
18. Carmine, F., Berto, R.: Contact with nature can help ADHD children to cope with their symptoms. The state of the evidence and future directions for research. *Vis. Sustain.* **14**, 1–11 (2020)
19. Evans, S.W., Owens, J.S., Bunford, N.: Evidence-based psychosocial treatments for children and adolescents with attention-deficit/hyperactivity disorder. *J. Clin. Child Adolesc. Psychol.* **43**(4), 527–551 (2014)
20. Evans, S.W., Owens, J.S., Wymbs, B.T., et al.: Evidence-based psychosocial treatments for children and adolescents with attention deficit/hyperactivity disorder. *J. Clin. Child Adolesc. Psychol.* **47**(2), 157–198 (2018)
21. Faraone, S.V., Biederman, J., Monuteaux, M.C., Doyle, A.E., Seidman, L.J.: A psychometric measure of learning disability predicts educational failure four years later in boys with attention-deficit/hyperactivity disorder. *J. Attention Disord.* **4**(4), 220–230 (2001). <https://doi.org/10.1177/108705470100400404>
22. Frolli, A., Ricci, M.C., Cavallaro, A., et al.: Virtual reality improves learning in children with ADHD. In: Proceedings of the IATED EDULEARN 2021, pp. 9229–9236 (2021)

23. Gongsook, P.: Time simulator in virtual reality for children with attention deficit hyperactivity disorder. In: Herrlich, M., Malaka, R., Masuch, M. (eds.) ICEC 2012. LNCS, vol. 7522, pp. 490–493. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-33542-6\\_61](https://doi.org/10.1007/978-3-642-33542-6_61)
24. Horbova, M., Andrunyk, V., Chyrun, L.: Virtual reality platform using ML for teaching children with special needs. In: CEUR Workshop Proceedings, vol. 2631, pp. 209–220 (2020)
25. Inglis, S.K., Carucci, S., Garas, P., et al.: Prospective observational study protocol to investigate long-term adverse effects of methylphenidate in children and adolescents with ADHD: the attention deficit hyperactivity disorder drugs use chronic effects (ADDUCE) study. *BMJ Open* **6**(4), e010433 (2016)
26. Jeffs, T.L.: Virtual reality and special needs. *Themes Sci. Technol. Educ.* **2**(1–2), 253–268 (2010)
27. Kyaw, B.M., Saxena, N., Posadzki, P., et al.: Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. *J. Med. Internet Res.* **21**(1), e12959 (2019)
28. Martin, A.J.: The role of ADHD in academic adversity: disentangling ADHD effects from other personal and contextual factors. *Sch. Psychol. Q.* **29**(4), 395 (2014)
29. Mühlberger, A., Jekel, K., Probst, T., et al.: The influence of methylphenidate on hyperactivity and attention deficits in children with ADHD: a virtual classroom test. *J. Atten. Disord.* **24**(2), 277–289 (2016)
30. Naslund, J.A., Aschbrenner, K.A., Araya, R., et al.: Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. *Lancet Psychiatry* **4**(6), 486–500 (2017)
31. Parsons, T.D., Bowerly, T., Buckwalter, J.G., Rizzo, A.A.: A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods. *Child Neuropsychol.* **13**(4), 363–381 (2007). <https://doi.org/10.1080/13825580600943473>
32. Powell, L., Parker, J., Harpin, V.: What is the level of evidence for the use of currently available technologies in facilitating the self-management of difficulties associated with ADHD in children and young people? A systematic review. *Eur. Child Adolesc. Psychiatry* **27**(11), 1391–1412 (2017). <https://doi.org/10.1007/s00787-017-1092-x>
33. Pozzi, M., Carnovale, C., Peeters, G., et al.: Adverse drug events related to mood and emotion in paediatric patients treated for ADHD: a meta-analysis. *J. Affect. Disord.* **238**, 161–178 (2018)
34. Romero-Ayuso, D., Toledano-González, A., Rodríguez-Martínez, M.D.C., et al.: Effectiveness of virtual reality-based interventions for children and adolescents with ADHD: a systematic review and meta-analysis. *Children* **8**(2), 70 (2021)
35. Radiani, J., Majchrzak, T.A., Fromm, J., et al.: A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. *Comput. Educ.* **147**, 103778 (2020)
36. Shamseer, L., Sampson, M., Bukutu, C., et al.: CONSORT extension for reporting N-of-1 trials (CENT) 2015: explanation and elaboration. *BMJ* **350**, h1793 (2015)
37. Spencer, T., Noyes, E., Biederman, J.: Telemedicine in the management of ADHD: literature review of telemedicine in ADHD. *J. Atten. Disord.* **24**(1), 3–9 (2020)
38. Gartner. Predicts 2022: 4 Technology Bets for Building the Digital Future (2022). <https://www.businesswire.com/news/home/20220207005085/en>. Accessed 06 Aug 2022
39. Genay, A.C.S., Lecuyer, A., Hachet, M.: Being an avatar “for real”: a survey on virtual embodiment in augmented reality. *IEEE Trans. Vis. Comput. Graph.* **28**, 5071–5090 (2021). <https://doi.org/10.1109/tvcg.2021.3099290>
40. Guo, H., Gao, W.: Metaverse-powered experiential situational English- teaching design: an emotion-based analysis method. *Front. Psychol.* **13**, 859159 (2022). <https://doi.org/10.3389/fpsyg.2022.859159>

41. Huang, H., Hwang, G.-J., Chang, S.-C.: Facilitating decision making in authentic contexts: an SVVR-based experiential flipped learning approach for professional training. *Interact. Learn. Environ.* 1–17 (2021). <https://doi.org/10.1080/10494820.2021.2000435>
42. Hwang, G.-J., Chien, S.-Y.: Definition, roles, and potential research issues of the metaverse in education: an artificial intelligence perspective. *Comput. Educ. Artif. Intell.* **3**, 100082 (2022). <https://doi.org/10.1016/j.caai.2022.100082>
43. Jeon, J.H.: A study on the principle of Metaverse composition with a focus on Roblox. *Korean Assoc. Vis. Cult.* **38**, 257–279 (2021). <https://doi.org/10.21299/jovc.2021.38.10>
44. Jovanović, A., Milosavljević, A.: VoRtex metaverse platform for gamified collaborative learning. *Electronics* **11**, 317 (2022). <https://doi.org/10.3390/electronics11030317>
45. Kang, Y.: Metaverse framework and building block. *J. Korea Inst. Inf. Commun. Eng.* **25**, 1263–1266 (2021). <https://doi.org/10.6109/JKIICE.2021.25.9.1263>
46. Kim, J.H., Lee, B.S., Choi, S.J.: A study on metaverse construction and use cases for non-face-to-face education. *J. Converg. Cult. Technol.* **8**, 483–497 (2022). <https://doi.org/10.17703/JCCT.2022.8.1.483>
47. Ko, H., Jeon, J., Yoo, I.: Metaverse platform-based flipped learning framework development and application. *J. Korean Assoc. Inf. Educ.* **26**, 129–140 (2022). <https://doi.org/10.14352/jkaie.2022.26.2.129>
48. Weng, T., Hsu, M.-H., Yang, D.-C.: A Study investigating the use of 3D computer animations of trigonometric functions to enhance spatial perception ability. *Int. J. Inf. Educ. Technol.* **7**, 23 (2017)
49. Lee, E.A.-L., Wong, K.W., Fung, C.C.: How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Comput. Educ.* **55**, 1424–1442 (2010)
50. Höffler, T.N.: Spatial ability: Its influence on learning with visualizations—a meta-analytic review. *Educ. Psychol. Rev.* **22**, 245–269 (2010)
51. Devon, A., Adrian, V.M.: Learning in virtual reality: effects on performance, emotion and engagement. *Res. Learn. Technol.* **26** (2018)

# **Security, Privacy and Ethics in Augmented Intelligence**



# Culture of Ethics in Adopting Learning Analytics

Dimitrios Tzimas<sup>(✉)</sup> and Stavros Demetriadis

Aristotle University of Thessaloniki, Thessaloniki, Greece  
[{detzimas,sdemetri}@csd.auth.gr](mailto:{detzimas,sdemetri}@csd.auth.gr)

**Abstract.** Learning analytics (LA) collects, analyzes, and reports large amounts of data about learners in order to improve learning in intelligent tutoring systems. Because LA ethics is an interdisciplinary field that addresses moral, legal, and social issues, institutions are responsible for implementing frameworks that address these concerns. Many ethical concerns apply to educational data sets of any size. However, in this study, we concentrate on big data, which increases the scale and granularity of the data collected. We present a synthesis on a growing subject of interest based on ethics regarding the capture of data by LA. This research aims twofold: (a) to extend the review of the scientific literature on LA ethics issues and (b) to identify emerging trends and answer open-field questions discussing three case studies. The following are the research questions for this study: what does LA ethics mean for educational stakeholders, and what are students' and teachers' perspectives on ethics as a factor in adopting LA? We developed a multi-stage design process that included a literature review, empirical research, and community involvement. The literature review identified 68 articles after searching journals and conferences. The selected articles were thoroughly examined using qualitative content analysis. The findings point to a lack of evidence-based guidelines on data ethics and the need to develop codes of practice to evaluate LA ethics policies. Finally, this work applies an ethical checklist to three case studies as an instructional design model for scholars, policymakers, and instructional designers, so partners can use LA responsibly to improve learning and teaching efficacy.

**Keywords:** Learning analytics adoption · Ethics · Co-design · Distance education · Data privacy and management · Stakeholders' perspectives

## 1 Introduction and Background

The growing availability of learning data necessitates research and debate on data ethics and students' rights and expectations. In this study, we conducted a literature analysis to identify emerging ethical issues related to learning analytics (LA). This content analysis will help educational stakeholders understand this area. The critical contribution of this work is a discussion of how LA's ethical concerns are addressed and resolved in intelligent tutoring systems (ITSs), discussing students' and teachers' perspectives, attitudes, and expectations in three case studies. Consequently, this paper can provide valuable guidelines and tracks of research and development policies. The background

section discusses LA, ethics, and the research questions. The method section presents the methodology, and the results section gives insights that propose trustworthy LA ethics. Finally, the discussion goes over the ideas that point to future directions.

### 1.1 Ethics and Learning Analytics

Data analytics is a mature technology and industry that is applied to real-world financial, business, and health systems. Big data analytics is the application of novel analytics techniques to large datasets in order to generate meaningful conclusions, make better decisions and recommendations, or evaluate models for improving organizational processes (Asamoah et al., 2017). Because we live in the big data era, all data is collected, expanding learning data's volume, velocity, and veracity (Prinsloo & Slade, 2017). Big data opens up new opportunities for supporting personalized learning (Kuo et al., 2022) based on data rather than rules. Furthermore, big data analytics uses analytical and data mining tools to massive datasets to better organizational goals and learner-centered education. LA is defined by the Society for Learning Analytics Research (SoLAR) (Lang et al., 2017) as the “measurement, collection, analysis, and reporting of big data about learners and their contexts and behaviors, for purposes of understanding and optimizing learning and the environments in which it occurs” (Siemens & Long, 2011, p. 33). Learning analytics is an interdisciplinary didactical technique for tutoring systems that employs computer science, artificial and augmented intelligence, and cognitive sciences principles to assist students, teachers, and educational institutions.

Ethics is a set of moral concepts concerned with what is right and ethos for people and society (Gray & Boling, 2016). Professional ethics is an application of ethics, and there are three types: deontological, consequentialist, and applied (Kitto & Knight, 2019). Ifenthaler and Tracey (2016) identified ethical concerns about educational data, and Siemens (2013) indicated that students fear data-based action stereotypes. Many of the ethical issues applies equally to learning data sets of any size. We focus on extensive data and improving data granularity. Finally, we apply rule-based consequentialism from deontological and consequentialist perspectives.

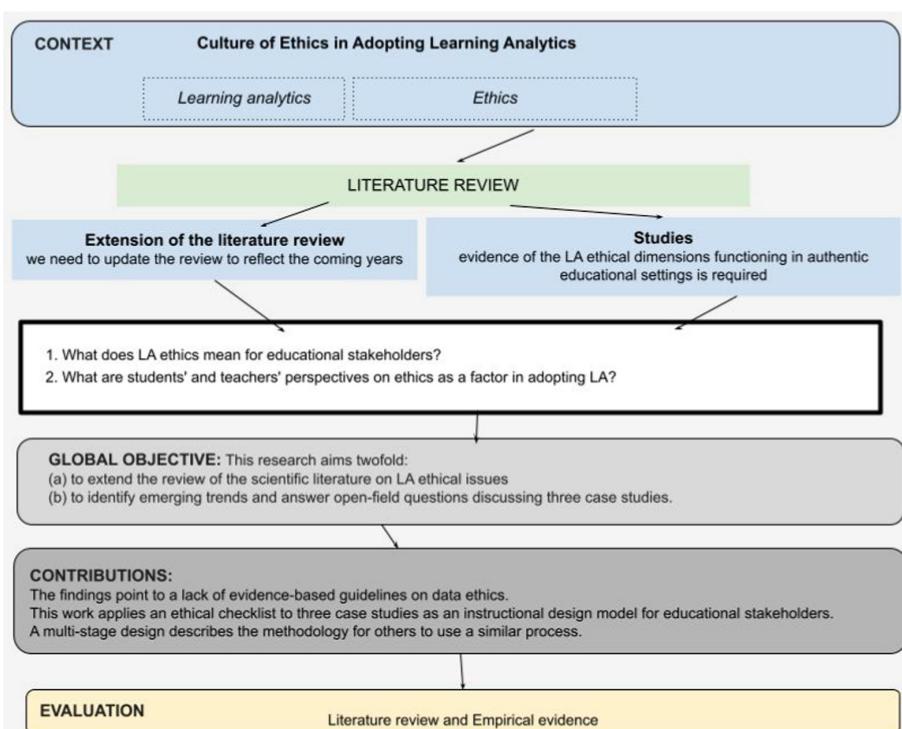
**Antitheses in the Literature.** Contradictions in technology, education, and policy are addressed in the literature (Braunack-Mayer et al., 2020; Tsai et al., 2019). All stakeholders have different points of view (Kitto & Knight, 2019), and there may be a power imbalance among them. Instructors are ethically responsible for using LA outcomes without harmful interventionism (Arnold & Sclater, 2017). Learners want guidance in their learning to increase engagement (Herder & Kawase, 2012), but they are concerned about data privacy (Ifenthaler & Schumacher, 2016; Jones & VanScoy, 2019). Furthermore, decision-makers employ deterministic behaviorism-based algorithms, but the learning approach requires stochastic models based on specialized learning theories (Fynn, 2016). Similarly, ethics varies globally (Willis et al., 2016). Thus, a contradiction relates to differing perspectives on ethical concerns between countries.

## 1.2 Research Questions

Considering their practical applications, our work extends previous studies (Tzimas & Demetriadis, 2021a; Tzimas & Demetriadis, 2021b). Figure 1 depicts the study's research context, goals, and contributions.

We use instructional design theory, specifically design layering, as a conceptual framework (Gibbons & Rogers, 2009). Our study focuses on the strategy and data management layers. The strategy layer specifies the roles of the agents, while data capture, analysis, and interpretation are all specified by the data management layer. Our work is expected to be a synopsis and guide for key educational stakeholders interested in learning more about the emerging field of LA ethics. Overall, this study is guided by the following research questions (RQs):

- First RQ: What does LA ethics mean for educational stakeholders?
- Second RQ: What are students' and teachers' perspectives on ethics as a factor in adopting LA?



**Fig. 1.** A diagram depicting the context, objectives, and contributions of the study

## 2 Method

### 2.1 Research Design

We devised a multi-stage design process that included a literature review, empirical research, and community participation. The LA literature was researched to determine current trends in the LA ethics field. We searched the literature context of IEEE Xplore Digital Library, Elsevier, Wiley Online Library, ACM, and Springer. The following journals were finally used in the research: Educational Technology Research & Development, Educational Technology & Society, Journal of Learning Analytics, British Journal of Educational Technology, British Educational Research Journal, International Journal of Technology Enhanced Learning, American Behavioral Scientist, Computers and Education, Theory and Research in Education, International Review of Research in Open and Distributed Learning, Journal of E-Learning and Knowledge Society, Journal of Electronic Resources Librarianship, and Australasian Journal of Educational Technology. The search term “Learning Analytics” was used in the abstracts, author keywords, and titles of the candidate articles. Searches were conducted for articles published in journals in English between 2011 and 2021.

Regarding the paper selection process, 583 articles initially met the criteria. After studying their abstracts and conclusions, we chose a corpus of 68 papers covering ethical principles for LA. The criterion for the final 68 papers was the study of LA ethical issues. The articles under consideration included critical analysis conceptual articles, literature reviews, and empirical studies.

For answering RQ2, we contrasted three case studies based on community involvement to discuss students’ and teachers’ perspectives on ethics as a factor in adopting LA.

## 3 Results

### 3.1 Ethical Issues in Learning Analytics (RQ1)

To answer the first research question, our classification presents ethical dimensions within the data management layer representing values about goals. We present a systematic review of how ethical aspects are dealt with in LA.

**Data Privacy and Ownership.** Personal data privacy includes how stakeholders collect and process personal data (Ifenthaler & Schumacher, 2016). A legal definition of privacy is a person’s right to control access to their data and to protect themselves from abuse. More than half of the articles mentioned privacy, emphasizing the issue’s importance. Big data is global, so many institutions lack control over storing learners’ data where different cultures and laws may apply (Willis et al., 2016).

Security concerns data protection in LA, which uses sensitive datasets (Jones & VanScy, 2019), while anonymization is insufficient for privacy concerns: “how little data it takes to reidentify individuals” (Haythornthwaite, 2017, p. 2). As a result, data analysts are interested in obscuring log data while it remains helpful in providing personalized services. Encrypting the data logs could be one method. Any qualitative

analysis of the data would be impossible in this situation. Another method for protecting learners' privacy is collecting data from many. Analyzing individual user behavior, however, would be difficult; anonymization implies a trade-off between privacy and fidelity (Herder & Kawase, 2012). According to Cruz et al. (2015), access management refers to granting access to protected data. Because students share much personal information requires authorization and trust (Ahn et al., 2021). Others have suggested that privacy may be sacrificed for educational benefits (Siemens, 2013).

In the same way, data ownership is a complicated legal and moral issue. The learner owns raw data; however, processed data no longer belongs to the learner. The term ownership refers to the data collected, processed, and output. The following questions were posed by Pardo and Siemens (2014): Can students choose how their information is shared? Who owns the data: the institutions that provide the infrastructure, the students, or the companies that develop educational products? Finally, Hoel et al. (2017) discussed students' "right to be forgotten" regarding data minimization. This focuses on data collection with a purpose, as more data does not equal better educational data (Ifenthaler & Tracey, 2016).

**Transparency and the Duty to Act.** Transparency entails making an informed decision to opt-in or out (Pargman & McGrath, 2021) and giving students self-control (Jones & VanScoy, 2019). It contains information about the collected data, the processing, and the time it will be stored (Pardo & Siemens, 2014). Informed consent is an individual's permission to take action based on data processing results (Sun et al., 2019). For example, schools must notify parents when their children's data is being analyzed. IRBs should approve informed consent from learners before participating in LA designs, and learners who decline to participate will not be penalized (Bellini et al., 2019).

According to Arnold and Sclater (2017), educational institutions are ethically obligated to obtain high-quality education data to provide the best possible support. If learners can opt-out, this may result in critical gaps in the dataset, reducing the efficiency of LA systems. Students' costs are high regarding fees and time spent on studies. As a result, educational institutions should support and encourage students (Scholes, 2016). Prinsloo and Slade (2017) state that educational agents are ethically responsible for acting when data obligates action. Furthermore, students share responsibility for doing their best to succeed (obligation to know) (Sun et al., 2019). Students, according to SRL, are active users who employ strategies (for example, metacognition, time management, and help-seeking).

**Paternalism.** From a pedagogical standpoint, allowing students to learn from previous experiences is critical without revealing their student profile "etched like a tattoo into their digital skins" (Mayer-Schonberger, 2011, p. 14). Academics may interpret a student's individualized historical data to classify them (Lawson et al., 2016). On the other hand, analysis based on individual student characteristics must not limit students' expectations and intellectual freedom (Chen & Zhu, 2019; West et al., 2020). Although data-driven learning can improve performance and retention, it can also lead to student labeling (Holloway, 2020), raising the concern that students may be stereotyped and creating norms for future surveillance (Braunack-Mayer et al., 2020; Scholes, 2016). LA-based interventions must be implemented following a learning theory and pedagogy (Roux

et al., 2022), such as self-regulated learning (SRL). Beattie et al. (2008) Proposed that predictions of learning behavior are not deterministic. Furthermore, Wintrup (2017) argued that knowledge of student profiles conceals risks that students will be negatively classified instead of enhancing diversity and inclusion in teaching and learning. For example, an institution's policy may exclude low-performance students.

**Fairness in Algorithms.** Errors in data analysis include data misinterpretation (human error) and misleading patterns (algorithm-based error) (Fynn, 2016). Noisy data and inaccurate models result in poor decisions. As a result, from a pedagogical standpoint, the outcomes may be detrimental to students. When making predictions for future learners, training data is critical. If the characteristics of the future population differ significantly from those of the present, such training will produce dubious results (Whitman, 2021). ML algorithms estimate parameters to fit a predetermined model (e.g., for sentiment analysis), which may not be suitable for human education and well-being. Instructors will sometimes use a domain-specific model with untrustworthy statistical methods, and even if they have aggregated data, they will analyze the results for an individual (Arnold & Sclater, 2017). Nonetheless, humans create the models, training data, and results, and their biases may influence the analysis results (Sarkar, 2019). According to Arnold and Sclater (2017), instructors fear losing students' autonomy in ITSs. Beattie et al. (2008) Argued that instructors could use predictive analytics to profile students' future performance inconveniently.

### 3.2 Synopsis of Findings

We present a summary of findings to provide evidence of our results and insights. After reviewing the 68 studies, a word list was sorted by frequency in descending order to highlight the main themes. Learners (145), HEIs (96), teachers (39), and instructional designers (18) are the most common themes in the stakeholder category. Similarly, in the category of ethical issues, the most common themes are privacy (115), the duty to act (17), profiling (14), and transparency (12). Finally, policy (38), legal (13), and GDPR (6) are other popular topics.

To make issues more manageable, we address open-ended questions in the included studies: What ethical axes should encourage the adoption of LA? (West et al., 2016); Would students be willing to have their data used if it prevented them from stopping out? (Arnold & Sclater, 2017); Can a student still opt out of a predictive model developed from a collection of data traces? (Drachsler & Greller, 2016).

Table 1 summarizes the ethical issues, the most prominent data management methods, and potential harm. In terms of privacy, institutions should ensure administrative efficiency, data minimization, and control. Furthermore, students should be able to opt out without repercussions and have the right to be forgotten. Concerning surveillance, stakeholders aim to avoid labeling and monitoring students while maintaining the objectivity of data and predictive models. The obligation to act and students' right to know should be applied as a moral imperative. Generally, this study looks into the overlapping ethical issues (see Fig. 2) identified by our research.

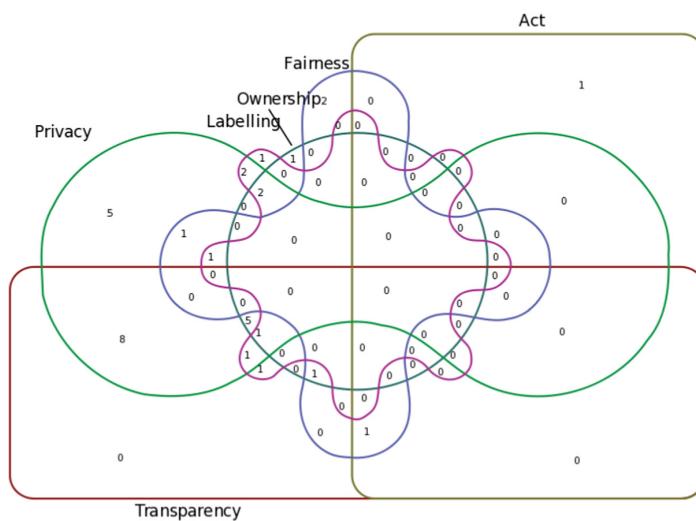
**Table 1.** Ethical concerns, data management methods, and potential harm.

Ethical issue	Methods	Harm
<i>Data ownership &amp; Privacy</i>	Make it clear who has access to the data; Create contracts with vendors that respect privacy and the GDPR; Data should be anonymized, and students should be informed about their analysis; Consider the differences in laws between countries; Students have the authority to correct incorrect information	Stakeholders will not trust the LA services; Learners' performance-related anxiety
<i>Surveillance</i>	Ensure that instructor feedback does not demotivate students; Labels should not be applied to students who are hindering their education; To model students as active users, use an instructional design	Learners are passive recipients who are afraid of bias and untrustworthy categorization
<i>Algorithmic fairness</i>	Consider the dynamic nature of a student's performance; Explain to students how the models produce consistent results; Inform the students that LA should not be used as the sole source of decision-making; Teach students how to interpret LA results critically	Learners will lose their autonomy; Institutions predict without fully comprehending the model
<i>Acting Obligation</i>	Consider the predictive value of LA; Use early warning systems to increase student motivation; Inform students about their progress and provide timely support; and encourage learners to self-intervention	There is no timely support
<i>Transparency</i>	Ensure that information is exclusively used for learning; A student can see what an institution sees and opt out of a data model	Learners will lose motivation to submit their analysis data

(continued)

**Table 1.** (continued)

Ethical issue	Methods	Harm
<i>Stakeholders</i>	Inform students about the importance of self-intervention; Provide LA training to teachers and data administrators; Establish communication channels between stakeholders (e.g., IRBs) in data interpretation	Stakeholders will have no means of communication

**Fig. 2.** The overlapping of ethical issues in the literature

### 3.3 Stakeholders' Perspectives on the Ethics of Adopting LA (RQ 2)

The purpose of this section is to involve the educational community and reveal how students and instructors discuss LA ethical issues and questions (Ahn et al., 2021). First, because of a lack of students' perspectives in the literature (West et al., 2020), we collected students' perceptions of adopting LA in tertiary education (Tzimas & Demetriadis, 2021b). We used questionnaires to gather student feedback on LA in order to develop a data-driven decision-making strategy. The aim was to elicit their perspectives on the LA experience and acceptance. We concluded that considering LA's ethical issues improved students' learning satisfaction and trust in LA. Following (Ahn et al., 2021) study, students contended that: the obligation to act is beneficial to their progress, they expect privacy to be implemented in all courses, and transparency would overcome students' resistance to LA adoption.

In another study (work in progress), we used PANDORA (Tzimas & Demetriadis, 2021a) ethical guidelines to mediate students' LA adoption in tertiary education. We

used ethics to build trust among stakeholders, specifically the methods listed below: Informed consent and anonymization; feedback motivates students given the dynamic nature of learning outcomes; informing students that LA is not the only factor in decision-making; and the obligation to act. An opinion-mining survey was used to collect student opinion data. We used content analysis to interpret data patterns. After responding to the open-ended question, “Feel free to comment on your experience with LA”, students’ perspectives were extracted: “I wish we had this kind of guidance in all our lessons”. As a result, students anticipate that their data will be used to inform support for reducing teacher resistance to LA adoption. The students then mentioned their lack of involvement in the design process because of the teacher and institutional-centric view on LA design. The following insight was derived from student statements: “LA should be tailored to my needs”; “It would be useful if I could configure LA based on my preferences”. In this regard, we propose that student perspectives be considered in LA-based decisions (co-design). Involving students in the LA design process may take some time. However, involving them through participatory methods can transform an unsuccessful prototype into a human-centered learning analytics (HCLA) system (Dimitriadis et al., 2021). Finally, inter-stakeholder communication leads to the agentic positioning of learners (Nichols, 2021) by understanding their priorities, values, needs, and constraints (Chen & Zhu, 2019; Verbert et al., 2020).

In a third study (work in progress), we investigated teacher needs and involvement, participatory and user-centered LA design in K-12 education. The following barriers making teachers resist LA acceptance have been identified: student categorization; privacy and personal data issues; trainee anxiety that their activities are being monitored; the issue of algorithm accuracy and precision; and the risk of stigmatization. According to the relevant excerpts: “We do not need so much analysis creating paternalism. Parents will be irritated”. “It takes time to adapt to LA. A relevant culture must be created.” “Students lose self-regulation. “Big brother is keeping an eye on them”; “Metrics and models do not explain anything”. Finally, teachers propose the following factors that underpin LA adoption: Training, deep stakeholder engagement, a transparent legal framework, and actionable LA.

## 4 Discussion and Conclusion

We emphasize that compliance with the rules and methods necessitates stakeholder awareness and communication. (Nevaranta et al., 2020). This data ethics instructional design encourages agents to be ethically prepared to discuss trust and accountability from policy, technology, and law perspectives (Christopoulos et al., 2021; Tzimas & Dimitriadis, 2021a). In our opinion, institutions should strike a balance between protecting learners’ privacy and collecting all necessary data to achieve their goal of personalized learning support. The rights to privacy and autonomy are not self-contained; stakeholders should consider them in conjunction with other rights, such as the right to know. Learners will resist rather than trust if there is no transparency. Furthermore, learners should be allowed to experiment and make mistakes without feeling watched.

To thoroughly understand data-driven decision-making, key stakeholders should understand what happens inside the black box of algorithms. Students want to know how

the LA models transparently produce results. As a result, institutions must develop cross-stakeholder communication strategies and ethical guidance to explain how complicated LA tools affect individuals.

Furthermore, the literature extracts tensions between innovation and operation (Tsai et al., 2019) and contradicts technology and regulation (Siemens, 2013). Technology and practice drastically evolve, while legal frameworks slowly evolve (Zúñiga-Prieto et al., 2020). Another point of contention is that LA focuses on existing data (Okkonen et al., 2020), but education should foster innovative and critical thinking (Tsai et al., 2019).

In addition, ethics mediates the adoption, efficacy, and impact of LA. It goes over the ethical issues surrounding LA and big data management in education. This research was framed by Gibbons and Rogers's (2009) instructional layering theory as a tool to discuss rethinking data ethics and developing credible solutions. The findings suggest that ethical analysis is a pedagogical requirement for a usable and easy-of-use LA environment. Furthermore, we state that the LA community focuses on LA ethics, as evidenced by increased published articles. Concerning stakeholders and governance, an open research question is how different countries' data management policies will converge for the more significant benefit of education.

Finally, the findings highlight the scarcity of empirically sound guidelines for LA ethics in ITSs. More case studies describing practices in data analytics ethics in higher education are needed. We need more research to validate these data management methods in different educational learner-centered settings for stakeholders and cultures. Finally, we plan to validate our insights on ethical guidance through expert input and field testing.

## References

- Ahn, J., Campos, F., Nguyen, H., Hays, M., Morrison, J.: Co-designing for privacy, transparency, and trust in K-12 learning analytics. In: ACM International Conference Proceeding Series, pp. 55–65 (2021). <https://doi.org/10.1145/3448139.3448145>
- Arnold, K.E., Slater, N.: Student perceptions of their privacy in learning analytics applications. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017, pp. 66–69 (2017). <https://doi.org/10.1145/3027385.3027392>
- Asamoah, D.A., Sharda, R., Hassan Zadeh, A., Kalgotra, P.: Preparing a data scientist: a pedagogic experience in designing a big data analytics course. *Decis. Sci. J. Innov. Educ.* **15**(2), 161–190 (2017). <https://doi.org/10.1111/dsji.12125>
- Beattie, S., Woodley, C., Souter, K.: Creepy analytics and learner data rights. *Ascielite.Org*, pp. 421–425 (2008)
- Bellini, C., De Santis, A., Sannicandro, K., Minerva, T.: Data management in learning analytics: terms and perspectives. *J. E-Learn. Knowl. Soc.* **15**(3), 133–144 (2019). <https://doi.org/10.20368/1971-8829/1135021>
- Braunack-mayer, A.J., Street, J.M., Tooher, R., Scharling-gamba, K.: Student and staff perspectives on the use of big data in the tertiary education sector : a scoping review and reflection on the ethical issues. *Rev. Educ. Res.* **90**(6), 788–823 (2020). <https://doi.org/10.3102/0034654320960213>
- Chen, B., Zhu, H.: Towards value-sensitive learning analytics design. In: ACM International Conference Proceeding Series, pp. 343–352 (2019). <https://doi.org/10.1145/3303772.3303798>
- Christopoulos, A., Mystakidis, S., Pellas, N., Laakso, M.-J.: ARLEAN: an augmented reality learning analytics ethical framework. *Computers* **10**, 92 (2021). <https://doi.org/10.3390/computers10080092>

- Cruz, G., Costa, A., Martins, P., Goncalves, R., Barroso, J.: Toward educational virtual worlds: Should identity federation be a concern? *J. Educ. Technol. Soc.* **18**, 27–36 (2015)
- Dimitriadis, Y., Martínez-Maldonado, R., Wiley, K.: Human-centered design principles for actionable learning analytics. In: Tsatsos, T., Demetriadis, S., Mikropoulos, A., Dagdilelis, V. (eds.) *Research on E-Learning and ICT in Education*, pp. 277–296. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-64363-8\\_15](https://doi.org/10.1007/978-3-030-64363-8_15)
- Drachsler, H., Greller, W.: Privacy and analytics – it's a DELICATE issue a checklist for trusted learning analytics. In: 6th Conference on Learning Analytics and Knowledge (2016). <https://doi.org/10.1145/2883851.2883893>
- Fynn, A.: Ethical considerations in the practical application of the Unisa socio-critical model of student success. *Int. Rev. Res. Open Dist. Learn.* **17**(6), 206–220 (2016)
- Gibbons, A.S., Rogers, P.C.: The architecture of instructional theory. In: Reigeluth, C.M., Carr-Chellman, A. (eds.) *Instructional-Design Theories and Models: Building a Common Knowledge Base*, vol. III, pp. 305–326. Lawrence Erlbaum Associates, Hillsdale (2009)
- Gray, C.M., Boling, E.: Inscribing ethics and values in designs for learning: a problematic. *Educ. Tech. Res. Dev.* **64**(5), 969–1001 (2016). <https://doi.org/10.1007/s11423-016-9478-x>
- Haythornthwaite, C.: An information policy perspective on learning analytics. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017*, pp. 253–256 (2017). <https://doi.org/10.1145/3027385.3027389>
- Herder, E., Kawase, R.: Considerations for recruiting contributions to anonymised data sets. *Int. J. Technol. Enhanc. Learn.* **4**(1), 85–98 (2012). <https://doi.org/10.1504/IJTEL.2012.048315>
- Hoel, T., Griffiths, D., Chen, W.: The influence of data protection and privacy frameworks on the design of learning analytics systems. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017*, pp. 243–252 (2017). <https://doi.org/10.1145/3027385.3027414>
- Holloway, K.: Big data and learning analytics in higher education: legal and ethical considerations. *J. Electron. Resour. Librariansh.* **32**(4), 276–285 (2020). <https://doi.org/10.1080/1941126X.2020.1821992>
- Ifenthaler, D., Schumacher, C.: Student perceptions of privacy principles for learning analytics. *Educ. Tech. Res. Dev.* **64**(5), 923–938 (2016). <https://doi.org/10.1007/s11423-016-9477-y>
- Ifenthaler, D., Tracey, M.W.: Exploring the relationship of ethics and privacy in learning analytics and design: implications for the field of educational technology. *Educ. Tech. Res. Dev.* **64**(5), 877–880 (2016). <https://doi.org/10.1007/s11423-016-9480-3>
- Jones, K.M.L., VanScoy, A.: The syllabus as a student privacy document in an age of learning analytics. *J. Doc.* **75**(6), 1333–1355 (2019). <https://doi.org/10.1108/JD-12-2018-0202>
- Kitto, K., Knight, S.: Practical ethics for building learning analytics. *Br. J. Edu. Technol.* **50**(6), 2855–2870 (2019). <https://doi.org/10.1111/bjet.12868>
- Kuo, R., Wasowski, R., Krahn, T., Chang, M.: LORD: a moodle plug-in helps to find the relations among learning objects. In: Crossley, S., Popescu, E. (eds.) *ITS 2022. LNCS*, vol. 13284, pp. 155–122. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_11](https://doi.org/10.1007/978-3-031-09680-8_11)
- Lang, C., Siemens, G., Wise, A., Gasevic, D.: Handbook of Learning Analytics (2017). <https://doi.org/10.18608/hla17>
- Lawson, C., Beer, C., Rossi, D., Moore, T., Fleming, J.: Identification of ‘at risk’ students using learning analytics: the ethical dilemmas of intervention strategies in a higher education institution. *Educ. Tech. Res. Dev.* **64**(5), 957–968 (2016). <https://doi.org/10.1007/s11423-016-9459-0>
- Mayer-Schonberger, V.: Delete: The virtue of forgetting in the digital age. Princeton University Press, Princeton (2011). [https://doi.org/10.1111/j.1540-5931.2011.921\\_1.x](https://doi.org/10.1111/j.1540-5931.2011.921_1.x)
- Nevaranta, M., Lempinen, K., Kaila, E.: Students’ perceptions about data safety and ethics in learning analytics. In: *CEUR Workshop Proceedings*, vol. 2737, pp. 23–37 (2020)

- Nichols, M.: Development of an approved learning analytics ethics position. *Open Learn.: J. Open Dist. e-Learn.* (2021). <https://doi.org/10.1080/02680513.2021.1986376>
- Okkonen, J., Helle, T., Lindsten, H.: Ethical considerations on using learning analytics in finnish higher education. In: Nazir, S., Ahram, T., Karwowski, W. (eds.) AHFE 2020. AISC, vol. 1211, pp. 77–85. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-50896-8\\_12](https://doi.org/10.1007/978-3-030-50896-8_12)
- Pardo, A., Siemens, G.: Ethical and privacy principles for learning analytics. *Br. J. Edu. Technol.* **45**(3), 438–450 (2014). <https://doi.org/10.1111/bjet.12152>
- Pargman, T.C., McGrath, C.: Mapping the ethics of learning analytics in higher education: a systematic literature review of empirical research. *J. Learn. Anal.* **8**(2), 123–139 (2021). <https://doi.org/10.18608/JLA.2021.1>
- Prinsloo, P., Slade, S.: An elephant in the learning analytics room. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017, pp. 46–55 (2017). <https://doi.org/10.1145/3027385.3027406>
- Roux, L., Nodenot, T., Etcheverry, P., Dagarret, P., Marquesuzaa, C., Lopisteguy, P.: A classification approach to recognize on-task student's behavior for context aware recommendations. In: Crossley, S., Popescu, E. (eds.) ITS 2022. LNCS, vol. 13284, pp. 161–170. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-09680-8\\_15](https://doi.org/10.1007/978-3-031-09680-8_15)
- Sarkar, D.: Text Analytics with Python (2019). <https://doi.org/10.1007/978-1-4842-4354-1>. ISBN 978-1-4842-4353-4
- Scholes, V.: The ethics of using learning analytics to categorize students on risk. *Educ. Tech. Res. Dev.* **64**(5), 939–955 (2016). <https://doi.org/10.1007/s11423-016-9458-1>
- Siemens, G.: Learning analytics. *Am. Behav. Sci.* **57**(10), 1380–1400 (2013). <https://doi.org/10.1177/0002764213498851>
- Siemens, G., Long, P.: Penetrating the fog: analytics in learning and education. *Educ. Rev.* **46**, 30–32 (2011). <https://doi.org/10.17471/2499-4324/195>
- Sun, K., Mhaidli, A.H., Watel, S., Brooks, C.A., Schaub, F.: It's my data! Tensions among stakeholders of a learning analytics dashboard. In: Conference on Human Factors in Computing Systems - Proceedings, pp. 1–14 (2019). <https://doi.org/10.1145/3290605.3300824>
- Tsai, Y.S., Poquet, O., Gašević, D., Dawson, S., Pardo, A.: Complexity leadership in learning analytics: drivers, challenges and opportunities. *Br. J. Edu. Technol.* **50**(6), 2839–2854 (2019). <https://doi.org/10.1111/bjet.12846>
- Tzimas, D., Demetriadis, S.: Ethical issues in learning analytics: a review of the field. *Educ. Technol. Res. Dev.* (2021a). <https://doi.org/10.1007/s11423-021-09977-4>
- Tzimas, D., Demetriadis, S.: The impact of learning analytics on student performance and satisfaction in a higher education course. In: Proceedings of the 14th International Conference on Educational Data Mining (EDM21), pp. 654–660. International Educational Data Mining Society (2021b)
- Verbert, K., Ochoa, X., De Croon, R., Dourado, R., De Laet, T.: Learning analytics dashboards: the past, the present and the future. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK 2020), pp. 35–40. Association for Computing Machinery, New York (2020). <https://doi.org/10.1145/3375462.3375504>
- West, D., Huijser, H., Heath, D.: Putting an ethical lens on learning analytics. *Educ. Tech. Res. Dev.* **64**(5), 903–922 (2016). <https://doi.org/10.1007/s11423-016-9464-3>
- West, D., Luzeckyj, A., Toohey, D., Vanderlelie, J., Searle, B.: Do academics and university administrators really know better? The ethics of positioning student perspectives in learning analytics. *Aust. J. Educ. Technol.* **36**(2), 60–70 (2020). <https://doi.org/10.14742/ajet.4653>
- Whitman, M.: Modeling ethics: approaches to data creep in higher education. *Sci. Eng. Ethics* **27**(6), 1–18 (2021). <https://doi.org/10.1007/s11948-021-00346-1>
- Willis, J.E., Slade, S., Prinsloo, P.: Ethical oversight of student data in learning analytics: a typology derived from a cross-continent, cross-institutional perspective. *Educ. Tech. Res. Dev.* **64**(5), 881–901 (2016). <https://doi.org/10.1007/s11423-016-9463-4>

- Wintrup, J.: Higher education's Panopticon? Learning analytics, ethics and student engagement. *High Educ. Pol.* **30**(1), 87–103 (2017). <https://doi.org/10.1057/s41307-016-0030-8>
- Zúñiga-Prieto, M.A., Ortiz, M., Ulloa, M., Jiménez, A.: Applying the LALA framework for the adoption of a learning analytics tool in Latin America: two case studies in Ecuador. In: CEUR Workshop Proceedings, vol. 2704, pp. 6–14 (2020)



# Promoting Ethical Uses in Artificial Intelligence Applied to Education

Ana Matias<sup>(✉)</sup> and Iraide Zipitria

The University of the Basque Country (UPV/EHU), San Sebastian-Donostia, Spain  
[amatias004@ikasle.ehu.eus](mailto:amatias004@ikasle.ehu.eus), [iraide.zipitria@ehu.eus](mailto:iraide.zipitria@ehu.eus)

**Abstract.** During the last decades, technology has created many opportunities to develop the world in diverse ways. Specifically, the growth of Artificial Intelligence (AI) can potentially change every aspect of human social interactions. In the educational field, numerous AI systems have been created using intelligent methods in order to develop novel teaching and learning solutions for the educational context. Most of these applications are powered by Machine Learning (ML) systems and algorithms, which use substantial amounts of data to work. Data reflects social reality and its existing power structures and biases, which is why, in order to guarantee ethical applications, it seems important to analyse issues such as the ones that are being considered in FATE. The development of ethical AI systems have been analysed by many researchers. However, especially in the context of education, biased systems can threaten the learning experience, perpetuating and supporting existing discriminatory practices. If we want to use AI for social good, algorithms must be fair and transmit positive and respectful values to students, while they develop their curricular capacities and psychological aspects. There is a chance to develop social-sensitive and ‘activist’ educational applications to mitigate social existent biases and empower the creation of a fairer and more respectful social world. This work will focus on how AI algorithms influence the learning experience of students as well as the threats the algorithms may provoke to that learning experience, specifically in the context of the discriminative data, which can perpetuate actual biased representations of social groups present in the data.

**Keywords:** Artificial Intelligence · Education · Ethics · Social Bias · Social Change

## 1 Introduction

As we know, life is always changing, as world development does. However, are we prepared to absorb technological changes? Nowadays, computers and technologies have changed our lives in many ways. Precisely, the development of Artificial

Intelligence (AI) is increasingly being applied in diverse contexts of the daily life occupying prominent positions. As AI is present in many simple activities of our daily life, such as leisure activities (e.g. Netflix, Alexa, Chess game), we can locate this intelligent technology in many other key areas of personal and social development, including education, which is an important social context of personality development [28]. According to the educational context in the digital era, technological advances are changing the ways we learn and teach in a tremendous and rapid way [26]. For example, interactive online learning applications such as Massive Open Online Course (MOOC) platforms are becoming more widely used within and outside the classrooms. In addition, nowadays students utilize tablets rather than copybooks to take notes, and educative tools and applications including Moodle or Google Classroom are gaining popularity among other innovative teaching practices [13]. Nevertheless, introducing AI systems in the educational structures lead to ethical issues to be taken into consideration. Artificial Intelligence applications work through methods such as ML and Deep Learning (DL) [7], smart systems that use large amounts of data to make predictions and provide answers. That data is collected from many sources of information as a representation of the meanings of society and the world. However, due to the stereotypes and prejudices present in human interactions, the sources of data used for AI training are frequently biased. Because of this, it is important to consider to what extent those smart systems can generate discriminative and unfair algorithms, and foster social inequalities and discriminative representations in society. Particularly in the case of education, there is already an open discussion about how algorithmic systems could influence students, instructors, institutions, and society as a whole [3, 42, 44, 48, 55]. As a matter of fact, scientists are already conducting research about the ethical conflicts of the working methods and programming of AI systems [1, 5, 6, 11, 15, 36, 48]. Several ethical codes have been developed to control these problematics [23, 29, 40]. FATE (Fairness, Accountability, Transparency and Ethics) is one of the most widely used guidelines in the context of education. Between others, it integrates the concept of Fairness, a term that refers to discrimination within the AI systems. In the educational context, AI is expected to enhance and enrich learning experiences and support the development of innovative teaching practices. The expression *smart education* refers to delivering effective education services using both new technologies' applications and advanced pedagogical practices, tools, and techniques [50]. This technology is being used to define specific and explicit forms of knowledge about education, including several psychological and social aspects which often remain implicit [25]. However, these implicit aspects are not always positive, for instance, some studies show the presence of social biases in the knowledge bases, reflecting the discriminative patterns in society [3, 31, 34]. Algorithmic bias impact educational systems in terms of gender, student race, nationality, ethnicity, urbanicity, native language, socioeconomic status, parental educational background, and whether a student has a parent in the military [2]. The role that education plays in aiding and abetting the reproduction of social exclusion and social inequality is of high significance [59]. Considering the

importance of providing a respectful and ethical education to scholars, it seems important to bear in mind the great responsibility of educational innovations to prevent these advances from perpetuating a vicious cycle of education inequalities [30]. Due to the potential use of large amounts of data (sometimes biased) for AI applications and the already existing inequalities in societal datasets [31, 45], it is important to prevent the perpetuation of the discrimination of different social groups. This paper will focus on how AI-based systems could really get involved in the education, not only in the facilities they provide from the point of view of the learning and teaching processes, but also, in the way it could neutralize existing unfair representations of individuals and brake the cycle of discrimination. That way, students will develop their capacities and personality in a respectful and egalitarian context, lowering the negative repercussions of integrating early discriminatory ways of thinking in the learning process. The focus will be on AI applied to the educational context since it is a crucial area for the development of individuals and society. Accordingly, the work will describe how the AI community suggests ethical guidelines for the employment of smart systems in the real world, focusing then on scholar systems. Subsequently, there are some examples of the consequences of applying biased systems and, finally, a discussion on how this technology could be applied to promote equality and fairness.

## 2 The Power of Artificial Intelligence (AI)

The introduction of AI and ML is catching on among many people, being used extensively in different devices, applications, and services, gaining widespread popularity in the daily life of society [56]. Today, AI is the most important technology of our era and, it has the most evidence-disrupting power, particularly with regards to ML techniques [7]. Despite that, there are also many domains in which AI systems may help meet the needs of society and communities, using them for social good, [14, 21, 51, 58, 60].

AI applications have achieved success in terms of technological developments, theoretical contributions, and impact on education [46]. Global movements can contribute to social issues using AI methods, such as the Amnesty International and Element AI, that help identifying and quantifying online abuse against women on Twitter [17]. Regarding other social questions such as climate change, there are ML applications developed to fight against it, accelerating the process of reducing emissions of electricity systems, by collecting data about the process and reporting on the deployment and operation of electricity system technologies [39]. There are many other ways to apply this technology to contribute to social issues (e.g. [18, 58, 63]). Anyhow, as education is one of the most important spheres of life, becoming a Fundamental Human Right. Today, AI methods have been applied into many academic contexts. Smart education is gaining important relevance in the global educational field [64]. Many functions of educational technology cover different genres of technology-based learning environments, such as computer-supported collaborative learning systems, Intelligent Tutoring Systems

(ITS), animated pedagogical agents, interactive simulations and games, and virtual environments [52]. Through different systems, it is being used for dropout prediction, automated essay scoring, graduate admissions and knowledge inference [2]. Thus, with the application of AI in educational institutions, educators can spend more time on students thanks to the automation of administrative tasks [56]. In addition, AI can analyse students' behavioral patterns, so it is possible to create personalized guidance or feedback for learners and teachers [36]. Among others, AI technology has focused on the students' learning characteristics, and it has developed adaptative learning environments and personalized learning [57]. As Roll and Wylie [46] relate, the education system focus is shifting from the product to the process, expanding its domain-knowledge view, including self-regulation, collaboration, and motivation. Students' learning characteristics are now important trends. For example, through smart environments learners can access digital resources and interact with systems that provide them with the convenient guidance, hints, suggestions, or supportive tools, no matter when or where the individuals wants to use them [32]. Additionally, adaptive learning systems try to place the student at the centre of the learning experience, making the content interactive and dynamic [20]. There has been a growing interest in explicit measurement of learning outcomes, and predictive AI models [53]. Thanks to all these advances, learners have begun to become more autonomous, and the different roles in the learning process are changing [53]. Consequently, it seems of real importance to question how fair remains the data applied for that kind of applications and which is its real potential in this specific context. Smart technologies seem to be nothing but positive steps in the educational development, however, the training data used for their functioning keeps unfairly biased information against certain subpopulations, including those with a particular sexual orientation, race, or gender [31]. Deep Learning (DL) is used for many approaches of ML [7], therefore, it is important to question how biased data could influence the learning and developmental process and what could be done to make fairer educational environments. In fact, the ITS and AIED communities have already opened a discussion about fairness in educational systems [3, 31, 56]. For instance, Kizilcec and Lee [34] have questioned how the increasing use of algorithmic systems in education may impact the scholar community. Accordingly, due to the importance of the learning and developmental process of individuals in the academic context [59], we must take into account the potential of biased data used in AI systems in order to perpetuate existing inequalities between different social groups [31]. The following section contains ethical guidelines developed to analyse the data.

### 3 Ethics in AI

This section is committed to show how scientists are already concerned about the ethical application of AI. First, the ethical discussions for AI systems in general, followed by the analysis applied on educational applications of AI.

### 3.1 Ethical Codes for AI

Talking about AI in general, there is already an increasing community concerned about how algorithmic systems that treat users in an unfair way may affect and impact society [44,47]. If the running data used to work is biased, then the AI systems will be a reflection of it. In areas such as education, this can provoke discriminative treatments among students, affecting their engagement and achievement of the knowledge [12,16,19,54,62]. Thus, is it ethical to overlook that? Considering the implicit and unconscious prejudices against some groups of people that masses have, developing egalitarian AI applications is a good and innovative opportunity to create a more equitable society.

As explained earlier, it seems clear that there are many aspects of smart education to consider. Socio-economic issues such as privacy, trust, and ethics must be taken into consideration in its implementations [56]. Consequently, as an attempt to control the ethical performance of the AI and ML systems, researchers have been motivated to formulate suggestions for AI ethics [24,53] including aspects such as the issue of privacy, transparency, accountability, and fairness. Some examples of these codes are: the Montreal Declaration of Responsible AI development [40], the RESET guide, a way to provide and discuss some fundamental principles for AI technologies focusing on accountability, data protection, safety/non-maleficence, efficacy, transparency, human agency, inclusiveness, fairness, explainability, and integrity [29]; and FATE, a tool based on issues of Fairness, Accountability, Transparency and Ethics (FATE) [23]. The various guidelines of AI ethics have a common goal, which is supporting and improving the ethical development, design, and deployment of AI [24]. The last guide example, FATE, has been created with a sociotechnical orientation. As it has been used to analyse many educational applications (e.g. [36] the focus will be placed on that last guideline.

### 3.2 FATE in AI Applied to Education

FATE can be applied on scholar systems in order to evaluate at which extent they can be used considering Fairness, Accountability, Transparency and Ethics. Firstly, the matter of ‘Fairness’ is related to ensuring equality among the recommended educational opportunities of learners. Educational gains and losses for students depend on the suggested courses, and, by extension, education influences in a significant way individuals’ learning experiences, which should not be undermined arbitrary decisions of recommender systems [36]. On the contrary, it seems important to study the influence of those applications on individuals. Secondly, the concept of ‘Accountability’ engenders how any harmful consequences or algorithmic failures of AI are addressed. To this effect, there are mechanisms for teachers’ knowhow, to support teacher identity, and to connect to institutional resources during the deployment or implementation of these systems in the case there exists any trouble [5]. In addition, ‘Transparency’ refers to user needs for understandable explanations of its working methods. Related to it, in a study carried out by Liao et al. [35], insights into user needs for explainability

are provided, as well as suggestions on how these needs should be understood, addressed, and prioritized. Finally, the notion of ‘Ethics’. In fact, research on algorithmic fairness in education can play a valuable role for social change [34]. However, as Holmes et al. [29], mentioned in their work, how the ethical dimensions are not a central area of focus for many researchers applying AI to education. AI innovations in the educational context have caused positive impacts [5, 56], there are ethical issues which must be analysed and, more specifically, negative societal effects that could be caused. One important consequence is the employment of biased data in the algorithmic systems of AI applied to education; which can treat different subpopulations differently, perpetuating or even exacerbating these biases [31]. Taking into account that the AI systems used, significantly influences the development and learning of students, this consideration becomes of real importance.

## 4 Social Biases

Inequities and inequalities in educational opportunities and outcomes existed long before the new technological advances. For example, achievement gaps and school segregation were already present in many educative contexts long before the digital era [34]. Many people experience discrimination because of social representations that stigmatize many groups basing on multiple social identities [12]. Here, the term discrimination refers to negative or unfair treatment based on a social identity such as gender, race, sexual orientation, or socioeconomic status [10]. For example, racial or ethnic discrimination is quite common among ethnic minority youth in schools and in other public spaces and several studies [22, 33, 61] show that the perpetrators can be peers, but also adult authority figures [4]. All types of discrimination (based on identities such as gender, race, sexual orientation, or socioeconomic status) in schools have negative effects for youths’ psychological, health, and academic outcomes [12]. It is related to having lower school engagement and academic achievement [12, 16, 19, 54, 62]. For all those reasons, it seems important to assure safe and inclusive academic spheres. However, where data is used predictively to assist decision making, it can have ethical consequences, discriminating some groups of populations. Data mining must be approached with care because, otherwise, it can contribute to the reproduction of existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society [45]. The validity of learning analytics algorithms, must be analysed with regards to their predictive accuracy for minorities [45], because sometimes algorithms make decisions that are skewed towards a particular group of people, what leads to unfair algorithm models [2]. Mitchell et al. [38] refer to the term ‘biased’ as different predictive performance across disadvantaged groups with a social axe, including race, gender, and class, in an unfair and unjustified way. Given that the past data may be biased, ML models further perpetuate or even exacerbate these biases. The resultant models treat different subpopulations differently, so it can be described as ‘unfair’ [31]. As a result, the use of

data-driven decision systems has raised concerns about fairness, bias, and discrimination against members of protected classes, such as racial, women, ethnic, seniors, religious, people with disabilities, national minorities, pre-existing medical conditions, and genetic vulnerabilities [34]. There is evidence and concerns about bias and unfairness of ML models applied in educational applications, including students' performance prediction, course recommendation, dropout prediction, knowledge tracing, automated essay scoring and student emotions detection [8,9,41,43]. For instance, Hutt, Duckworth, Gardner and D'Mello [31] showed inherent biases in college applications predicting on-time college graduation. To answer to all these issues, some researchers have developed methods for constructing fair ML systems. For example, He, Burghardt and Lerman [27] described a method to create interpretable representations used by ML models in order to build fairer accurate models of data. That method removes correlations between data and any protected variable that could make biased associations in the algorithms. Another example is 'Sensitive Loss', a discrimination-aware DL method that prevents discriminatory effects by automatic systems of biased face recognition algorithms [49]. Researchers have also introduced tools that can assess the amount of fairness in a system. That is the case of 'Aequitas', a toolkit that produces reports from the obtained data in the different models with regards to several bias and fairness metrics for different population subgroups [37]. There are already ML scientists that are concerned about the consequences of the use of unfair systems, and, in fact, it represents a step forward to prevent discriminatory practices observed in automatic systems. However, observing how do society represent the different subgroups, it is also a good opportunity to promote a more equitable and respectful society. It could be interesting for AI researchers to create activist and more sensible applications in the context of education, to really educate individuals and fight discrimination already present in schools through explicit and visible methods.

## 5 Conclusions and Discussion

As seen, technological advances around AI are having important impacts in human life, being used to bring facilities in diverse contexts. They are fast becoming an invaluable part of the human-development toolkit, generating considerable social changes and benefits. In fact, researchers are already focusing on AI implementations for social good. This AI systems with educational applications are gradually found in different institutions and contexts and, therefore, it is important to make sure this technology can really offer an adequate education to students. As technology have developed in gigantic proportions, we must be conscious about how society works in order to ensure an adequate implementation of that technology in life. In the case of AI systems applied to education, it seems to be a good opportunity to visualize social bias through data to try to make fair applications that mitigate discrimination and, consequently, promote a respectful education. Due to the high influence of technologies in society, and, more precisely, academic, and personal consequences of that discrimination reflexed in the sphere of education, it seems important to assure fair and

equality-sensitive systems, so students can develop and learn positive values and interactions. Learning is a psychological process that has its major impact during the development of individuals. There is a high vulnerability of impact of AI scholar tools, because they are applied in a context of development of individuals and cognitive processes, where learning is being enhanced. That is the reason why it is very important to control the consequences of new technologies in the academic context, not only focusing on the effectiveness of the tools, but also in its influence on the students. As technologies are ever more being present in our daily life, it is a good opportunity to make the best of it and try to change pre-existing social issues. Therefore, the use of AI can be a good opportunity to visualize these implicit practices and biased social representations, in order to produce some step to enhance society to change. Using AI systems, we could re-educate individuals on their implicit social bias. As a case and point, that could be a good scholar practice, that would promote ethics and respect among students and, thus, among their future social interactions. In fact, the present work shows how the ITS and AIED communities are already creating ways to guarantee fair AI systems. That way, systems will neutralize social biases and stop promoting the perpetuation of the discrimination cycle in educational settings. Nevertheless, perhaps another approach could be to take advantage of the knowledge about the existence of discriminative information in the systems to try to tackle that problem from the beginning of the sequence. That is, the information in the algorithms is biased because machines acquire biased data from society, which means that societal meanings are not neutral nor fair. Since community applying AI to education is focused on how systems benefit teaching and learning activities, it would be interesting to expand the academic use of these applications to actively make students reflect about stereotypes and prejudices among some social groups. That way, smart systems could help develop critical values and make people take responsibility of negative social practices, increasing their sensitivity to social issues during their developmental stages. By doing this, as well as offering fair implicit predictions from AI systems, which are already being studied, the AI systems in education could be focused specifically in promoting an equal and respectful view of society, through explicit and active ways. New technologies could be used to visualize and make individuals more sensible and responsible of the biased social representations of society, which damage and disadvantage minorities. To sum up, computer sciences develop in an exponential way, and new advances shock society without even understanding the totality of these discoveries. They are increasingly being present in our life, and it is important to adjust the coexistence between social reality and artificial applications. Clearly, technological appliances on society can be of high benefit for many activities, but it must also be seriously analysed, so it does not threaten societal functioning in negative ways. As well as controlling ethics of AI systems, such as the unequal predictions of AI applications about students. To close up, AI technologies' impacts on the educational area truly has the promise for transforming social values and reduce the actual and unfair discrimination suffered by some social subgroups.

## References

1. Aiken, R.M., Epstein, R.G.: Ethical guidelines for AI in education: starting a conversation. *Int. J. Artif. Intell. Educ.* **11**, 163–176 (2000)
2. Baker, R.S., Hawn, A.: Algorithmic bias in education. *Int. J. Artif. Intell. Educ.* **32**, 1052–1092 (2021)
3. Bayer, V., Hlosta, M., Fernandez, M.: Learning analytics and fairness: do existing algorithms serve everyone equally? In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 71–75. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78270-2\\_12](https://doi.org/10.1007/978-3-030-78270-2_12)
4. Benner, A.D., Graham, S.: The antecedents and consequences of racial/ethnic discrimination during adolescence: does the source of discrimination matter? *Dev. Psychol.* **49**(8), 1602–1613 (2013)
5. Bhimdiwala, A., Neri, R.C., Gomez, L.M.: The FATE of AIED, chap. Advancing the Design and Implementation of Artificial Intelligence in Education through Continuous Improvement. *IJAIED Special Issue*. Springer, Cham (2021)
6. Bogina, V., Hartman, A., Kuflik, T., Shulner-Tal, A.: The FATE of AIED, chap. Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics. *IJAIED Special Issue*. Springer, Cham (2021)
7. Borges, A.F.S., Laurindo, F.J.B., Spinola, M.M., Goncalves, R.F., Mattos, C.A.: The strategic use of artificial intelligence in the digital era: systematic literature review and future research directions. *Int. J. Inf. Manag.* **57**, 102225 (2021)
8. Bridgeman, B., Trapani, C., Attali, Y.: Considering fairness and validity in evaluating automated scoring. Technical report, National Council on Measurement in Education (NCME), Educational Testing Service (ETS), San Diego, USA (2009)
9. Bridgeman, B., Trapani, C., Attali, Y.: Comparison of human and machine scoring of essays: differences by gender, ethnicity, and country. *Appl. Meas. Educ.* **25**, 27–40 (2012)
10. Brown, C.S., Bigler, R.S.: Children's perceptions of discrimination: a developmental model. *Child Dev.* **76**, 533–553 (2005)
11. Busuioc, M.: Accountable artificial intelligence: holding algorithms to account. *Public Adm. Rev.* **81**(5), 825–836 (2020)
12. Byrd, C.M., Carter-Andrews, D.J.: Variations in students' perceived reasons for, sources of, and forms of in-school discrimination: a latent class analysis. *J. Sch. Psychol.* **57**, 1–14 (2016)
13. Chen, X., Xie, H., Hwang, G.J.: A multi-perspective study on artificial intelligence in education: grants, conferences, journals, software tools, institutions, and researchers. *Comput. Educ. Artif. Intell.* **1**, 100005 (2020)
14. Cowls, J.: AI for social good: whose good and who's good? Introduction to the special issue on artificial intelligence for social good. *Philos. Technol.* **34**(1), 1–5 (2021)
15. Curzon, J., Kosa, T.A., Akalu, R., El-Khatib, K.: Privacy and artificial intelligence. *IEEE Trans. Artif. Intell.* **2**(2), 96–108 (2021)
16. Datu, J.A.D.: Everyday discrimination, negative emotions, and academic achievement in filipino secondary school students: cross-sectional and cross-lagged panel investigations. *J. Sch. Psychol.* **68**, 195–205 (2018)
17. Delisle, L., Kalaitzis, A., Majewski, K., de Berker, A., Marin, M., Cornebise, J.: A large-scale crowd-sourced analysis of abuse against women journalists and politicians on twitter. In: AI for Social Good NeurIPS 2018 Workshop 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montreal (2018)

18. Donner, J., et al.: Stages of design in technology for global development. *Computer* **41**, 34–41 (2008)
19. Dotterer, A.M., McHale, S.M., Crouter, A.C.: Sociocultural factors and school engagement among African American youth: the roles of racial discrimination, racial socialization, and ethnic identity. *Appl. Dev. Sci.* **13**(2), 61–73 (2009)
20. El Ghouch, N., El Mokhtar, E.N., Seghroucheni, Y.Z.: Analysing the outcome of a learning process conducted within the system. *Int. J. Emerg. Technol. Learn.* **12**(3), 43–56 (2017)
21. Floridi, L., et al.: Ai4people, an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* **28**, 689–707 (2018)
22. Greene, M.L., Way, N., Pahl, K.: Trajectories of perceived adult and peer discrimination among black, Latino, and Asian American adolescents: patterns and psychological correlates. *Dev. Psychol.* **42**(2), 218–238 (2006)
23. Hakami, E., Hernandez-Leo, D.: How are learning analytics considering the societal values of fairness, accountability, transparency and human well-being?: a literature review. In: Martinez-Mones, A., Alvarez, A., Caeiro-Rodriguez, M., Dimitriadis, Y. (eds.) LASI-SPAIN 2020: Learning Analytics Summer Institute Spain 2020: Learning Analytics. Time for Adoption?, pp. 121–141. Aachen: CEUR, Valladolid, Spain (2020)
24. Hallamaa, J., Kalliokoski, T.: Ai ethics as applied ethics. *Front. Comput. Sci.* **4**, 776837 (2022)
25. Hamal, O., El Faddouli, N.E., Alaoui Harouni, M.H., Lu, J.: Sustainability. *Artif. Intell. Educ.* **5**, 1–11 (2022)
26. Hashim, H.: Application of technology in the digital era education. *Int. J. Res. Couns. Educ.* **1**(2), 1–5 (2018)
27. He, Y., Burghardt, K., Lerman, K.: A geometric solution to fair representations. In: 2020 AAAI/ACM Conference, pp. 279–285. ACM, New York (2020)
28. Heaven, P.C.L., Leeson, P., Ciarrochi, J.: Personality development at school: assessing a reciprocal influence model of teachers' evaluations and students' personality. *J. Res. Pers.* **43**(5), 815–821 (2009)
29. Holmes, W., et al.: The FATE of AIED, chap. Ethics of AI in Education: Towards a Community-Wide Framework, pp. 1–23. IJAIED Special Issue. Springer, Cham (2021)
30. Holsinger, D.B., Jacob, W.J.: Inequality in education: a critical analysis. In: Holsinger, D.B., Jacob, W.J. (eds.) Inequality in Education. CERC Studies in Comparative Education, vol. 24, pp. 1–33. Springer, Dordrecht (2008). [https://doi.org/10.1007/978-90-481-2652-1\\_1](https://doi.org/10.1007/978-90-481-2652-1_1)
31. Hutt, S., Duckworth, A.L., Gardner, M., D'Mello, S.K.: Evaluating fairness and generalizability in models predicting on-time graduation from college applications. In: The 12th International Conference on Educational Data Mining (EDM 2019), pp. 79–88. EDM, Montreal, Canada (2019)
32. Hwang, G.-J.: Definition, framework and research issues of smart learning environments - a context-aware ubiquitous learning perspective. *Smart Learn. Environ.* **1**(1), 1–14 (2014). <https://doi.org/10.1186/s40561-014-0004-5>
33. Kessler, R.C., Mickelson, K.D., Williams, D.R.: The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. *J. Health Soc. Behav.* **40**(3), 208–230 (1999)
34. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: Holmes, W., Porayska-Pomsta, K. (eds.) The Ethics of Artificial Intelligence in Education, pp. 1–30. Taylor and Francis (2022)

35. Liao, V., Gruen, D.M., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15. Association for Computing Machinery, New York (2020)
36. Marras, M., Boratto, L., Ramos, G., Fenu, G.: The FATE of AIED, chap. Equality of Learning Opportunity via Individual Fairness in Personalized Recommendations. IJAIED Special Issue. Springer, Cham (2021)
37. Mehrabi, N., Morstatter, F., Saxena, N., Galstyan, A., Lerman, K.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**(6), 115 (2021)
38. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: choices, assumptions, and definitions. *Annu. Rev. Stat. Appl.* **8**, 141–163 (2021)
39. Monteleoni, C.S., McQuade, S., Schmidt, G.A.: Climate informatics: accelerating discovering in climate science with machine learning. *Comput. Sci. Eng.* **15**, 32–40 (2013)
40. de Montreal, U.: Montreal declaration for a responsible development of artificial intelligence. Technical report, Universite de Montreal (2018)
41. Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., Baker, R.: Who's learning? Using demographics in EDM research. *J. Educ. Data Mining* **12**(3), 1–30 (2020)
42. Pinkwart, N.: Another 25 years of AIED? Challenges and opportunities for intelligent educational technologies of the future. *Int. J. Artif. Intell. Educ.* **26**(2), 771–783 (2016)
43. Qian, H., Rangwala, H.: Towards fair educational data mining: a case study on detecting at-risk students. In: Rafferty, A.N., Whitehill, J., Cavalli-Sforza, V., Romero, C. (eds.) The 13th International Conference on Educational Data Mining (EDM 2020), pp. 431–437. EDM (2020)
44. Raji, I.D., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: FAT\* 2020: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, pp. 33–44. ACM (2020)
45. Riazy, S., Simbeck, K.: Predictive algorithms in learning analytics and their fairness. In: Pinkwart, N., Konert, J. (eds.) DELFI 2019, pp. 223–228. Gesellschaft fur Informatik e.v., Boon, Germany (2019)
46. Roll, I., Wylie, R.: Evolution and revolution in artificial intelligence in education. *Int. J. Artif. Intell. Educ.* **26**(2), 582–599 (2016)
47. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019)
48. Schiff, D.: Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI Soc.* **36**(1), 331–348 (2021)
49. Serna, I., Morales, A., Fierrez, J., Obradovich, N.: Sensitive loss: improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artif. Intell. (AIJ)* **305**, 103682 (2022)
50. Singh, H., Miah, S.J.: Smart education literature: a theoretical analysis. *Educ. Inf. Technol.* **25**(4), 3299–3328 (2020). <https://doi.org/10.1007/s10639-020-10116-4>
51. Skaug Saetra, H.: A framework for evaluating and disclosing the ESG related impacts of AI with the SDGS. *Sustainability* **13**(15), 8503 (2021)
52. Spector, J.M., Merrill, M.D.: Section i foundations. In: Spector, J.M., Merrill, M.D., Elen, J., Bishop, M. (eds.) *Handbook of Research on Educational Communications and Technology*, pp. 3–113. Springer, New York (2014)
53. Srinivasan, V.: AI and learning: a preferred future. *Comput. Educ. Artif. Intell.* **3**, 100062 (2022)

54. Stone, S., Han, M.: Perceived school environments, perceived discrimination, and school performance among children of Mexican immigrants. *Child. Youth Serv. Rev.* **27**, 51–66 (2005)
55. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. In: *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)* (2019)
56. Tahiru, F.: Ai in education: a systematic literature review. *J. Cases Inf. Technol.* **23**(1), 1–20 (2021)
57. Tang, K.Y., Chang, C.Y., Hwang, G.J.: Trends in artificial intelligence-supported e-learning: a systematic review and co-citation network analysis (1998–2019). *Interact. Learn. Environ.* (2021)
58. Tomasev, N., et al.: Otake-matsuura: Ai for social good: unlocking the opportunity for positive impact. *Nat. Commun.* **11**, 2468 (2020)
59. Tzanakis, M.: Bourdieu's social reproduction thesis and the role of cultural capital in educational attainment: a critical review of key empirical studies. *Educate J.* **11**(1), 76–90 (2011)
60. Vinuesa, R., et al.: The role of artificial intelligence in achieving the sustainable development goals. *Nat. Commun.* **11**, 233 (2020)
61. Way, N., Rosenbloom, S.R.: Experiences of discrimination among African American, Asian American, and Latino adolescents in an urban high school. *Youth Soc.* **35**(4), 420–451 (2004)
62. Wong, C.A., Eccles, J.S., Sameroff, A.: The influence of ethnic discrimination and ethnic identification on African American adolescents' school and socioemotional adjustment. *J. Pers.* **71**(6), 1197–1232 (2003)
63. Xie, M., Jean, N., Burke, M., Lobell, D., Ermon, S.: Transfer learning from deep features for remote sensing and poverty mapping. In: *Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*, pp. 3929–3935. Association for the Advancement of Artificial Intelligence, Standford, USA (2016)
64. Zhu, Z.-T., Yu, M.-H., Riezebos, P.: A research framework of smart education. *Smart Learn. Environ.* **3**(1), 1–17 (2016). <https://doi.org/10.1186/s40561-016-0026-2>

# **Applied Natural Language Processing**



# Improving Collaboration via Automated Intelligent Nudges

Solomon Ubani<sup>(✉)</sup> and Rodney Nielsen

University of North Texas, Denton, TX 76203, USA

{solomon.ubani, rodney.nielsen}@unt.edu

**Abstract.** Micromanagement is one of the major obstacles of effective collaboration and therefore, pair programming. This is the first research to investigate a computational method that provides intelligent real-time nudges to avoid micromanagement with the aim of improving collaboration. We conducted human experiments using the computational method and statistically evaluated the effect of the method. Our method (implemented as a video chat system) resulted in 24% less micromanagement during pair programming when compared to scenarios without our method.

**Keywords:** ITS · CITS · Collaboration

## 1 Introduction

Remote pair programming (RPP) is an Agile software engineering practice that involves two programmers working together to develop software remotely. Like in traditional pair programming, RPP involves two actors—the programmer typing the code (the driver) and the programmer observing the code to catch any errors in the code or find room for improvement (the navigator). Optimally, both actors should switch roles regularly and articulate their questions and ideas [1].

Effective pair programming enhances code quality and improves the coding and communication skills of both programmers. Unfortunately, effective pair programming is sometimes hindered by poor communication and imbalanced dynamics. A specific obstacle that stems from both poor communication and imbalanced power dynamics is micromanagement<sup>1</sup>. Böckeler and Siessegger [2] noted that micromanagement does not leave room for a pair partner to think through the problem and can make pair programming both frustrating and counterproductive.

Some examples of micromanagement found in our data include:

- “Make sure it compiles and works as expected before we continue”.
- “Get rid of this, I guess just start seeding a random number”.
- “Just make sure it works”.

<sup>1</sup> Micromanagement is defined as “to manage or control with excessive attention to minor details” ([Dictionary.com](#)).

- “I already have a pretty good idea on how to do this so I guess I can just tell you what I’m doing as a do it or you know if you wanna add info”

The aim of this research is to implement and evaluate a computational method that provides automatic intelligent nudges with the aim of enhancing communication and balancing power by reducing micromanagement during collaboration. Thaler and Sunstein [3] defined a nudge as an intervention that modifies the actions of people in a foreseeable way without restricting their choices or incentivizing them. They noted that a nudge is easily avoidable. An example of a nudge is using posters to encourage people to eat healthily. However, a ban on junk food is not a nudge. Saghai [4] and Parkinson et al. [5] indicated that nudges influence people by cognitively stimulating them to exhibit desired behavior. In this paper, we research a novel computational method that provides automated intelligent nudges with the aim of improving collaboration called CollabAssist.

The main contributions of this research include:

This is the first research to investigate a computational method that provides real time nudges to users to avoid micromanagement. Specifically, we implement and evaluate CollabAssist as a video chat system that automatically detects micromanagement in pair-programming interactions within the context of student learning assignments and nudges students to avoid such issues.

**Research Question:** Do automated intelligent nudges provided by CollabAssist improve collaboration? Specifically, does CollabAssist lead to less micromanagement during pair programming when compared to scenarios without CollabAssist? In this paper, we describe the implementation of CollabAssist and the experiments designed to answer these research questions.

## 2 Related Work

### Application of Nudges in Education

Edwards and Li [6] utilized gamification to provide nudges that helped students develop self-regulatory skills. Dimitrova et al. [7] and Mitrovic et al. [8] employed nudges in a video-based learning system with the aim of increasing user engagement with educational videos without constraining users. Results showed that the nudges were able to foster desired behavior such as higher user engagement and better user experience. Stoyanova and Krämer [9] utilized nudges to resolve conflicts during group collaboration. The results showed that the system generated nudges were more effective than human peers in resolving conflicts. In this research work, we provide nudges with the aim of reducing micromanagement during pair programming.

### Feedback on Collaboration

Mercer [10] opined that productive and effective collaboration cannot occur without adequate support through feedback. This is supported by a study by Zakaria et al. [11] that evaluated the conversations of students during a pair-programming session and discovered that, organically, students use Cumulative (uncritical) talk more than Exploratory

(critical, constructive) talk. Zakaria et al. [12] successfully utilized human interventions to achieve more Exploratory talk by providing feedback during collaboration. In this research, we investigate how effective it is to utilize automated nudges to provide feedback on collaboration.

### Classifying Peer Interaction

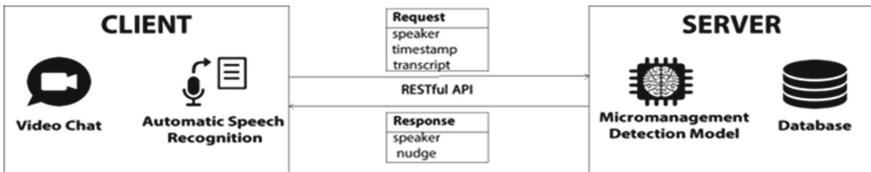
Some researchers have employed machine learning for dialogue analysis of student peer interaction. This includes, among others, models to detect off-task behavior [13], bullying [14], disruptive talk [15], micromanagement [16], disputational talk [17], exclusive language [18], and during collaboration. In this work, we build on micromanagement models developed by Ubani and Nielsen [16] to automatically detect micromanagement during pair programming.

### Collaborative Intelligent Tutoring Systems

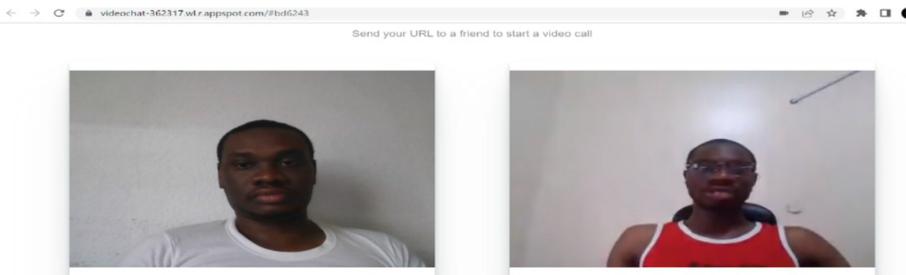
Collaborative Intelligent Tutoring Systems (CITS) aim to improve collaboration using conversational agents [19], collaborative agents [20], peer tutoring assistants [21], and interaction analysis toolkits [22]. Jeong et al. [23] noted that identifying and addressing disruptive collaborative states such as micromanagement can improve the quality of collaborative dialogue and consequently lead to more productive collaboration. Ubani and Nielsen [24] highlighted the need for more CITS that provide support after verbal interaction analysis. The CITS developed in this paper facilitates pair programming by detecting micromanagement and, where appropriate, providing real time nudges to reduce micromanagement. The aim of the system is to improve collaboration and, consequently, Computer Science skills. CollabAssist was implemented as a video chat system to be used in combination with a remote collaborative software development system such as Google Colab or Microsoft Live Share.

## **3 System Implementation**

The video chat is a cloud-based application based on the client-server architecture. Each student's video chat client connects to a central server that hosts the classifier trained to detect micromanagement. The web technologies used for the development of the video chat client include Javascript, AJAX, HTML5 and CSS3. We used IBM Watson's speech-to-text software to convert audio input from the user's microphone to text. WebRTC was used to enable real time video and voice communication between two users. The technologies used for the development of the video chat server include MySQL, Python, and NodeJS. The client communicates with the server synchronously through a RESTful API. The server hosts the machine learning model that detects micromanagement in transcripts of the users (see Machine Learning Model). When micromanagement is detected on the server, it relays that to the client which then provides a nudge to the user. The architecture of CollabAssist is displayed in Fig. 1. The client user interface of CollabAssist is shown in Fig. 2.



**Fig. 1.** Architecture of CollabAssist.



**Fig. 2.** User Interface of CollabAssist.

## 4 Machine Learning Model

The machine learning model was trained by fine-tuning a pretrained transformer model, BERT. We chose BERT because it required 27% less computational resources than newer models to deploy in production environments while also demonstrating similar classification performance when compared to newer transformer models like DeBERTa. BERT was fine-tuned with a training time of three epochs, batch size of 64, using the AdamW optimizer with a learning rate of  $5 \times 10^{-5}$ . We used categorical cross entropy as the loss function and set the maximum sequence length to 75 tokens. Before tokenization, we preprocessed each dialog span<sup>2</sup> to eliminate repeated word disfluencies and incorrect word cases stemming from the speech-to-text software. Full details about the model and the data used to train the model can be found in the paper written by Ubani and Nielsen [16].

Since the primary aim of the computational model is to improve collaboration by detecting micromanagement and providing occasional nudges to students to avoid its use, a higher precision is more important than a higher recall. This is because, in practice, we do not need to nudge a student at every single occurrence of micromanagement—it is enough to provide occasional corrective nudges to the students. In this research, we hypothesize that one to three nudges would be adequate to achieve the desired effect of less micromanagement. Hence, the model should retrieve one to three instances of micromanagement for each student. We also want to avoid false positives as much as possible and only nudge students when the model is confident since false positives may

<sup>2</sup> Audio of every user was captured on the client side every 10 s. A dialog span in this research is a 10 s period that included at least a word from that user.

be counterproductive. To this end, we introduce a novel metric called the precision over  $C$ , where  $C$  is the confidence score of the model’s prediction.

We evaluated the model investigated by Ubani and Nielsen [16] using the precision over  $C$  metric on the test dataset described by Ubani and Nielsen [16]. The model achieved a precision over  $C = 90\%$  of 0.69 while retrieving a minimum of three instances of micromanagement for each student on the test dataset.

The confidence score of the model is not well calibrated. This is consistent with other Neural Network-based models. Guo et al. [25] explored some Neural Network confidence score calibration techniques which include binning methods, matrix and vector scaling, and temperature scaling that might improve the calibration of our model. However, calibrating the model is beyond the scope of this research since, in practice, our system only provides occasional nudges when it meets a high confidence threshold. Hence, the overconfidence of our model does not have a deleterious effect on our system.

## 5 Experiments

### 5.1 Participants

The study took place in an undergraduate class in one of the 50 largest Universities in the USA. There were 30 participants, and they were all between the ages of 18 and 35. All participants were either of the Caucasian or Asian race. Six (20%) of the participants were female while 80% were male. Every participant had at least basic knowledge of computer programming. The students participated to earn extra credit in their class. The protocol for this study was approved by the university’s Institutional Review Board.

### 5.2 Pair Programming Sessions and Tasks

There were two programming tasks of fairly equal difficulty. In each programming task, the students used Visual Studio Live Share to write code together in real time and used our audio-video software to communicate with each other.

### 5.3 Method

In our experiments, we explored two scenarios: a no-nudge control versus an intelligent-nudge treatment.

**No-Nudge Control:** In this scenario, the students pair programmed with CollabAssist but they received *no* nudges.

**Intelligent-Nudge Treatment:** In this scenario, the students engaged with the full capabilities of CollabAssist. The CollabAssist system analyzed students’ dialogue as they pair programmed and nudged students to avoid micromanagement whenever it was over 90% confident that the student was micromanaging their partner. Table 1 shows the distinct nudges that were given to students by CollabAssist. Figure 3 shows one of the four nudges that were given to students when CollabAssist detected micromanagement. The student received one of the nudges in Table 1 selected at random. A given student never received the same nudge more than once in a pair-programming session if the CollabAssist system detected micromanagement four times or less.



**Fig. 3.** A nudge displayed to the user

**Table 1.** All distinct nudges used by the system.

	Nudge
1	Plan together
2	Ask for opinions
3	Negotiate decisions
4	Think collaboratively

#### 5.4 Condition Assignments and Paired t-Test Dyads

The first pair-programming session served as a baseline to estimate each student's rate of micromanagement. In the first pair-programming session, we randomly paired all the students. No student received a nudge. At the end of this session, we utilized the micromanagement detection model to classify all the dialog spans of each student. We then ranked the students by the number of dialog spans classified as micromanagement, breaking ties based on their rate of micromanagement<sup>3</sup>. Then, for each dyad of adjacently ranked students, we randomly placed one in the treatment (intelligent-nudge) group and the other in the control (no-nudge) group.

The student dyads described above were designated as the eventual paired t-test dyads, resulting in 14 paired t-test dyads. In the second pair-programming session, we generated programming pairs by randomly pairing students within their assigned groups (treatment versus control). One control-group pair withdrew from the experiment before the second session. Hence, the associated paired t-test dyads were excluded from the final analysis.

---

<sup>3</sup> Dialog spans classified as micromanagement/dialog spans that had at least one word.

## 5.5 Study Data

To answer the research questions of our study and evaluate CollabAssist two annotators manually labeled the dialog spans of the second pair-programming session as *Micromanagement*, *Good Collaboration* or *Other*<sup>4</sup>. The two annotators achieved a Cohen's Kappa inter-rater reliability score of  $K = 0.84$  characterized as almost perfect agreement by Landis and Koch [26]. At a more fine-grained level, the annotators achieved a Cohen's Kappa inter-rater reliability score of  $K = 0.62$  for *Micromanagement* and  $K = 0.64$  for *Good Collaboration* characterized as substantial agreement by Landis and Koch [26]. A third analyst served as the adjudicator. The 28 students that participated in session 2 had a total of 2535 ( $M = 90.5$ ,  $SD = 32.8$ ) dialog spans in the adjudicated dataset. A total of 103 ( $M = 3.7$ ,  $SD = 2.6$ ) dialog spans were labeled as *Micromanagement*, 106 ( $M = 3.8$ ,  $SD = 2.5$ ) dialog spans were labeled as *Good Collaboration* while the remaining 2326 dialog spans were labeled as *Other*. The pair programming sessions lasted an average of 31 min.

## 5.6 Model Evaluation

The model achieved a Precision over  $C = 0.90$  of 0.68 while also retrieving at least one instance of micromanagement for each student. Error analysis of the model revealed that the model never erroneously classified *Micromanagement* as *Good Collaboration* or vice-versa. This was reassuring since it means that the students were never nudged to avoid micromanagement at a time when they exhibited desirable collaboration etiquette.

## 5.7 Experiment Evaluation

To evaluate CollabAssist, we utilized quantitative methods to compare the number of session-2 dialog spans labeled as micromanagement between treatment and control groups that were matched for micromanagement in a baseline session. We hypothesized that in the second pair-programming session, the students that receive intelligent nudges from CollabAssist would use statistically less micromanagement ( $p < 0.05$ ) than the students that received no nudges.

# 6 Results

Did the students in the experimental/treatment group have less micromanagement than the students in the control group?

The evaluation of students' dialog spans in the second pair programming sessions revealed that indeed the students in the treatment group achieved significantly less micro-management when compared to the students in the control group (paired t-test;  $p = 0.025$ ). Specifically, the average number of dialog spans including micromanagement for students in the treatment group was 24% less than in the control group. The students

---

<sup>4</sup> We followed the guidelines defined by Ubani and Nielsen [16] but used the category *Other* rather than *Irrelevant* for dialogue that was neither indicative of good collaboration nor micromanagement.

in the treatment group had fewer instances of micromanagement in all but four paired t-test dyads. In one of the four paired t-test dyads, the control student had fewer instances while in the other three paired t-test dyads, the control and treatment student had an equal number of micromanagement instances.

## 7 Discussion

The quantitative evaluation of CollabAssist provides evidence of its effectiveness in reducing micromanagement during remote pair programming sessions. Additionally, we discuss some qualitative evaluations of the system with the aim of providing helpful insights future researchers could build on.

- Did False Positive instances of micromanagement have a deleterious effect?

Examining the four paired t-test dyads where there was not a reduction in micromanagement in the treatment group, two of the students in the treatment group received a nudge that stemmed from a false positive instance of micromanagement in the first half of the programming session. Perhaps, this implies that a false positive nudge early in the session led the students to distrust the system and ignore future nudges. Consequently, an increase in the precision of the model may lead to better performance of CollabAssist in improving collaboration by reducing micromanagement.

- How much additional reduction of micromanagement could be achieved by an improvement in model recall?

Examining only the four paired t-test dyads where there was not a reduction in micromanagement in the treatment group, three of the students in the treatment group received their first nudge in the second half of the programming session. Reducing the confidence threshold of the model to 80%, retroactively, revealed that two of the three students would have received a nudge in the first half of the programming session. This could have led to more reduction in micromanagement in the remainder of the programming session. However, as seen in the previous bullet point, a lower model confidence threshold will lead to more false positives which would be deleterious. Future researchers could explore other methods to improve the recall of the model without significantly reducing the precision of the model.

A 10% absolute decrease in the confidence threshold of the model would not have had an effect on the other one student. This could be because the total number of dialog spans the student had was 15 less than the mean number of dialog spans of all the students. This implies relatively low participation from the student. Future researchers could also explore nudging students who are not actively participating in the programming session.

- Is there a difference in the nature of the micromanagement that does occur after the four different types of nudges we provided?

We did not detect evidence that there was a difference in the nature of micromanagement that occurs after any specific nudge shown in Table 1. However, this is likely

because of the stochastic nature of the nudges that the system provided. The stochastic nudges produced by the system did not take into account the type of micromanagement that occurred. On close inspection of the data from this research, the following are some common types of micromanagement that were discovered.

Interruptions: One programmer interrupts the other while they are either talking or driving the programming session without giving them the opportunity to complete their thought or statement.

Example: “Make sure it compiles and works as expected before we continue”.

Directives: One programmer giving explicit instructions or orders to the other without allowing for input or discussion.

Example: “Get rid of this, I guess just start seeding a random number”.

Unsolicited advice: One programmer providing unsolicited advice or suggestions to the other without being asked or without allowing for input or discussion.

Example: “Just make sure it works”.

Driver role usurping: The navigator abruptly assumes the role of the driver without discussing or agreeing to it first.

Example: “I already have a pretty good idea on how to do this so I guess I can just tell you what I’m doing as a do it or you know if you wanna add info”.

It is noteworthy that the types of micromanagement listed above are not mutually exclusive. However, future researchers would benefit from tailoring the nudges to the type of micromanagement that occurred. This would enable proper evaluation of the effect of the different nudges provided by the system.

- Did the students discussing the nudges during the sessions have an effect?

There is evidence that two of the 14 students in the treatment group discussed a nudge they received. However, the sample size ( $N = 2$ ) is too small to achieve an adequate statistical power of a test. The limited sample size probably stemmed both from a small number of participants as well as having short programming sessions that spanned about 30 min. We hypothesize that if the students discussed the nudges and the behavior that triggered them, it could lead to increased awareness of their own behavior and an increased effort to change it. Additionally, if the students ignored the nudges or did not take them seriously, it could lead to less or no change in their behavior. Future researchers with a larger number of participants and longer programming sessions could investigate the effect of discussing the nudges during the pair programming session and how that could impact the reduction of micromanagement.

It is noteworthy that while CollabAssist currently provides feedback to reduce micromanagement, it is a robust and model-independent system that can provide feedback on other forms of collaborative behaviors. For example, future researchers can explore using CollabAssist to reduce off task behavior [13], bullying [14], disruptive talk [15], disputational talk [17], or exclusive language [18] during collaboration.

## 8 Conclusion and Future Work

Effective collaborative learning is well known to lead to learning gains. However, collaborative learning when executed ineffectively can be frustrating and counterproductive.

This research provides an effective method of improving collaboration by reducing micromanagement. This method is also feasible in remote environments. Our current work introduces a new Collaborative Intelligent Tutoring System, CollabAssist, that uses a machine learning model to detect micromanagement in real time and then nudges students to avoid such micromanagement. The experiments conducted to evaluate the system revealed the system's effectiveness in reducing micromanagement; treatment students had 24% fewer dialog spans with micromanagement than their matched-peers in the control group. Our qualitative evaluation also provides insights that could be used to improve the system in the future. Future researchers can conduct a longitudinal study that investigates the effect that CollabAssist has over multiple programming sessions. Additionally, future researchers can investigate the impact CollabAssist has across genders and races.

## References

1. Hughes, J., Walshe, A., Law, B., Murphy, B.: Remote pair programming. In: CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education, vol. 2, pp. 476–483 (2020). <https://doi.org/10.5220/0009582904760483>
2. Böckeler, B., Siessiger, N.: On pair programming. Martin Fowler (2020)
3. Thaler, R. H., Sunstein, C. R.: *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin (2009)
4. Saghai, Y.: Salvaging the concept of nudge. *J. Med. Ethics* **39**(8), 487–493 (2013)
5. Parkinson, J.A., Eccles, K.E., Goodman, A.: Positive impact by design: the wales centre for behaviour change. *J. Posit. Psychol.* **9**(6), 517–522 (2014)
6. Edwards, S.H., Li, Z.: A proposal to use gamification systematically to nudge students toward productive behaviors. In: Koli Calling 2020: Proceedings of the 20th Koli Calling International Conference on Computing Education Research, pp. 1–8 (2020)
7. 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 (2017)
8. Mitrovic, A., Gordon, M., Piotrkowicz, A., Dimitrova, V.: Investigating the effect of adding nudges to increase engagement in active video watching. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) *AIED 2019. LNCS (LNAI)*, vol. 11625, pp. 320–332. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-23204-7\\_27](https://doi.org/10.1007/978-3-030-23204-7_27)
9. Stoyanova, F., Krämer, N.: Please don't shoot the messenger! Prompts in online learning groups-influences of nudging messages' sender and publicness on recipients' perception and attribution (2019)
10. Mercer, N.: *The Guided Construction of Knowledge: Talk Amongst Teachers and Learners. Multilingual Matters* (1995)
11. Zacharia, Z., et al.: Collaborative talk across two pair-programming configurations. In: *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings*, 2019 13th International Conference on Computer Supported Collaborative Learning (CSCL), vol. 1 (2019)
12. Zakaria, Z., et al.: Two-computer pair programming: exploring a feedback intervention to improve collaborative talk in elementary students. *Comput. Sci. Educ.* **32**(1), 3–29 (2022)
13. Carpenter, D., et al.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) *AIED 2020, Part I. LNCS (LNAI)*, vol. 12163, pp. 55–66. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-52237-7\\_5](https://doi.org/10.1007/978-3-030-52237-7_5)

14. Nikiforos, S., Tzanavaris, S., Kermanidis, K.-L.: Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *J. Comput. Educ.* **7**(4), 531–551 (2020). <https://doi.org/10.1007/s40692-020-00166-5>
15. Park, K., et al.: Detecting disruptive talk in student chat-based discussion within collaborative game-based learning environments. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 405–415 (2021)
16. Ubani, S., Nielsen, R.: Detecting micromanagement during pair programming. In: 2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 01–07. IEEE (2021)
17. Ubani, S., Nielsen, R.: Classifying different types of talk during collaboration. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) AIED 2022, Part II. LNCS, vol. 13356, pp. 227–230. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-11647-6\\_40](https://doi.org/10.1007/978-3-031-11647-6_40)
18. Ubani, S., Nielsen, R., Li, H.: Detecting exclusive language during pair programming. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
19. Tegos, S., Demetriadis, S., Tsatsos, T.: A configurable conversational agent to trigger students' productive dialogue: a pilot study in the CALL domain. *Int. J. Artif. Intell. Educ.* **24**, 62–91 (2014)
20. Casamayor, A., Amandi, A., Campo, M.: Intelligent assistance for teachers in collaborative e-learning environments. *Comput. Educ.* **53**(4), 1147–1154 (2009)
21. 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**, 279–306 (2011)
22. Michailidis, N., Kapravelos, E., Tsatsos, T.: Interaction analysis for supporting students' self-regulation during blog-based CSCL activities. *J. Educ. Technol. Soc.* **21**(1), 37–47 (2018)
23. Jeong, H., Hmelo-Silver, C.E.: Seven affordances of computer-supported collaborative learning: how to support collaborative learning? How can technologies help? *Educ. Psychol.* **51**(2), 247–265 (2016)
24. Ubani, S., Nielsen, R.: Review of collaborative intelligent tutoring systems (CITS) 2009–2021. In: 2022 11th International Conference on Educational and Information Technology (ICEIT), pp. 67–75. IEEE (2022)
25. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
26. Landis, J.R., Koch, G.G.: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 363–374 (1977)



# Preliminary Performance Assessment on Ask4Summary's Reading Methods for Summary Generation

Rita Kuo<sup>1</sup> , Maria F. Iriarte<sup>2</sup> , Di Zou<sup>3</sup> , and Maiga Chang<sup>2</sup>

<sup>1</sup> Utah Valley University, Orem, UT, USA

<sup>2</sup> Athabasca University, Edmonton, AB, Canada

[maiga.chang@gmail.com](mailto:maiga.chang@gmail.com)

<sup>3</sup> The Education University of Hong Kong, Tai Po, Hong Kong

**Abstract.** Ask4Summary creates summary for students' questions based on text-based learning materials. This study conducts a preliminary assessment on Ask4Summary's performance in terms of generating summaries with different subsets of course materials (e.g., supplement academic papers in PDF only, notes and slides in Word and PowerPoint only, and everything the teacher provides for the students) read and processed by two reading methods: the built-in algorithm based on Python NLTK and AWS Comprehend Keyphrase Extraction and Syntax Analysis. The course materials of a graduate level Academic Writing in English course in an Asian university and twenty-six common questions that students may ask in the class are provided by the course instructor. Each of the questions are read via the two methods and Ask4Summary generates the summaries with the six different datasets created by: (1) Python NLTK reading the academic papers in PDF only; (2) Python NLTK reading notes and slides in Word and PowerPoint format only; (3) Python NLTK reading every course materials; (4) AWS Comprehend reading academic papers in PDF only; (5) AWS Comprehend reading notes and slides in Word and PowerPoint format only; and (6) AWS Comprehend reading every course materials. For the 312 queries (i.e., ask 26 questions in 6 datasets with 2 methods analyzing the questions) made, 117 queries successfully generated the summary, where only 2 of them were read by AWS Comprehend. Among the rest of 115 summaries, 67 of them are from the datasets created via the built-in algorithm and 48 are from the datasets created by AWS Comprehend.

**Keywords:** Language Learning · NLTK · AWS · Natural Language Processing · Learning Materials

## 1 Introduction

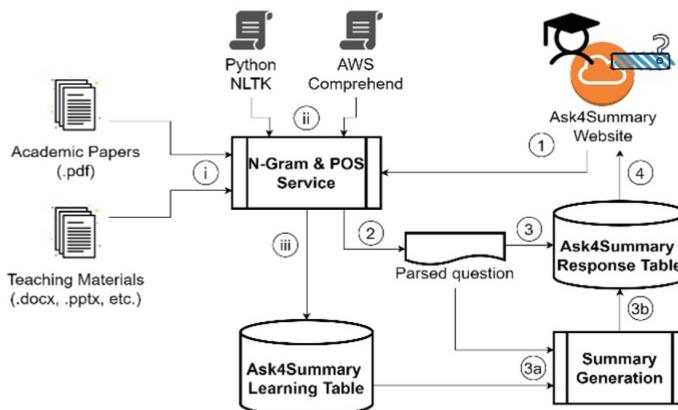
Information overload is caused by the highly increased education resources and makes users in the online learning environment spend huge time in searching for the suitable education resources [6]. Helping students retrieve important information from the education resources becomes an important research area in the educational technology domain,

such as multi-document, user-specific, and innovative text summarization applications [2]. The research team has developed the Ask4Summary [4, 5] to automatically responds students' questions by generating summaries from the course content to help students quickly retrieve the important information, especially when online learning is getting more popular after the pandemic.

To enhance the service, we would like to know which Natural Language Processing (NLP) toolkit and types of course content can generate the summaries successfully. Therefore, the research team worked with the teacher in the Academic Writing in English course in an Asian university to evaluate the system. Section 2 briefly introduces the NLP toolkits used in the service and the system workflow of Ask4Summary. The evaluation method is sketched in Sect. 3. Section 4 explains the analysis results based on the collected data. A brief summary and the future works are concluded in Sect. 5.

## 2 Ask4Summary

The research team created an Ask4Summary website<sup>1</sup> for a graduate level Academic Writing in English course in an Asian university to evaluate the performance of summary generation. Ask4Summary first reads and processes (see Stage i in Fig. 1) the text-based materials that the instructor used in the course. The materials include supplement academic papers in PDF and notes and slides in Word and PowerPoint.



**Fig. 1.** The system workflow of Ask4Summary.

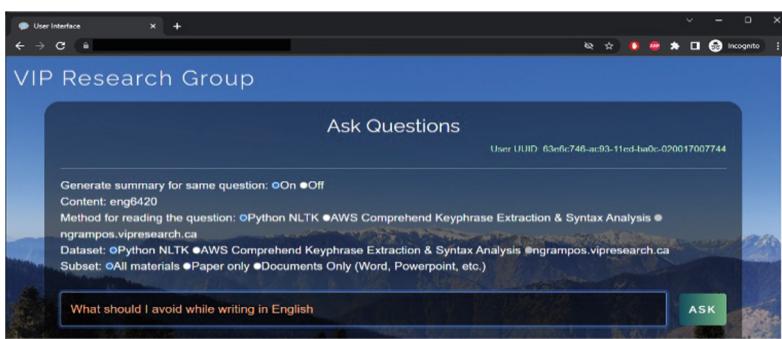
Ask4Summary identifies the valid N-grams according to their Part-of-Speech (PoS) tags [3]. The research team selects two widely used NLP toolkit/service to support the built-in algorithm, which are Python NLTK [1] and AWS Comprehend Keyphrase Extraction and Syntax Analysis<sup>2</sup> (see Stage ii in Fig. 1). With the NLTK and AWS Comprehend's help, Ask4Summary stores the extracted N-grams in the Ask4Summary

<sup>1</sup> <https://eng6420.ask4summary.vipresearch.ca/>.

<sup>2</sup> <https://aws.amazon.com/comprehend/features/>.

Learning Table with the original content (see Stage iii in Fig. 1). The toolkit used in the feature extraction is also recorded in the Ask4Summary Learning Table so teachers or students are able to select which dataset generated by different toolkits they would like to use to generate the summary.

After the text-based learning materials are read and processed, teachers and students can use Ask4Summary (see Stage 1 in Fig. 1) on the website. They can enter their course related question as Fig. 2 shows. The users can choose which “brain” they want Ask4Summary to use for generating summaries. The “brains” are different Ask4Summary Learning Tables created earlier with either NLTK or AWS Comprehend method reading different subsets of learning materials: the academic paper in PDF format only, notes and slides in Word and PowerPoint format only, and all course materials.



**Fig. 2.** Users can ask questions related to the Academic Writing in English course on the Ask4Summary website.

There is another table – Ask4Summary Response Table – used to store the questions users asked before as well as the questions’ N-grams & PoS tags and the generated summaries. After a question’s valid N-grams have been extracted (see Stage 2 in Fig. 1), Ask4Summary first checks whether or not similar question in Ask4Summary Response Table has been asked before (see Stage 3 in Fig. 1). If similar question exists, Ask4Summary simply retrieves the past generated summaries and delivers to the users (see Stage 4 in Fig. 1). On the other hand, it uses Cosine Similarity to find the top Y sentences in the top X documents, where X and Y are pre-defined and can be adjusted, are related to the question for generating the summaries (see Stage 3a in Fig. 1).

The generated summary will be saved in the Ask4Summary Response Table (see Stage 3b) and dispatches the summary to the user (see Stage 4). Figure 3 shows the summary generated by the system with the reading method Python NLTK regarding the question “What should I avoid while writing in English.” The Ask4Summary first uses Python NLTK to analyze the question entered. According to the selection of the “brain” types in Fig. 2 – that is, using Python NLTK to read all course material and storing the data into the Ask4Summary Learning Table, the generated summary is: “1) Avoid informal verbs. 2) Avoid colloquial expressions. Avoid informal or colloquial expressions such as a lot of in your written assignments. 3) Avoid clichés. Avoid using the expressions below in your written work.” Users can give feedback regarding their

perceived relevance (scaled from 0 to 10) and perceived satisfaction (scaled from 0 to 10) toward the generated summary.

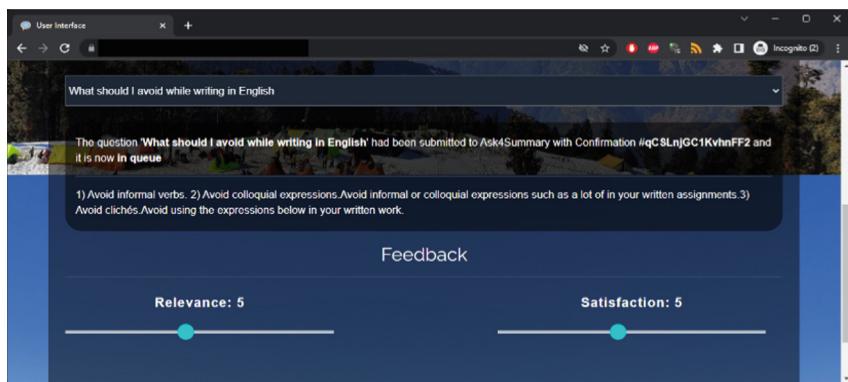


Fig. 3. The summary of the question generated by the Ask4Summary website.

### 3 Evaluation Design

The Academic Writing in English course is used for assessing the success rate of generating summaries for the course related questions. The course has 45 documents fed into Ask4Summary, include academic papers in PDF and the notes and slides in Word and PowerPoint described in Sect. 2. The course instructor also provides 26 common questions that students in the class usually ask, such as “*what are the differences between academic and non-academic genres*,” “*what is plagiarism*,” and “*what are the common stages of the conclusion*.”

In the evaluation, Ask4Summary generates the summary from six “brains” pre-created by using: (1) Python NLTK reading the academic papers in PDF format only; (2) Python NLTK reading notes and slides in Word and PowerPoint format only; (3) Python NLTK reading every course materials; (4) AWS Comprehend reading academic papers in PDF only; (5) AWS Comprehend reading notes and slides in Word and PowerPoint format only; and (6) AWS Comprehend reading every course materials. The summary generation results are saved as the format shown in Table 1.

Each query has its own *id* generated by the service and the questions are saved in the *question* field. The method used in reading question is saved in the *read\_tool* column and the method used in reading and processing the course materials is saved in the *content\_tool* column – the reading method could be *NLTK* (Python NLTK) or *AWS* (AWS Comprehend). Moreover, the course material subset used for generating summaries is recorded in the *source* column, which annotates *paper* (academic papers in PDF format only), *document* (notes and slides in Word and PowerPoint format only), or *all* (all the course materials). The column, *success*, indicates whether or not Ask4Summary generated the summary successfully and the generated summaries will be kept in the *summary* field.

**Table 1.** The generated summary records for queries (partial).

Id	question	read_tool	content_tool	source	success	summary
2	What is genre	NLTK	NLTK	paper	yes	Biber (2006), for example, shows us that ...
25	What are the differences between academic and non-academic genres	AWS	NLTK	all	no	
42	What is nominalization	NLTK	AWS	document	yes	Nominalization refers to ...
...	...		...	...	...	

## 4 Preliminary Assessment Results

First, the success rate in generating summaries is only 37.50% (117 of 312 queries); only 2 of the success queries were using AWS Comprehend as reading method for processing the questions. A successful query in this preliminary assessment means that Ask4Summary can generate summary for the asked question without considering how relevance and/or how satisfaction the generated summary could bring to the user who asked the question.

The research applied Chi-square test on the dataset with SPSS 28.0 showing that using Python NLTK to read questions has significantly higher successful rate (73.72%, 115 of 156) in terms of generating summaries than using AWS Comprehend (1.28%, 2 of 156) where  $\chi^2(1, n = 312) = 174.619, p < .001$ . Among the 117 success queries, 67 summaries were generated from the datasets that were built by using Python NLTK to read course materials and 50 of them were built by using AWS Comprehend.

Because the success queries that use AWS Comprehend to read questions are too few (i.e., only two), the preliminary assessment then focuses on the 156 queries that use Python NLTK to read the 26 questions and would like to see if there is a significant difference in terms of successfully generating summaries with different datasets built by using Python NLTK and AWS Comprehend reading methods, based on subsets of materials. The Chi-square test result is  $\chi^2(1, n = 156) = 11.944, p < .001$ . According to the data in Table 2, the datasets – despite of which subset of materials were using – built by Python NLTK performs better than AWS Comprehend.

To figure out which subset of materials is better for Ask4Summary to generate summaries for the 26 common questions that students would ask in the course, despite of the methods used to read subsets of materials the Chi-square test is applied in the queries that only use Python NLTK reading questions. The Chi-square test reveals significant differences –  $\chi^2(2, n = 156) = 6.419, p = .040$  – on generating summaries with the datasets built by using different subsets of course materials, *paper* (academic papers), *document* (notes and slides), and *all* (all the course materials). The result only shows

**Table 2.** The Chi-square test for the success rate in data sets using different NLP toolkits in the built-in algorithm when reading questions using Python NLTK

	No	Yes	Total
NLTK	11 (14.10%)	67 (85.90%)	78
AWS	30 (38.46%)	48 (61.54%)	78
Total	41	115	156

that there is a significant difference among the subsets but not telling us which subset is significant having better performance than the others.

The Bonferroni correction is applied on the original Chi-square test and the results are listed in Table 3. Both *paper* and *document* categories are annotated with a subscript – *a*, indicating that there is no significant difference on having success summary generation rates between *paper* and *document* as material source. Similarly, there is no significant difference between using documents only and using all materials as the source when Bonferroni correction annotates a subscript – *b* – on both *document* and *all* categories. However, the category *paper* and *all* have different subscript annotations, suggesting that there is a significant difference between the use of academic papers only and the use of all materials; the results show that the use of all materials has better performance with 82.69% success rate than only using academic papers (61.54%) while generating summaries.

**Table 3.** The Bonferroni correction applied in Chi-square test when comparing the success rate of generating summaries with different subsets of materials.

		Paper	Document	All	Total
No	Count	20 <sub>a</sub>	12 <sub>a, b</sub>	9 <sub>b</sub>	41
	% within generation	38.46%	23.08%	17.31%	26.28%
Yes	Count	32 <sub>a</sub>	40 <sub>a, b</sub>	43 <sub>b</sub>	115
	% within generation	61.54%	76.92%	82.69%	73.72%
Total	Count	52	52	52	156
	% within generation	100.0%	100.0%	100.0%	100.0%

We would further like to investigate the influence that subsets might have impact on the success rate of summary generation when considering reading methods separately. While still only considering the use of Python NLTK to read question, the Chi-square test shows that there is no significant difference on the success rate of generating summary among those subsets all read by using Python NLTK earlier with  $\chi^2(2, n = 78) = 2.752$ ,  $p = .253$ . Similarly, the Chi-square test also shows that there is no significant difference among the subsets read by using AWS Comprehend with  $\chi^2(2, n = 78) = 4.225$ ,  $p = .0.121$ . Table 4 lists the Bonferroni correction. All the three subsets are annotated with

a subscript – *a*, indicating that there is no significant difference among the subset after applied the Bonferroni correction on the Chi-square test.

**Table 4.** The Bonferroni correction applied to Chi-square tests when comparing the success rate of generating summaries with different subsets read by different reading methods.

			Paper	Document	All	Total
NLTK	No	Count	6 <sub>a</sub>	3 <sub>a</sub>	2 <sub>a</sub>	11
		% within generation	23.08%	11.54%	7.69%	14.10%
	Yes	Count	20 <sub>a</sub>	23 <sub>a</sub>	24 <sub>a</sub>	67
		% within generation	76.92%	88.46%	92.31%	85.90%
	Total	Count	26	26	26	78
		% within generation	100.0%	100.0%	100.0%	100.0%
AWS	No	Count	14 <sub>a</sub>	9 <sub>a</sub>	7 <sub>a</sub>	30
		% within generation	53.85%	34.62%	26.92%	38.46%
	Yes	Count	12 <sub>a</sub>	17 <sub>a</sub>	19 <sub>a</sub>	48
		% within generation	46.15%	63.38%	73.08%	61.54%
	Total	Count	26	26	26	78
		% within generation	100.0%	100.0%	100.0%	100.0%

## 5 Discussion and Future Works

According to the analysis results of the preliminary assessment, using AWS Comprehend to read questions has a very low success rate (1.28%) in generating summaries. It is because the questions are usually short and AWS Comprehend cannot properly extract and identify key phrases from short sentences. When only considers the use of Python NLTK to read questions, the overall success rate of generating summaries for the 26 course related questions is increasing to 73.72% (115 of 156) from 37.50%. The datasets built by using reading method Python NLTK also have better performance (85.90% success rate for generating summaries) compared to AWS Comprehend's 61.54% – despite of which material subset was used. This higher summary generation rate might be caused due to the use of same method for reading the questions and course materials.

The assessment also shows that Ask4Summary has higher success rate in generating summaries from the subset pre-built with more course materials: it has a significantly higher success rate (82.69%) from the subset built with both academic papers and teaching materials than from the subset built with only academic papers (61.54%). However, when we consider the method used for reading the subsets of course materials separately, we find that there is no significant difference in terms of the success rate between the use of academic paper only and the use of everything. The Chi-square tests and Bonferroni correction results (see Table 4) on one hand show the course instructor that no material

is more than the others and all materials included in her course are equally important for her students and lectures. On the other hand, the results also show more materials read by Ask4Summary no matter which reading method it uses, better chance it can generate summaries for student questions.

The follow-up works the research team right now works on is to ask the course instructor to evaluate the summaries generated by both Ask4Summary and ChatGPT in terms of the perceived relevance and satisfaction toward a question-summary pair service. Through the follow-up works we can figure out (1) either Python NLTK or AWS Comprehend is more suitable for Ask4Summary's summary generation algorithm, (2) which subset of course materials can help Ask4Summary to generate more relevant and better summaries, and (3) the course instructor's perceptions toward Ask4Summary and ChatGPT and how different the perceptions would be. We can also analyze those summaries that receive lower perceived relevance and satisfaction from the teacher to understand how to improve the Ask4Summary algorithm.

**Acknowledgements.** The authors acknowledge the support of Athabasca University's IDEA Lab and Mitacs Globalink program.

## References

1. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol (2009)
2. El-Kassas, W.S., Salama, C.R., Refea, A.A., Mohamed, H.K.: Automatic text summarization: a comprehensive survey. *Expert Syst. Appl.* **165**, 113679 (2021)
3. Pal, S., Chang, M., Iriarte, M.F.: Summary generation using natural language processing techniques and cosine similarity. In: Abraham, A., Gandhi, N., Hanne, T., Hong, T.P., Nogueira Rios, T., Ding, W. (eds.) Intelligent Systems Design and Applications. ISDA 2021. LNNS, vol. 418, pp. 508–517. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-96308-8\\_47](https://doi.org/10.1007/978-3-030-96308-8_47). [removed for blind review]
4. Saleh, M., Chang, M., Iriarte, M.F.: Ask4Summary automatically responds student's question with a summary assembled from course content. In: Proceedings of the 30th International Conference on Computers in Education (ICCE 2022), Kuala Lumpur, Malaysia (Hybrid), vol. 2, pp. 413–418 (2022). [removed for blind review]
5. Saleh, M., Iriarte, M.F., Chang, M.: Ask4Summary: a summary generation moodle plugin using natural language processing techniques. In: Proceedings of the 30th International Conference on Computers in Education, (ICCE 2022), Kuala Lumpur, Malaysia (Hybrid), vol. 1, pp. 549–554 (2022). [removed for blind review]
6. Taratukhina, Y.V., Bart, T.V., Vlasov, V.V.: Machine learning models of information recommendation system on individualization of education. *Educ. Resour. Technol.* **2**(2), 7–14 (2019)



# Joint Topic Model with Selected Side Information for Inter-University Syllabus Analysis Focusing on the Variety of Instructional Approaches

Kentaro Shiga<sup>1</sup> and Naoyuki Morimoto<sup>2(✉)</sup>

<sup>1</sup> Graduate School of Information Science and Technology, Osaka University,  
Osaka, Japan

[shiga.kentaro@ist.osaka-u.ac.jp](mailto:shiga.kentaro@ist.osaka-u.ac.jp)

<sup>2</sup> Graduate School of Engineering, Mie University, Mie, Japan  
[morimoto@info.mie-u.ac.jp](mailto:morimoto@info.mie-u.ac.jp)

**Abstract.** In recent university education, instructional approaches such as active learning have been varied, and inter-university opportunities to take classes have expanded due to the spread of MOOCs. Therefore, students have to refer to many syllabi, which significantly increases their burden when taking classes. In order to support syllabus browsing, a method of inter-university analysis of syllabi should be useful. However, the way universities describe the course syllabus and its related instructional approaches are not uniform among universities, it is not sufficient to simply search by apparent words in the syllabus texts, and it is necessary to capture latent relationships among words. In order to solve these problems, the authors propose Joint Topic Model with Selected Side Information (SS-JTM) to obtain relationships between the contents and instructional approaches of classes of several universities by selecting instructional approaches to be used as side information from syllabi. Functional extension of topic models has the possibility of performance degradation, but the results of evaluation experiments using Web syllabi from several universities have confirmed that SS-JTM performs as well as the baseline topic models such as LDA or JTM, and thus the functionality has been successfully extended.

**Keywords:** instructional approaches · Syllabus analysis · Course recommendation systems · Topic modeling · LDA

## 1 Introduction

Active learning is a generic term for instructional approaches that incorporate the active participation of learners, as opposed to one-way lecture-style education by teachers. Group discussions, debates, and group work in the classroom are also effective methods of active learning. Active learning is increasingly being introduced into university education [1, 11]. Also, in addition to the common

credit transfer systems, inter-university learning opportunity has been increased and varied with the spread of MOOCs (Massive Open Online Courses) such as edX<sup>1</sup>, Coursera<sup>2</sup> etc., which allow students to take classes at universities around the world via the Internet. There are many studies that are aware of the importance of active learning or MOOCs. Jahnke et al. [12] proposed a strategy to support active learning using digital technology. Furthermore, against the backdrop of the recent spread of COVID-19, Uetaka et al. [18] conducted the study on effective methods to achieve active learning in online classes with multiple students.

With such various instructional approaches and expansion of inter-university learning opportunities, students have to refer to many syllabi, which significantly increases their burden when taking classes. For example, some students prefer individual study while others prefer group study, therefore the preferred class is different for each student. In order to support syllabus browsing, a method of inter-university analysis of syllabi should be useful. However, the way universities describe the course syllabus and its related instructional approaches are not uniform among universities, it is not sufficient to simply search by apparent words in the syllabus texts, and it is necessary to capture latent relationships among words.

In this paper, the authors propose an approach to obtain the latent relationship between instructional approaches and learning contents from syllabi for inter-university course recommendation and syllabus analysis. Specifically, using text information obtained from syllabi, the proposed method classifies and relates each class based on learning contents and instructional approaches. This allows, for example, the proposed method to recommend classes that have a strong relationship with group learning to students who like group learning. Inspired by the study of course recommendation systems using topic models [2], our study focuses on the Joint Topic Model (JTM) [15], which is a kind of extended model of Latent Dirichlet Allocation (LDA) [5]. In the example shown in Fig. 1, words such as “Grammar” and “Reading” that represent the learning content are combined into word information, and words representing the instructional approach, such as “Group” and “Exercise”, are combined into side information.

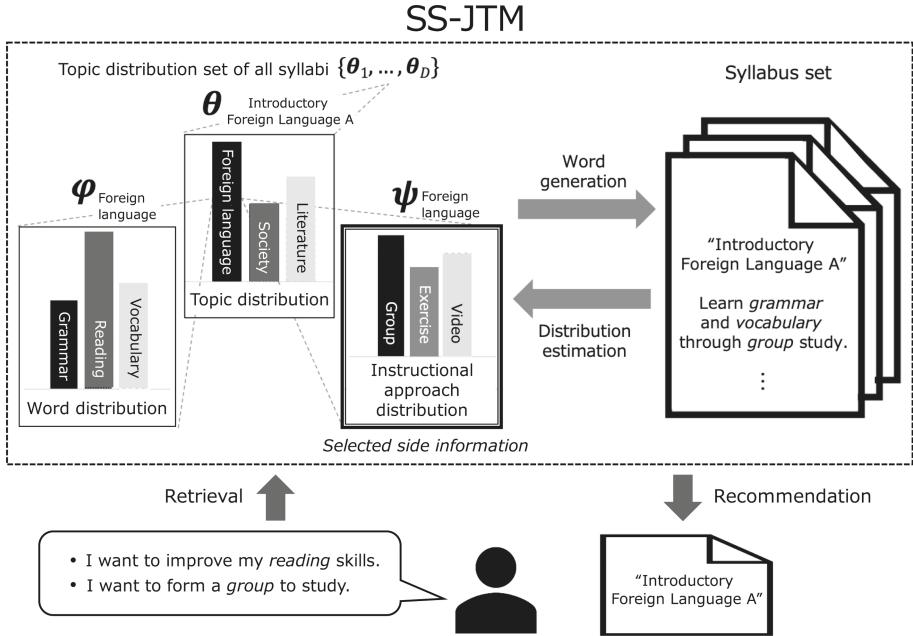
However, it is difficult to select appropriate side information because there is no unified vocabulary to describe various instructional approaches among universities. Side information is needed to describe the entire classes in syllabi, and the meaning of each piece of side information should be unique to increase its interpretability. In this study, the authors define selected side information as *a combination of side information that encompasses the entire syllabus and has little overlap*, and propose the SS-JTM (Joint Topic Model with Selected Side Information), which obtains and utilizes side information from syllabi. The contributions of this study are as follows.

- In the inter-university syllabus analysis that takes into account the various instructional approaches, it is an issue that the instructional approaches

---

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

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



**Fig. 1.** The framework of SS-JTM

are not unified among multiple universities. Therefore, simply merging the instructional approaches of multiple universities results in inappropriate duplication of the meaning of instructional approaches as side information. To address this issue, the authors proposed an algorithm for selecting the instructional approaches to be used as side information. SS-JTM can obtain the latent relationship between learning contents and instructional approaches among multiple universities.

- Although extensions to topic models may worsen performance [17], the results of evaluation experiments using Web syllabi from multiple universities demonstrate that SS-JTM successfully extends the functionality for selecting side information while maintaining the same performance as the baseline topic models.

## 2 Related Work

A solution for syllabus analysis by Matsuda et al. [14] is a supervised model of LDA, which analyzed syllabi of major universities. It became clear that there has been a regional bias in the computer science curriculum. Differences in the tendency to lean toward software or hardware, theory or application, and so on, were observed. The result would suggest a potential barrier to the development of a model curriculum among multiple universities. Another solution by Fréchet et al. [9] is a data-driven method such as variational inference or advanced use of

regular expressions to build a syllabus of political science field. These methods aimed to help new instructors build their own course syllabus.

A solution for course recommendation by Bakhshinategh et al. [4] is based on graduation attributes, which are criteria set by each university and represent the knowledge and skills that students should acquire before graduation. However, the course recommendation in the study by Bakhshinategh et al. [4] is dependent on the style of each university, and it is necessary to adjust the recommendation model accordingly. The purpose of this study differs from those of the inter-university course recommendation systems envisioned in the study by Bakhshinategh et al. [4]. Another solution by Apaza et al. [2] is a course recommendation system that applies LDA to syllabi of multiple universities. While course recommendations based on text analysis as in the study by Apaza et al. [2] can be applied to multiple universities, course recommendations based on grade prediction have challenging issues that it is difficult to recommend classes to first-year students with little course history, and personal information such as grades must be used.

As a solution to the information overload in online education, Jiang et al. [13] proposed the course recommendation using LDA. They emphasized the personalized recommendation focusing on the interests of each student, which was close to the motivation of our study. However, our study pays attention not only to students' interests but also to how to make use of the various instructional approaches and opportunities to take classes at multiple universities.

### 3 Preliminaries

#### 3.1 Topic Model

A topic is a category of meaning created from the co-occurrence of words (occurrence in the same document). This co-occurrence means not only the co-occurrence that appears in the document but also the latent co-occurrence that does not appear directly in the document. A document clustering method using topics is called topic models. Topic models in the broad sense include Latent Semantic Analysis (LSA) [7], which uses singular value decomposition of the word-document matrix. On the other hand, a topic model in the narrow sense is a probabilistic model for generating a set of documents expressed in bag-of-words and is represented by Latent Dirichlet Allocation (LDA) [5]. Topic models such as LDA can be considered superior to models such as LSA in the following two aspects [17]. First, LDA does not have negative values in document vectors like LSA, so it is easy for humans to interpret because a positive value can be intuitively interpreted as a probability value. Second, LDA enables one word to be assigned to multiple topics, contrary to LSA.

In the case of a syllabus, as shown in Fig. 1, each topic represents a grouping similar to an academic category, such as “Foreign language” or “Literature”. If a “Foreign language” topic consists of “Grammar” and “Reading”, then the appearance of “Grammar” in a syllabus can be considered as a latent occurrence of “Reading”, since “Reading” is generated from the “Foreign language” topic.

**Table 1.** Mathematical notation in Algorithms 1 and 2

Descriptions	Symbols
Number of documents	$D$
Number of topics	$K$
Document index	$d \in [D]$
Topic index	$k \in [K]$
Number of word kinds (Vocabulary)	$V$
Number of side information kinds	$U$
Length of the document $d$	$N_d$
Number of side information in the document $d$	$M_d$
Word set of the document $d$	$\mathbf{w}_d = \{w_{d1}, \dots, w_{dN_d}\}$
Side information set of the document $d$	$\mathbf{x}_d = \{x_{d1}, \dots, x_{dM_d}\}$
Dirichlet distribution hyperparameters	$\alpha, \beta, \gamma$
Dirichlet distribution	$\text{Dirichlet}(\cdot)$
Categorical distribution	$\text{Categorical}(\cdot)$
Word distribution of the topic $k$	$\phi_k = \{\phi_{k1}, \dots, \phi_{kV}\}$
Side information distribution of the topic $k$	$\psi_k = \{\psi_{k1}, \dots, \psi_{kU}\}$
Topic distribution of the document $d$	$\boldsymbol{\theta}_d = \{\theta_{d1}, \dots, \theta_{dK}\}$
$n$ -th word topic in the document $d$	$z_{dn}$
$m$ -th side information topic in the document $d$	$y_{dm}$

Table 1 shows an explanation of the symbols used in Algorithm 1 and Algorithm 2 in this chapter.

### 3.2 Latent Dirichlet Allocation (LDA)

LDA is a probabilistic model that generates a set of documents based on the assumption that each document consists of a unique topic distribution and each topic consists of a unique word distribution. The generation process is shown in Algorithm 1.

Here, the categorical distribution is a probability distribution of a variable that takes one value from several discrete values. The Dirichlet distribution is the probability distribution of the parameters of the categorical distribution and is expressed as in Eq. 1.

$$\text{Dirichlet}(\boldsymbol{\phi}|\boldsymbol{\beta}) = \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \phi_v^{\beta_v-1} \quad (1)$$

Categorical Distribution Parameters	: $\boldsymbol{\phi} = \{\phi_1, \dots, \phi_V\}$
Dirichlet Distribution Parameters	: $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_V\}$
Gamma Function	: $\Gamma(\cdot)$

---

**Algorithm 1** The generation process of LDA

---

```

1: for topics  $k = 1, \dots, K$  do
2:   Generating word distributions  $\phi_k \sim \text{Dirichlet}(\beta)$ 
3: end for
4: for documents  $d = 1, \dots, D$  do
5:   Generating topic distributions  $\theta_d \sim \text{Dirichlet}(\alpha)$ 
6:   for words  $n = 1, \dots, N_d$  do
7:     Generating the word's topic  $z_{dn} \sim \text{Categorical}(\theta_d)$ 
8:     Generating the word  $w_{dn} \sim \text{Categorical}(\phi_{z_{dn}})$ 
9:   end for
10: end for

```

---

In the second line of Algorithm 1,  $\phi_k \sim \text{Dirichlet}(\beta)$ , a word distribution  $\phi$  is generated for each topic according to the Dirichlet distribution with  $\beta$  as the hyperparameter. Similarly, in the fifth line, a topic distribution  $\theta$  is generated for each document according to the Dirichlet distribution with  $\phi$  as the parameter. In LDA, the word  $w$  is the observed variable (the variable that can be observed on the document), and  $\phi$  and  $\theta$  are latent variables (hidden variables that cannot be observed on the document). LDA updates the parameters of the model by iteratively generating  $w$  with  $\phi$  and  $\theta$ , and Bayesian estimation of  $\phi$  and  $\theta$  based on the posterior distribution.

### 3.3 Joint Topic Model (JTM)

JTM is an extension of LDA for a set of documents with side information. The generation process is shown in Algorithm 2. JTM iteratively generates side information and estimates its predictive distribution in the same way as word information, based on the assumption that each topic has its own distribution of side information in addition to the word distribution. The parameter estimation of JTM uses the collapsed Gibbs sampling method [10]. JTM is capable of obtaining the relationship between word information and side information for each topic. The study by Mimno et al. [15], for example, proposed a multilingual document analysis method that applied JTM to English documents using German documents as side information and extracted topics with German-English correspondences without using a dictionary. The method was based on a JTM with German documents as side information.

**Algorithm 2** The generation process of JTM

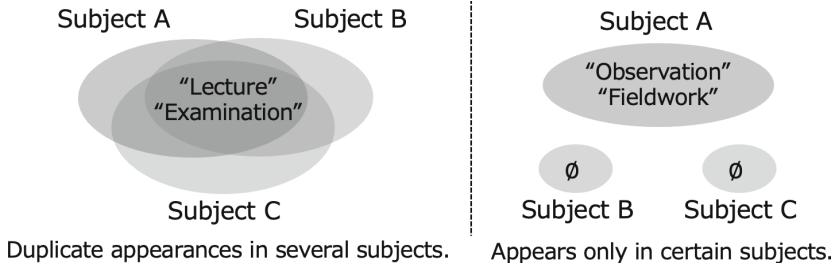
---

```

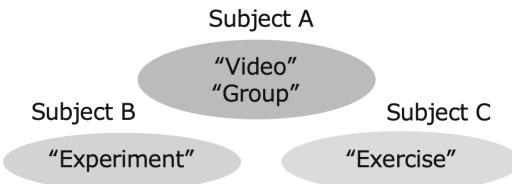
1: for topics  $k = 1, \dots, K$  do
2:   Generating word distributions  $\phi_k \sim \text{Dirichlet}(\beta)$ 
3:   Generating side information distributions  $\psi_k \sim \text{Dirichlet}(\gamma)$ 
4: end for
5: for documents  $d = 1, \dots, D$  do
6:   Generating topic distributions  $\theta_d \sim \text{Dirichlet}(\alpha)$ 
7:   for words  $n = 1, \dots, N_d$  do
8:     Generating the word's topic  $z_{dn} \sim \text{Categorical}(\theta_d)$ 
9:     Generating the word  $w_{dn} \sim \text{Categorical}(\phi_{z_{dn}})$ 
10:    end for
11:   for side information  $m = 1, \dots, M_d$  do
12:     Generating the side information's topic  $y_{dm} \sim \text{Categorical}(\theta_d)$ 
13:     Generating the side information  $x_{dm} \sim \text{Categorical}(\psi_{y_{dm}})$ 
14:   end for
15: end for

```

---



**Fig. 2.** Example of inappropriate side information ( $\emptyset$  indicates a subject that does not include any instructional approaches).



**Fig. 3.** Example of selected side information

## 4 Proposed Method

### 4.1 Joint Topic Model with Instructional Approaches as Side Information

In this study, the authors avoid the problem in the study [2] by focusing on instructional approach preferences instead of course history and grade informa-

tion, whose usage is generally restricted from the viewpoint of privacy protection. Specifically, the authors propose a system that obtains relationships between learning contents and instructional approaches from the distribution of words and side information in each topic obtained by JTM, and recommends courses that are strongly related to the preferred instructional approaches of the students. For example, the “foreign language” topic in Fig. 1 is strongly related to group learning. Even if the syllabus does not contain any words related to the instructional approach of classes, JTM has the advantage of being able to predict the instructional approach of classes based on the distribution of word information.

## 4.2 Selection of Side Information

If the targeted syllabus is limited to a specific university, it is relatively easy to select side information to be used as the instructional approach. On the other hand, when multiple universities are targeted as in this study, which aims at inter-university syllabus analysis, the classification of instructional approaches is not standardized, so it is necessary to select side information. Therefore, this study defines selected side information as *a combination of side information that encompasses the entire syllabus and has little overlap*. Inappropriate side information is a combination of instructional approaches that appear in most syllabi such as {Lecture, Examination} as shown in the left side of Fig. 2, or a combination of class forms that appear in only some classes such as {Observation, Fieldwork} as shown in the right side of Fig. 2, and therefore cannot represent the whole classes. On the other hand, the selected side information is a combination of instructional approaches that appear in all subjects and have little overlap with other subjects, as shown in Fig. 3.

Since no attempt has been made to refine the side information of JTM, the proposed SS-JTM method defines an algorithm to refine the side information as shown in Eq. 2.

$$\arg \max_{\mathbf{x} \in \mathcal{P}(\mathbf{X})} \sum_{i \neq j}^D \frac{|\{\mathbf{w}_i \cup \mathbf{w}_j\} \cap \mathbf{x}| - \lambda |\mathbf{w}_i \cap \mathbf{w}_j \cap \mathbf{x}|}{|\{\mathbf{w}_i \cup \mathbf{w}_j\} \cap \mathbf{x}| + 1} \quad (2)$$

Side information set	:	$\mathbf{X} = \{x_1, \dots, x_U\}$
Penalty hyperparameter	:	$\lambda$

First, a power set  $\mathcal{P}(\mathbf{X})$  is obtained for the set of side information  $\mathbf{X}$  before selection, and then the element of  $\mathcal{P}(\mathbf{X})$  that maximizes the objective function, i.e., the subset  $\mathbf{x}$  of  $\mathbf{X}$  that is most appropriate as the selected side information, is obtained. Specifically, the number of side information that appears in both document  $i$  and document  $j$ ,  $|\mathbf{w}_i \cap \mathbf{w}_j \cap \mathbf{x}|$  is subtracted from the number of side information that appears in document  $i$  or document  $j$ ,  $|\{\mathbf{w}_i \cup \mathbf{w}_j\} \cap \mathbf{x}|$ , as a penalty for duplication. In this experiment, the authors set  $\lambda = 2$  and give maximum penalty points to the combinations of side information that match in

document  $i$  and document  $j$ , and maximum penalty points to the combinations that are exclusive. This process is repeated for all document combinations.

## 5 Experiments

When extensions are made to topic models, as a trade-off for the achievement of the purpose of the extension, the performance of the model may deteriorate. In order to compare the performance of the proposed SS-JTM with the two baseline models, which are LDA and JTM, the authors have conducted an evaluation experiment on the 1420 syllabus documents of all the general education classes in the 2021 academic year at the University of Tokyo<sup>3</sup> and Mie University<sup>4</sup>. MeCab<sup>5</sup>, a Japanese morphological analysis system, is used for preprocessing the document data.

### 5.1 Parameter Settings

First, the authors have estimated the appropriate range of the number of topics in LDA based on four major indices. The upper panel of Fig. 4 shows the two indices that should be minimized (Griffiths2004 [10], CaoJuan2009 [6]), and the lower panel shows the two indices that should be maximized (Arun2010 [3], Deveaud2014 [8]). The estimation resulted in  $K = 11$  as the appropriate number of topics for LDA. However, since this value may not be universally appropriate for all models, evaluations have been conducted in the range of 3 to 18 topics in the experiments described later. In addition, the hyperparameters of the Dirichlet distribution ( $\alpha, \beta, \gamma$ ) are set to 0.1, and the maximum number of training iterations is set to 200.

### 5.2 Evaluation Metrics

Coherence [16] is a measure of the quality of each topic, and is defined as follows: the higher the value, the stronger the semantic coherence of the topics. The definition of Coherence is shown in Eq. 3.

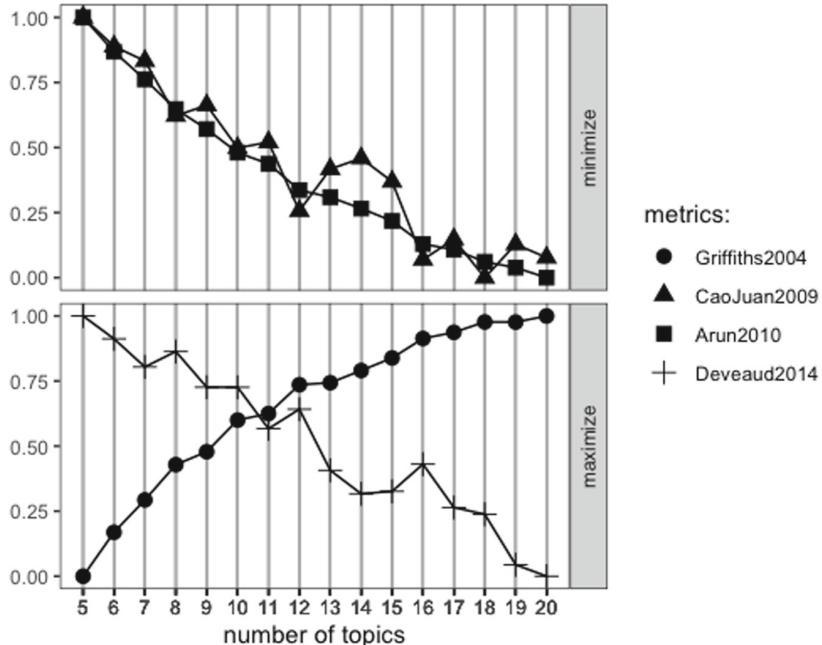
$$\text{Coherence}_k = \sum_{n \neq m}^N \log \frac{D(v_n^{(k)}, v_m^{(k)}) + 1}{D(v_m^{(k)})} \quad (3)$$

Top $N$ words that occur frequently	$: \{v_1^{(k)}, \dots, v_N^{(k)}\}$
Number of documents where $v_n^{(k)}$ appears	$: D(v_n^{(k)})$
Number of documents where $v_n^{(k)}$ and $v_m^{(k)}$ appear	$: D(v_n^{(k)}, v_m^{(k)})$

<sup>3</sup> <https://catalog.he.u-tokyo.ac.jp>.

<sup>4</sup> <https://syllabus.mie-u.ac.jp>.

<sup>5</sup> <http://taku910.github.io/mecab/>.



**Fig. 4.** Results of estimating the appropriate number of topics

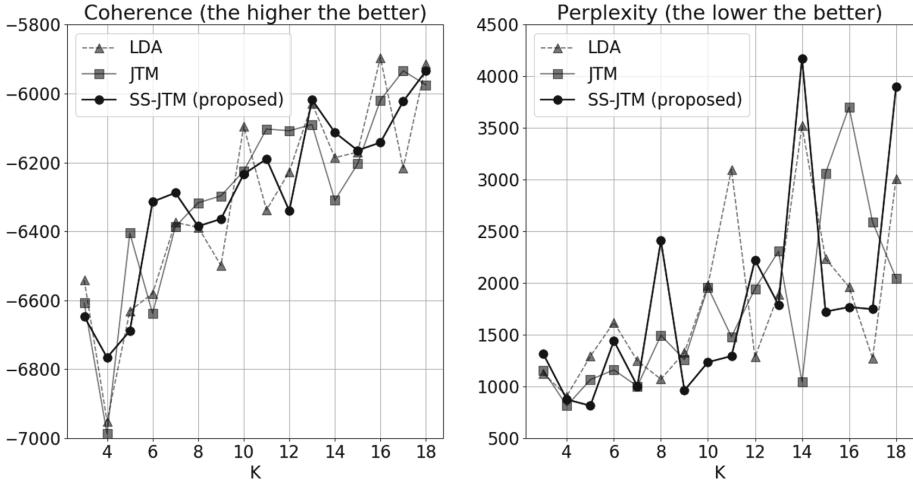
Specifically, coherence is higher when the words forming a topic actually co-occur well with each other. In general, the higher the number of topics, the higher the coherence. Performance comparison among different models is based on average coherence per topic, as shown in Sect. 6.2. In the experiment,  $N$  is set as 50.

Perplexity [5] indicates that the smaller the value, the higher the predictive performance of general stochastic models not limited to topic models. The definition of perplexity is shown in Eq. 4.

$$\text{Perplexity}(\mathbf{W}^{\text{test}} | M) = \left( \prod_{d=1}^D \frac{1}{\log p(\mathbf{w}_d^{\text{test}} | M)} \right)^{\frac{1}{\sum_{d=1}^D N_d}} \quad (4)$$

Test document set	: $\mathbf{W}^{\text{test}}$
Probability model (topic model)	: $M$
Likelihood of test document $d$ in $M$	: $p(\mathbf{w}_d^{\text{test}}   M)$

Intuitively, perplexity is a number representing the average number of branches per word. 13% of the total number of words in the entire syllabus is used for tests in the calculation of perplexity.



**Fig. 5.** The results of the experiments

## 6 Results

### 6.1 Selection of Side Information

The side information before the selection was determined by referring to the general classification of instructional approaches. The results of the selection are shown below.

#### Side information before selection

Lecture, Exercise, Experiment, Practical skill, Practical training, Group, Video, Fieldwork, Seminar, Observation, Examination, Omnibus

#### Selected side information

Lecture, Seminar, Practical training, Group, Video, Omnibus

After the selection, “Practical skill” and “Experiments”, which have a large overlap in meaning with “Exercises” and “Practical training”, and “Fieldwork” and “Observation”, which have a small distribution overall, are eliminated. As a result of the side information refinement, the selected side information is obtained that can explain many subjects with little overlap in meaning as a instructional approach.

### 6.2 Performance Evaluation

The results of the experiment are shown in Fig. 5. Coherence shows that SS-JTM has the worst value in the case where  $K = 4$ , which is better than the baseline for the same number of topics. As for perplexity, the SS-JTM is noticeably inferior to the baseline in the case where  $K = 14$  and  $18$ , while it exceeds the baseline models in the case where  $K = 11$ , as mentioned in Sect. 5.1. One of the reasons

for the unstable perplexity in the range of  $K \geq 7$  is that the overlap of side information among topics increases due to the six types of side information in the SS-JTM. In some cases, Perplexity is very similar even though K is different, but the reason for this is not well understood at this time. It is a topic for future research.

The above results demonstrate that the proposed SS-JTM method performs as well as the baseline models while extending the function of side information selection.

## 7 Conclusion

In this study, the authors proposed SS-JTM for inter-university syllabus analysis focusing on the various instructional approaches. SS-JTM obtains relationships between learning contents and instructional approaches by selecting side information from combinations of instructional approaches in syllabi of multiple universities. SS-JTM will help to achieve more personalized learning in course recommendations or curriculum developments. Although extensions to the topic model may worsen performance, SS-JTM succeeded in maintaining the same level of performance as the baseline models. In addition, SS-JTM may be applicable not only to syllabi, but also to general data for which side information is not selected.

One of the future tasks is to implement a course recommendation and visualization system based on the proposed method, and to quantitatively and qualitatively evaluate how many useful latent relationships can be discovered. For this purpose, it is necessary to investigate the quantitative definition of latent relationship since it is not self-evident.

**Acknowledgement.** The Web syllabus used in the experiment was obtained from the University of Tokyo and Mie University. This research work was supported by JSPS KAKENHI Grant Number JP22K12312.

## References

1. Abraham, Y.S.: Importance of active learning in an undergraduate course in construction scheduling. In: Proceedings of ASEE (2020)
2. Apaza, R.G., Cervantes, E.V., Quispe, L.C., Luna, J.O.: Online courses recommendation based on LDA. In: Proceedings of the 1st Symposium on Information Management and Big Data (SIMBig 2014), vol. 1318, pp. 42–48 (2014)
3. Arun, R., Suresh, V., Madhavan, C.E.V., Murty, M.N.: On finding the natural number of topics with latent dirichlet allocation: some observations. In: Proceedings of PAKDD, pp. 391–402 (2010)
4. Bakshinategh, B., Spanakis, G., Zaïane, O.R., ElAtia, S.: A course recommender system based on graduating attributes. In: Proceedings of CSEDU, pp. 347–354 (2017)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)

6. Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S.: A density-based method for adaptive LDA model selection. *Neurocomputing* **72**(7–9), 1775–1781 (2009)
7. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**(6), 391–407 (1990)
8. Deveaud, R., SanJuan-Ibekwe, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* **17**(1), 61–84 (2014)
9. Fréchet, N., Savoie, J., Dufresne, Y.: Analysis of text-analysis syllabi: building a text-analysis syllabus using scaling. *PS Polit. Sci. Polit.* **53**(2), 338–343 (2020)
10. Griffiths, T., Steyvers, M.: Finding scientific topics. *PNAS* **101**(Suppl. 1), 5228–5235 (2004)
11. Hartikainen, S., Rintala, H., Pylväs, L., Nokelainen, P.: The concept of active learning and the measurement of learning outcomes: a review of research in engineering higher education. *Educ. Sci.* **9**(4), 276 (2019)
12. Jahnke, I., Meinke-Kroll, M., Todd, M., Nolte, A.: Exploring artifact-generated learning with digital technologies: advancing active learning with co-design in higher education across disciplines. *Technol. Knowl. Learn.* **27**(1), 335–364 (2022)
13. Jiang, X., Bai, L., Yan, X., Wang, Y.: LDA-based online intelligent courses recommendation system. *Evol. Intell.* 1–7 (2022)
14. Matsuda, Y., Sekiya, T., Yamaguchi, K.: Curriculum analysis of computer science departments by simplified, supervised LDA. *J. Inf. Process.* **26**, 497–508 (2018)
15. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of EMNLP, pp. 880–889 (2009)
16. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of EMNLP, pp. 262–272 (2011)
17. Sato, I., Okumura, M.: Statistical latent semantic analysis based on topic model, 272 p. Corona Publishing Co. Ltd. (2015). (in Japanese)
18. Uetake, T., Majima, T., Aoki, A., Baba, S.: A trial of active learning method for business management education in online environment. In: Proceedings of HCII, vol. 1421, pp. 132–137 (2021)



# Plug & Play with Deep Neural Networks: Classifying Posts that Need Urgent Intervention in MOOCs

Laila Alrajhi<sup>1,2(✉)</sup> and Alexandra I. Cristea<sup>1</sup>

<sup>1</sup> Computer Science, Durham University, Durham, UK

{laila.m.alrajhi, alexandra.i.cristea}@durham.ac.uk

<sup>2</sup> Educational Technology, King Abdulaziz University, Jeddah, Saudi Arabia

**Abstract.** Urgent intervention in learner forum posts have recently occupied a very important role in research in Massive Open Online Course (MOOC) environments. Intervening in time may make the difference between a learner dropping out or staying on a course. However, due to the typical extremely high learner-to-instructor ratio in MOOCs, it is very challenging – if not sometimes impossible – for the instructor to monitor all the existing posts and identify which need immediate intervention, to encourage retention. Current approaches are based on shallow machine learning and deep learning. Whilst deep learning methods have been shown to be most accurate in many domains, the exact architecture can be very domain-dependent. In spite of their sheer size and representation power, deep neural networks are known to perform better when a problem is divided into the right sub-problems. These sub-problems can be further assembled together, to answer to the original problem, in what we intuitively call a ‘plug & play’-like fashion, similarly to puzzles – via *hybrid (deep) neural networks*. Hence, in this paper, we address this problem by proposing a classification model for identifying when a given post needs intervention from an instructor, based on hybrid neural networks. We represent words using two different methods; *word2vec*: that capture the word’s semantic and syntactic characteristics; and *transformer model* (BERT): which represents each word according to its context. Then we construct different architectures, integrating various deep neural networks (*DNNs*) -‘*word-based*’ or ‘*word-character based*’, as we expected that adding additional character-sequence information may increase performance. For word-based, we apply convolutional neural network (*CNN*) and/or different types of recurrent neural networks (*RNN*); in some scenarios we added *attention*. This is to present a comprehensive answer to the character-sequence question in particular, as well as to the *urgency of intervention need prediction in MOOC forums*, in general. Experimental results demonstrate that using BERT rather than word2vec as a word embedding enhances performance in different models (the optimal result is the *CNN + LSTM + Attention* model based on BERT at word-level). Interestingly, adding word-character input does not improve the performance, as it does for word2vec.

**Keywords:** Instructor Intervention · MOOC · Text Classification · Hybrid Model · Deep Learning · BERT · word2vec

## 1 Introduction

Massive open online courses (MOOCs) are a subset of information systems called open online learning environments with largescale enrolment [1]; they offer high-quality education and different courses, in different domains, and they are provided by different universities, institutions and ventures [2]. Most of these courses are offered at no cost (free) or extremely cheaply [3], and some have no prerequisite requirements; as a result, these courses attract hundreds of thousands of learners [4] to improve their life-long learning and increase their knowledge at their convenience [5]. These learners live all over the world and have diverse knowledge backgrounds, abilities, motivations and even education systems.

The completion rates for courses under the MOOCs model are extremely low, at only around 10% [6]. There are several ongoing debates about the reasons for these low completion rates. One of the most critical factors that have been identified is missing direct interaction in terms of face-to-face communication, support and collaboration, which leads online learners to feel isolated and suffer from a lack of meaningful human interaction compared with other educational environments. Because of this, some learners are stuck, confused, need clarification and may struggle to stay in the course; if these issues are not addressed, it can ultimately lead to dropout [5, 7].

From the instructor's perspective, intervention to address learners' questions in online learning is a central and essential teaching activity [8]. While instructors have limited time, they try to assist, encourage and motivate learners and respond to their questions as much as possible. However, due to the tremendous number of learners enrolled and extremely high ratio of learners-to-instructors, the instructors need to be selective in their interventions [2]. In addition, the massive amounts of posts present in MOOCs, most of which are just general discussion and social connections that do not involve any urgent issues or require intervention, mean that it is extremely difficult and time-consuming for instructors to effectively monitor and review all the existing posts, which may number in the millions, and find cases where it is necessary to engage in meaningful interactions, resolve issues and provide feedback.

At the same time, deep learning models have been developed for text classification [9], with a great amount of success. However, these models cannot deal directly with the words, working instead on word embeddings, to produce word vectors [10]. Currently, using Bidirectional Encoder Representations from Transformers (BERT) [11] as a word embedding has become popular, because it produces word embeddings based on their context - unlike existing word embedding models, which embed each word in a single vector, without taking into account the different contexts of use [12]. In addition, [13] proved that using BERT to represent words improved the performance for detecting urgent posts.

However, whilst deep neural networks (DNNs) have been successful, the exact model that is appropriate has been found to be often domain-specific. For instance, [14] found that the best model for text classification depends on the extent to which it is important to understand the semantic meaning of whole sentences.

The umbrella research question (RQ0) of this paper is *What is the preferable combination between different deep learning models, to construct the best predictor model for the need of instructor intervention?* We answer by applying the ‘plug & play’

technique for word-based and word-character based input, based on different embeddings (word2vec [15] and BERT). The sub-question (RQ1) is: *Do word-character based approaches outperform word-based approaches for the post urgency problem and is this different on BERT as word embedding, compared to more traditional models?*, such as word2vec.

The key contributions of this paper are the following: (i) *analysing and exploring for the first time to MOOC posts data in terms of length (number of words and characters per posts)*, (ii) *constructing different simple and hybrid deep learning models, by applying ‘plug & play’ techniques, to establish good combinations in terms of performance*, (iii) *applying an attention mechanism that considers word-based input only, by using a separate attention score for every word, according to their importance* (iv) *for the first time, showing the quality of BERT and its sufficiency when using word-based only, without adding word-character based input*.

## 2 Related Works

### 2.1 Instructor Intervention Prediction Models

Chaturvedi et al. [2] were among the first researchers to present the problem of prediction for instructor intervention in the MOOC forums environments, framing it as a binary prediction task; they labelled data automatically, as follows: positive, if instructors had posted replies; negative otherwise. They proposed three models (Logistic Regression, Linear Chain Markov Model and Global Chain Model) to determine whether or not an instructor would intervene to threads or posts. They concluded that it is important to use a thread structure in predicting instructor intervention behaviour [2]. Unlike us, they did not train straight on the text of the posts.

A similar study Chandrasekaran et al. [16] added prior knowledge about the type of forum (thread-level) as a feature, in addition to linguistic content features, to predict when intervention in forum posts was needed as a binary classifier.

It should be noted that the method used in the two previous studies for labelling data is inaccurate, as there may be some posts that required intervention from the instructor but he/she did not intervene (due to missing them, having no time, etc.); similarly, there could be interventions where it was not needed. On the other hand, the following studies, similarly to our current study, used the Stanford MOOCPosts dataset, which was manually labelled by three consultants and published in [17], as we explain in Sect. 3.1.

In terms of the transfer learning model and cross-domain MOOC forum post classification, Bakharia [18] conducted preliminary research on cross-domain classification, by training different classifiers to classify posts onto three different categories, specifically: confusion, urgency and sentiment. She then validated them across different unseen domain areas, using the Stanford MOOCPosts dataset. She found low cross-domain classification accuracy, but she mentioned that transfer learning should be given more consideration. In a follow-up study, Wei et al. [19] proposed a model for classification and transfer learning for cross-domain MOOC forum posts, based on deep neural networks, as a combination of a (*CNN + LSTM*) called ConvL, to identify three tasks: confusion, urgency and sentiment in posts. They used the same Stanford MOOCPosts dataset.

Almatrafi et al. [20] also used the Stanford MOOCPosts dataset to build a generalised model to identify reliably urgent posts, regardless of the content of the course, by implementing different linguistic features and metadata as features, to train different shallow ML models. Their results were satisfactory. The latter however did not use deep learning. In addition, Sun et al. [21] distinguished potentially important urgent posts by presenting an improved recurrent convolutional neural network (*RCNN*) to obtain the contextual information. Another study by Alrajhi et al. [22] focused on predicting urgent posts and created a multidimensional deep learner intervention model that combines different dimensions with text. All these research focus on word level as input to classification model, in contrast our research add word-character level.

Guo et al. [23] used deep learning models combining (*CNN + GRU*) that extract semantic information and structural information to detect posts that needed urgent responses, by applying attention, to develop a hybrid character/word neural network with the same dataset. Khodeir [13] utilised instead a multi-layer (*Bi-GRU*) based on BERT as an embedding layer to classify learners urgent posts. She used BERT as a word embedding to represent words in their context.

The idea of our research was inspired by the last two researches; however, we applied the idea of *plug & play in deep learning* based on *word2vec* or *BERT*, for different word-based or word-character based inputs, towards a more *comprehensive view on constructing models to predict the urgency of intervention need in MOOC forums*.

## 2.2 Towards Plug & Play: Combinations in Deep Learning

There are two fundamental deep learning architectures, namely convolutional NNs (*CNNs*) [24] and recurrent NNs (*RNNs*) [25]. Popular deep RNNs are the long short-term memory (*LSTM*) [26], gated recurrent units (*GRU*) [27] and bidirectional RNN (*Bi-RNN*) [28]. For text classification, a CNN is known to be better at extracting local and position-invariant features, while a RNN is effective at modelling units in sequence. This is due to their different architectures, since CNNs are hierarchical, while RNNs are sequential [14].

Whilst deep learning has been proven to be performant, recently, a great amount of research has focused on combining two or more types of DNNs, to produce a more effective combination model (ensemble). This is specifically prominent in the computer vision field [29–31]. Recently, many researchers applied the combination of different DNN models on text analysis and classification, as well. For example, Wang et al. [32] introduced a technique to combine CNN and RNN models for sentiment analysis of short texts and their results showed that it leads to improvements in accuracy. Lai et al. [33] proposed RCNN by applying RNN to capture contextual information, followed by a CNN to obtain the final representation for sentence classification.

Another study, Zhang et al. [34], combined CNN and GRU to detect hate speech on Twitter. Their model is outperforming on 6 out of 7 datasets between 1 and 13% in terms of F1-score. Also, as mentioned in Sect. 2.1, Wei et al. [19] proposed a framework for transfer learning based on CNN and LSTM and showed the effectiveness of their model on the Stanford MOOCPosts dataset.

All previous studies have merged different types of DNNs for the word-level only. In addition to combining different layers at word-level, there are other researches that

combine characters with words as input. For example, Liang et al. [35] used word-level and character-level representation as input to classify informal text. Their results are competitive with other studies in the SemEval-2010 Task8 and outperform on the KBP-SF48 dataset, by better learning character features. Also, Yenigalla et al. [36] proposed a method to integrate both character- and word-based models for text classification, to address the problem of the unseen word in word-based models. Their results showed that the accuracy was improved.

Guo et al. [23] proposed an attention-based model that concatenates word-level and character-level representation, to extract the semantic and structure information. They clarified that the MOOC posts contain a lot of noise and they can overcome this problem by adding character-level to capture this special information.

As we mentioned earlier, we built our study based on Guo et al.'s research. However, they used semantic and structural information to classify posts that need intervention. They learned semantic information by applying CNN, while they learned structure by using the last hidden state of the GRU. Then, they used an attention mechanism to learn the weights of the word-character representations. In contrast, in our model we use CNN to extract local features and investigate different types of RNN (plug & play) to model units in sequence, by returning all the hidden states to the attention, to allocate weights for every word. We apply this attention mechanism first to word-based input, then we added character-based input, to improve noisy data, such as misspellings. In addition, we represent words using two methods: (1) BERT, as in Khodeir [13] in contrast to (2) Guo et al., [23] as they used word2vec (google-news Vectors).

## 3 Methodology

### 3.1 Dataset

As mentioned above, we use the Stanford MOOCPosts dataset [17] to develop a model and predict the urgency for instructor intervention in MOOC forums. Many researchers have worked on this dataset [18, 20, 23]. The dataset consists of 29,604 posts in total, collected from 11 Stanford University online courses from 3 domain areas (Humanities/Sciences, Medicine and Education). Humanities/Sciences contains 6 courses, Medicine contains 4 courses and Education contains one course. Each post on these courses is categorised along six dimensions, namely, 'confusion', 'sentiment', 'urgency', 'question', 'answer' and 'opinion'. Three human coders have created a gold dataset, by categorising the data firstly along three binary dimensions (opinion, question and answer) and the rest (sentiment, confusion and urgency) onto 7- point scales (1–7). We use binary classification, as proposed by [23], by converting the urgency dimension into a binary value (with posts with score > 4 as 1 and 0 for the remainder). Thus, the forum posts are classified into binary categories of instructor interventions, according to the expressed urgency:

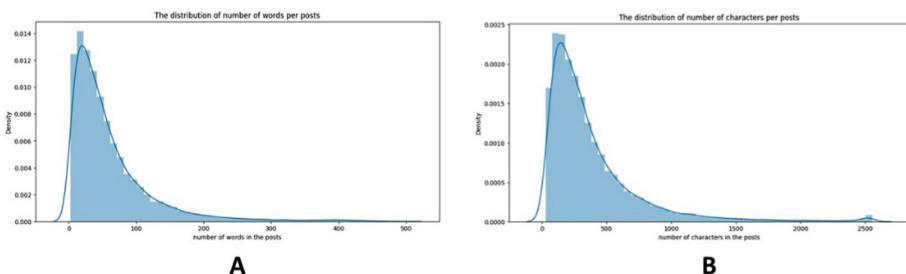
- *Urgent intervention*: requires immediate intervention;
- *No intervention*: no response is required.

We further applied pre-processing, including data cleaning and removing the integer numbers [20], special characters. Then we removed all previous automated anonymisation tags (e.g., <ZipRedac>, <nameRedac>, <redacted> ... etc.) [19]. Next, we converted the text to lowercase. The stopwords were kept because, as [37] noted, models can improve if they include stopwords. After that we prepared the final input, by adding the name and the domain of the course to the text input. We followed [23] who argued that to understand the information contained in the post, one should connect the course and domain information of the post to the text of the post. Thus, we removed 13 posts with an empty course name; this produced 29,584 posts.

Then, we split the data into *training*, *validation* and *testing* as follows: training and testing into (80%, 20%) by using stratified sampling to select a sample that is representative of different classes (urgent intervention and no intervention). After that, we split the training data into training and validation sets (80%, 20%, respectively).

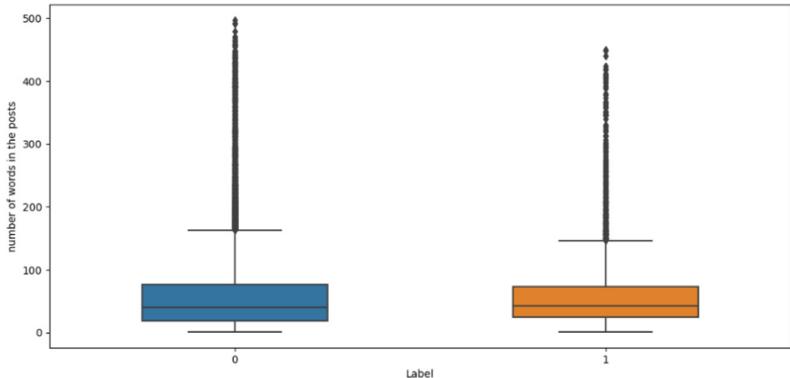
### 3.2 Exploring the Dataset

As an essential step, we explore and analyse the data set, to understand the data. We focus here on number of words and characters in all posts, because in deep learning models we should specify the length of the input sequence to these models. As we mentioned in Sect. 3.1, the input of these models is the text in the post and the name and the domain of the course, therefore we have to consider this information in the following calculations. As shown in Fig. 1, (A) is the distribution of the number of words per posts (mean = 60.36 words, minimum = 2 words and maximum = 498 words). (B) is the distribution of the number of characters per posts (mean = 380.28 characters, minimum = 29 characters and maximum = 2556 characters).



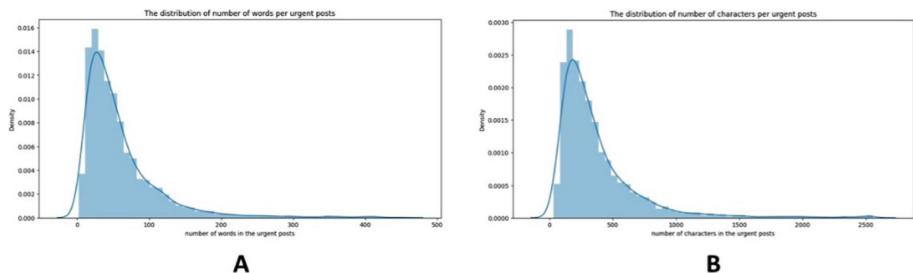
**Fig. 1.** Distributions of posts (A) Number of words per posts - (B) Number of characters per posts.

Then, to understand the behaviour of learners and how many words they write when they need urgent intervention (Label = 1) or not (Label = 0), we visualise the representation of the number of words per label, as we see in Fig. 2. To check if there are a statistical significant differences between the two populations, we use Mann-Whitney test. We found that  $p < 0.05$ , meaning they are statistically significantly different in terms of the length of posts (number of words).



**Fig. 2.** Box plot for the number of words per posts that were written by learners needing intervention (Label = 1) or not needing intervention (Label = 0).

Following, we analyse specifically the posts written by learners who need urgent intervention. Depicted in Fig. 3 (A) is the distribution of the number of words per urgent posts (mean = 59.63 words, minimum = 2 words and maximum = 450 words). On the right side (B) is the distribution of the number of characters per urgent posts (mean = 372.12 characters, minimum = 32 characters and maximum = 2556 characters).

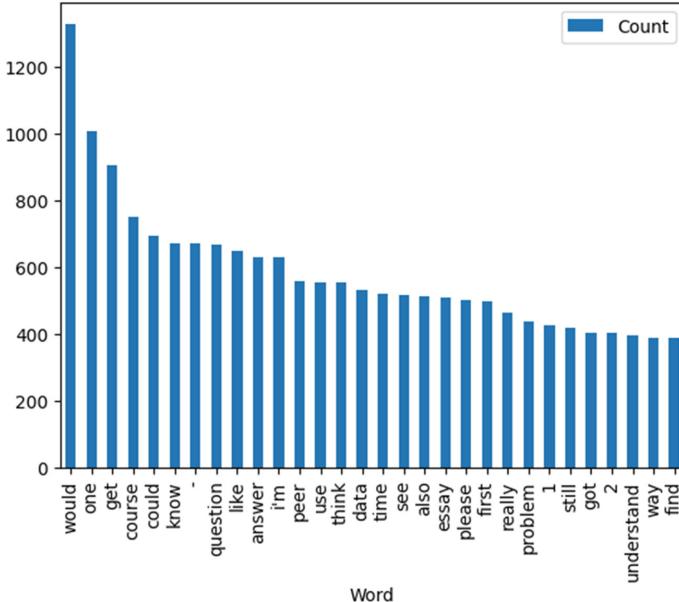


**Fig. 3.** Distributions of urgent posts (A) Number of words per urgent posts - (B) Number of characters per urgent posts.

Finally, to find out which words were most frequently used by learners in urgent posts, we calculate the top 30 frequency words in these posts, after removing stopwords; see Fig. 4.

### 3.3 Prediction Models

The input of these models is the post from each learner and the output is the classification: if this post needs intervention or not, according to its urgency (binary prediction). We represent words as numerical forms (word embedding) using *word2vec*, as in [23], which converts words into vectors that depict semantics; and *BERT*, as in [13], which generates contextual representations for each words. We implement two different training models, which are as follows: (1) with *word-based input*. (2) with character-based



**Fig. 4.** The top 30 frequency words in urgent post

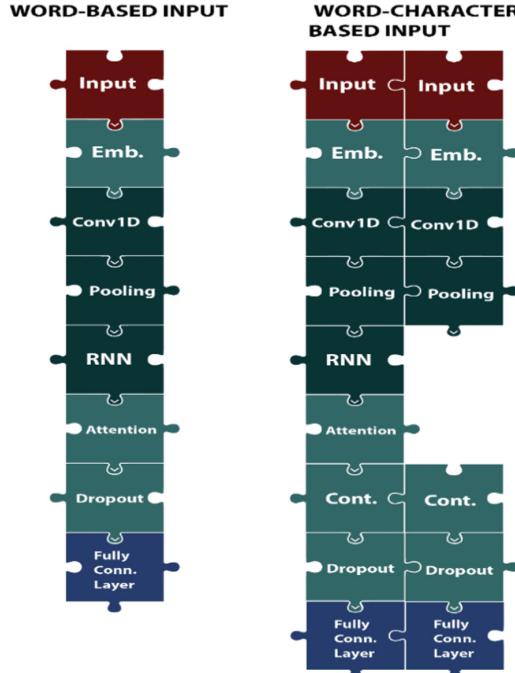
input in addition to the word-based, to configure what we call *word-character based input*. Figure 5 shows the general architectures for these two cases (word-based input and word-character based input). We select (Word2vec or BERT) as a word embedding, CNN to extract local complex context features, RNN to model units in sequence and learn feature structures, and attention mechanism to give higher weight to keywords. In word-character based models, in addition to the DNN layers for words, we applied additionally CNN, to select the features for characters.

After we construct every model (word based or word-character based) based on different word embeddings, we train it by using the Adam optimiser, as in [23]. We use the binary crossentropy as a loss function, because our problems involve binary decisions. We set the batch size to 64 and we employ early stopping, to stop training after 5 epochs, when there is no progress in performance.

### 3.3.1 The Word-Based Input

We construct different basic and combined deep learning models, to selected features, by applying plug & play, i.e., CNN, RNN (LSTM, Bidirectional-LSTM, GRU and Bidirectional-GRU) as they are a modification of RNN. These models are based on word2vec or BERT. In the following, we explain a general word-based architecture; however, during the implementation, we remove some layers and add other layers, such as different types of RNN and attention layers.

In word2vec we apply (Word2vec GoogleNews-vectors-negative300) as in [23], as they showed that this renders better results than GloVe on the Stanford dataset. In



**Fig. 5.** Deep Learning as a Puzzle: General architectures for two cases (word-based input and word-character based input).

addition to pre-trained embedding, we train word embeddings during the NN training (fine-tuning).

In BERT, we use the BERT tokeniser to tokenise sentences into tokens using ‘bert-base-uncased’. In BERT, an original word is split into tokens. These tokens represent words and the words that are not part of the vocabulary are represented as sub-words and characters. Also, we add special tokens [CLS] and [SEP]; [CLS] – is inserted at the start of the text and [SEP] – is inserted at the end; or if there are more sentences, it is used to separate these sentences [38].

We constrain the maximum length of each post, by following [13], to 512 tokens, since 512 tokens is the maximum model input size for BERT. Thus, we pad sequences of less than 512 with zeroes and trim the rest of the tokens which are more than 512, to ensure that each sequence has the same length.

As Fig. 5 shows, the first layer is an input layer, followed by the embedding layer, mapping words onto vectors. The output from these layers is passed onto the 1D convolution layer as input (with *128 filters*, kernel size of  $\{3,4,5\}$  and Rectified Linear Unit ‘ReLU’ as activation function) as in [23] to derive interesting features. Then, the produced features are further compressed, by using a pooling layer (*max pooling*). These features feed into one of different RNN (LSTM, Bidirectional-LSTM, GRU and Bidirectional-GRU) layer with 128 hidden units, which helps to find the relationship between words.

The next layer is the *attention* layer, which learns the weighting for each word. We use attention with context mechanism [39]. That is, through a series of mathematical formulas, a context vector is randomly initialised and multiplied by each word, to generate the importance score.

Next, for the following drop-out layer, we use drop-out with a rate of 0.5 as recommended in [23], by randomly dropping out nodes during the training phase, as well as an early stopping mechanism, to alleviate overfitting. Then, the fully connected layer is used to classify the output as: 1- needs urgent intervention or 0 – no intervention - by calculating the probability ( $P$ ):

$$P = \begin{cases} 1, & \text{if } > .5 \\ 0, & \text{if } \leq .5 \end{cases} \quad (1)$$

### 3.3.2 The Word-Character Based Input

In the second case, we add characters in addition to the words. We selected 1024 as the length of each input post as recommended in [23], and due to the fact that most posts have a number of characters less than 1024. Next, we encode characters and use character embedding as per the Johnb30/py\_crepe code on GitHub, which is a reimplemented version of the Crepe character-level convolutional neural net model that was originally shown in [40].

To select the most important features, we use a standard 1D convolutional and pooling layer, respectively. The convolution with filter = 128 and kernel size of {5,7,9} and the activation function is ‘ReLU’, following recommendations in [23], and pooling is global max pooling. The selected features are concatenated with the features from the word-based input, and the rest follows the processing explained in the word-based input section.

## 4 Results and Discussion

In this section we present the results we obtained after training every model. We are reporting precision ( $P$ ), recall ( $R$ ) and F1-score ( $F1$ ) for each class and the balance accuracy ( $BA$ ) for the word-based approach in our various deep learning models (first row), and then report the combined (word and character-based results) (second row), as shown in Table 1. In this research we compare models based on the ( $BA$ ) score, because it is a widely used metric for binary classification of imbalanced datasets [41].

From these results, we find that the  $BA$  score for models with BERT as a word embedding outperformed all the models based on word2vec. That means it is better to represent words using BERT. The best value from all the models is **0.875** and the answer to RQ0 is *CNN + LSTM + Attention* model based on BERT at word-level.

In terms RQ1, the (word-based versus word-character based), we observe that if word2vec is used as a word embedding, word-character based often outperforms base ones (word-based) (these improvements are statistically significant using (McNemar’s test:  $p < 0.05$ ) as shown in Table 1). In contrast, for models with BERT as a word embedding, there is no improvement between different bases (word only and word-character).

**Table 1.** The performance results of word2vec and BERT as word embedding for the word-based and word-character based approach for different models (P,R,F1, BA %) and P value (P.V) comparing word-based versus word-character based for each model.

Model	Word Embedding	Level	No Intervention (0)			Intervention (1)			BA	P.V
			P	R	F1	P	R	F1		
CNN	Word2vec	Word	0.89	0.96	0.92	0.74	0.51	0.60	0.732	<b>P ≤ 0.05</b>
		Word + Char	0.92	0.95	0.94	0.77	0.66	0.71	0.807	
	BERT	Word	0.93	0.96	0.95	0.82	0.69	0.75	0.826	P > 0.05
		Word + Char	0.92	0.97	0.95	0.84	0.66	0.74	0.815	
CNN + GRU	Word2vec	Word	0.90	0.95	0.92	0.72	0.55	0.62	0.749	<b>P ≤ 0.05</b>
		Word + Char	0.93	0.96	0.95	0.82	0.68	0.74	0.823	
	BERT	Word	0.95	0.94	0.95	0.76	0.79	0.77	0.865	<b>P ≤ 0.05</b>
		Word + Char	0.95	0.96	0.95	0.80	0.76	0.78	0.860	
CNN + Bidirectional (GRU)	Word2vec	Word	0.92	0.93	0.93	0.68	0.67	0.68	0.798	<b>P ≤ 0.05</b>
		Word + Char	0.92	0.96	0.94	0.80	0.65	0.71	0.803	
	BERT	Word	0.94	0.96	0.95	0.82	0.74	0.78	0.851	<b>P ≤ 0.05</b>
		Word + Char	0.94	0.97	0.95	0.84	0.75	0.79	0.856	
CNN + GRU + Attention	Word2vec	Word	0.92	0.93	0.93	0.69	0.67	0.68	0.800	<b>P ≤ 0.05</b>
		Word + Char	0.93	0.96	0.95	0.81	0.70	0.75	0.829	
	BERT	Word	0.95	0.96	0.95	0.80	0.76	0.78	0.859	P > 0.05

(continued)

**Table 1.** (*continued*)

Model	Word Embedding	Level	No Intervention (0)			Intervention (1)			BA	P.V
			P	R	F1	P	R	F1		
		Word + Char	0.94	0.97	0.95	0.83	0.74	0.79	0.854	
CNN + Bidirectional (GRU) + Attention	Word2vec	Word	0.90	0.95	0.93	0.74	0.56	0.64	0.755	<b>P ≤ 0.05</b>
		Word + Char	0.93	0.96	0.94	0.80	0.67	0.73	0.816	
	BERT	Word	0.95	0.94	0.95	0.77	0.80	0.78	0.872	<b>P ≤ 0.05</b>
		Word + Char	0.95	0.96	0.95	0.81	0.78	0.80	0.868	
CNN + LSTM	Word2vec	Word	0.81	1.00	0.90	0.00	0.00	0.00	0.5	<b>P ≤ 0.05</b>
		Word + Char	0.92	0.97	0.95	0.82	0.66	0.73	0.814	
	BERT	Word	0.95	0.94	0.95	0.77	0.79	0.78	0.869	<b>P ≤ 0.05</b>
		Word + Char	0.95	0.95	0.95	0.80	0.78	0.79	0.869	
CNN + Bidirectional (LSTM)	Word2vec	Word	0.89	0.96	0.93	0.77	0.51	0.61	0.738	<b>P ≤ 0.05</b>
		Word + Char	0.93	0.95	0.94	0.77	0.69	0.73	0.821	
	BERT	Word	0.94	0.96	0.95	0.81	0.76	0.78	0.857	P > 0.05

(continued)

**Table 1.** (*continued*)

Model	Word Embedding	Level	No Intervention (0)			Intervention (1)			BA	P.V
			P	R	F1	P	R	F1		
		Word + Char	0.95	0.96	0.95	0.81	0.77	0.79	0.865	
CNN + LSTM + Attention	Word2vec	Word	0.92	0.94	0.93	0.71	0.65	0.68	0.795	<b>P ≤ 0.05</b>
		Word + Char	0.92	0.97	0.94	0.83	0.62	0.71	0.794	
	BERT	Word	0.95	0.94	0.95	0.77	0.81	0.79	<b>0.875</b>	<b>P ≤ 0.05</b>
		Word + Char	0.95	0.95	0.95	0.80	0.79	0.80	0.874	
CNN + Bidirectional (LSTM) + Attention	Word2vec	Word	0.91	0.95	0.93	0.73	0.61	0.66	0.777	<b>P ≤ 0.05</b>
		Word + Char	0.92	0.97	0.95	0.83	0.66	0.74	0.815	
	BERT	Word	0.95	0.95	0.95	0.79	0.78	0.78	0.863	P > 0.05
		Word + Char	0.95	0.95	0.95	0.78	0.79	0.78	0.868	

Also, the difference between these models is not always statistically significant. Which means BERT is good enough to represent words without any support.

## 5 Conclusion

Instructor intervention is a highly important issue in the MOOC environment. Many researchers have tried to detect when an intervention is needed in MOOC post forums, by implementing different prediction models, which have rendered different levels of accuracy. This research has explored MOOC posts needing urgent instructor attention and intervention or not, by analysing the textual contents of learners' posts and the information about courses. To reach this goal, the current study attempts to discover the best way to construct deep learning models, by using different inputs (word-based or word-character based) based on word2vec or BERT as a word embedding. Then we presented a combination of models, by applying the *plug & play* technique, which can solve a large problem by dividing it into sub-problems. This concretely means adding different inputs, stacking multiple layers, connecting layers, changing the dimensions in other layers, etc. Here, we use it to predict the urgency of posts in the Stanford

MOOCPosts dataset. Our conclusion is that using BERT as word-embedding is better to be used as standalone, without any addition of input as character.

Our approach for future work will involve utilising other datasets, as well as other environments, and evaluating whether combining other NN models can increase the performance. We also plan to implement character-input only such as CharacterBERT [42] and CharBERT [43].

## References

1. Arguello, J., Shaffer, K.: Predicting speech acts in MOOC forum posts. In: Ninth International AAAI Conference on Web and Social Media (2015)
2. Chaturvedi, S., Goldwasser, D., Daumé III, H.: Predicting instructor's intervention in MOOC forums. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers (2014)
3. Yang, T.-Y., et al.: Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE J. Sel. Topics Signal Process.* **11**(5), 716–728 (2017)
4. Wise, A.F., Cui, Y., Vytasek, J.: Bringing order to chaos in MOOC discussion forums with content-related thread identification. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM (2016)
5. Yang, D., et al.: Exploring the effect of confusion in discussion forums of massive open online courses. In: Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM (2015)
6. Crossley, S., et al.: Combining click-stream data with NLP tools to better understand MOOC completion. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM (2016)
7. Kizilcec, R.F., Halawa, S.: Attrition and achievement gaps in online learning. In: Proceedings of the Second (2015) ACM Conference on Learning@ Scale (2015)
8. Chandrasekaran, M.K., et al.: Using discourse signals for robust instructor intervention prediction. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
9. Minaee, S., et al.: Deep learning-based text classification: a comprehensive review. *ACM Comput. Surv. (CSUR)* **54**(3), 1–40 (2021)
10. Rani, S., Kumar, P.: Deep learning based sentiment analysis using convolution neural network. *Arab. J. Sci. Eng.* **44**(4), 3305–3314 (2019)
11. Devlin, J., et al.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
12. Mazari, A.C., Boudoukhani, N., Djeffal, A.: BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Comput.*, 1–15 (2023)
13. Khodeir, N.A.: Bi-GRU Urgent classification for MOOC discussion forums based on BERT. *IEEE Access* **9**, 58243–58255 (2021)
14. Yin, W., et al.: Comparative study of cnn and rnn for natural language processing. arXiv preprint [arXiv:1702.01923](https://arxiv.org/abs/1702.01923) (2017)
15. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
16. Chandrasekaran, M.K., et al.: Learning instructor intervention from mooc forums: early results and issues. arXiv preprint [arXiv:1504.07206](https://arxiv.org/abs/1504.07206) (2015)
17. Agrawal, A., et al.: YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In: The 8th International Conference on Educational Data Mining (2015)

18. Bakhtaria, A.: Towards cross-domain mooc forum post classification. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. ACM (2016)
19. Wei, X., et al.: A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information* **8**(3), 92 (2017)
20. Almatrafi, O., Johri, A., Rangwala, H.: Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Comput. Educ.* **118**, 1–9 (2018)
21. Sun, X., et al.: Identification of urgent posts in MOOC discussion forums using an improved RCNN. In: 2019 IEEE World Conference on Engineering Education (EDUNINE). IEEE (2019)
22. Alrajhi, L., Alharbi, K., Cristea, A.I.: A Multidimensional deep learner model of urgent instructor intervention need in MOOC forum posts. In: Kumar, V., Troussas, C. (eds.) ITS 2020. LNCS, vol. 12149, pp. 226–236. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-49663-0\\_27](https://doi.org/10.1007/978-3-030-49663-0_27)
23. Guo, S.X., et al.: Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums. *IEEE Access* **7**, 120522–120532 (2019)
24. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
25. Elman, J.L.: Finding structure in time. *Cogn. Sci.* **14**(2), 179–211 (1990)
26. Hochreiter, S., Schmidhuber, J., Elvezia, C.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
27. Chung, J., et al.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014)
28. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
29. Zhao, C., Han, J.G., Xu, X.: CNN and RNN based neural networks for action recognition. In: *Journal of Physics: Conference Series*. IOP Publishing (2018)
30. Ullah, A., et al.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* **6**, 1155–1166 (2017)
31. Tsironi, E., et al.: An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. *Neurocomputing* **268**, 76–86 (2017)
32. Wang, X., Jiang, W., Luo, Z.: Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (2016)
33. Lai, S., et al.: Recurrent convolutional neural networks for text classification. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015)
34. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: Gangemi, A., et al. (eds.) *The Semantic Web*, pp. 745–760. Springer International Publishing, Cham (2018). [https://doi.org/10.1007/978-3-319-93417-4\\_48](https://doi.org/10.1007/978-3-319-93417-4_48)
35. Liang, D., Xu, W., Zhao, Y.: Combining word-level and character-level representations for relation classification of informal text. In: *Proceedings of the 2nd Workshop on Representation Learning for NLP* (2017)
36. Yenigalla, P., Kar, S., Singh, C., Nagar, A., Mathur, G.: Addressing unseen word problem in text classification. In: Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F. (eds.) *Natural Language Processing and Information Systems*, pp. 339–351. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91947-8\\_36](https://doi.org/10.1007/978-3-319-91947-8_36)
37. Wise, A.F., et al.: Mining for gold: identifying content-related MOOC discussion threads across domains through linguistic modeling. *Internet High. Educ.* **32**, 11–28 (2017)
38. Clark, K., et al.: What does bert look at? an analysis of bert's attention. arXiv preprint [arXiv:1906.04341](https://arxiv.org/abs/1906.04341) (2019)

39. Yang, Z., et al.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
40. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems (2015)
41. Alamri, A., et al.: MOOC next week dropout prediction: weekly assessing time and learning patterns (2021)
42. Boukkouri, H.E., et al.: CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. arXiv preprint [arXiv:2010.10392](https://arxiv.org/abs/2010.10392) (2020)
43. Ma, W., et al.: CharBERT: character-aware pre-trained language model. arXiv preprint [arXiv: 2011.01513](https://arxiv.org/abs/2011.01513) (2020)



# Personalized Feedback Enhanced by Natural Language Processing in Intelligent Tutoring Systems

Christos Troussas<sup>(✉)</sup> , Christos Papakostas , Akrivi Krouskas ,  
Phivos Mylonas , and Cleo Sgouropoulou

Department of Informatics and Computer Engineering, University of West Attica, Egaleo,  
Greece

{ctrouss, cpapakostas, akrouskas, mylonasf, csgouro}@uniwa.gr

**Abstract.** This paper proposes a novel approach for enhancing feedback in intelligent tutoring systems (ITSs) for Java programming using natural language processing (NLP). The proposed approach overcomes the limitations of traditional rule-based feedback generation systems and provides more personalized and relevant feedback to learners. The architecture includes three main components: a natural language parser (that takes as input comments and/or questions of the user that can be inserted through a text box in the user interface.), a feedback generator, and a feedback evaluator. The natural language parser is responsible for converting the unstructured text input of the learner into structured data, which can be analyzed for generating feedback. The feedback generator component then processes this data and generates personalized feedback for the learner based on their specific needs. Finally, the feedback evaluator component assesses the quality of the generated feedback and determines its helpfulness to the learner. The evaluation results are promising, indicating that using NLP techniques can improve the overall performance of intelligent tutoring systems and provide a more personalized learning experience for students.

**Keywords:** natural language processing · intelligent tutoring systems · feedback parser · feedback generator · feedback evaluator

## 1 Introduction

Intelligent tutoring systems (ITS) are computer-based systems that provide personalized and adaptive instruction to learners [1–5], aiming to improve the learning outcomes and the efficiency of the educational process. Feedback to users of ITSs is a critical issue, as they can receive information about their performance and guidance towards achieving their learning objectives [6–9]. To provide high-quality feedback delivery, ITSs must first collect data about learners' performance, such as their responses to questions or their interaction with the learning environment. This data is then analyzed to identify their strengths and weaknesses; the system is able to determine the most appropriate feedback for their specific needs. Despite the challenges of generating high-quality feedback, ITS

have shown great potential in improving learning outcomes and providing personalized instruction to learners. As technology continues to advance, ITSs are expected to become more sophisticated and effective in providing feedback and guidance to learners.

Recent advances in natural language processing (NLP) have greatly improved the feedback generation process in intelligent tutoring systems (ITS) [10–13]. NLP can help to analyze and understand the natural language input provided by learners, such as their code comments in a computer programming course or explanations of their problem-solving approach, and use that information to provide more personalized and relevant feedback.

Leveraging state-of-the-art NLP techniques to analyze the learners' natural language input and generate personalized feedback is a promising way to enhance the feedback generation module in ITS, especially for complex domains such as programming languages. By using NLP techniques, the system can provide feedback that is tailored to learners' specific needs and level of understanding, leading to more effective learning outcomes.

In view of the above, in this paper, we propose three components incorporated in our intelligent tutoring system for Java programming to support the process of feedback delivery to learners: the natural language parser, feedback generator, and feedback evaluator. These components work together to provide a comprehensive and effective learning experience for the learner, as follows:

- The natural language parser is responsible for analyzing the learner's natural language input, such as their code comments or explanations of their problem-solving approach, and extracting relevant information. This information is then passed on to the feedback generator, which uses this information to generate personalized feedback that is tailored to the learner's specific needs and level of understanding.
- The feedback generator takes into account various factors, such as the learner's past performance, their learning objectives, and their preferred learning style, to generate feedback that is relevant, engaging, and useful.
- The feedback evaluator then assesses the effectiveness of the feedback and provides feedback to the feedback generator, allowing it to adjust its approach and improve the quality of feedback provided to the learner.
- The synergy between these three components is crucial for the success of the intelligent tutoring system. By leveraging the power of natural language processing, the natural language parser can extract valuable information from the learner's input, allowing the feedback generator to generate personalized and relevant feedback. The feedback evaluator then assesses the quality of this feedback, allowing the system to continually improve and provide more effective feedback to the learner.

The three components of the intelligent tutoring system work together in a synergistic manner to provide a more effective and personalized learning experience for the learner, ultimately improving their understanding and proficiency in Java programming.

Our approach has several advantages over traditional feedback generation approaches in ITSs. Firstly, the use of natural language processing techniques allows for more tailored feedback that is better suited to each learner's specific needs and level of understanding. Finally, by providing more useful and actionable feedback, this approach can

help to improve the overall effectiveness of the ITS. This can lead to better learning outcomes for the learners and can help them to achieve their educational and professional goals.

## 2 Related Work

The traditional rule-based approach for providing feedback in ITSs has limitations, as it relies on pre-defined rules and does not take into account the nuances of the students' response or the context in which it was provided [14–16]. This can lead to generic and unhelpful feedback that does not address the students' specific needs.

In recent years, there has been a growing interest in using NLP techniques to analyze student responses and provide personalized feedback in ITSs. By analyzing the students' natural language input, more relevant and targeted feedback can be provided.

Recent research has shown promising results in using deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for natural language understanding in ITS for computer programming education [17]. In [18], the authors presented computer-assisted system for creating customized multiple-choice questions with different levels of difficulty and covering various types of skills such as vocabulary, grammar, and reading comprehension. In [19], the authors have introduced a novel technique to automatically offer feedback for basic programming exercises. The presented system determines the smallest possible modifications required to rectify the errors in a student's incorrect solution, giving them an accurate understanding of the extent to which their solution was incorrect and informing them about their mistakes.

Analyzing further the related literature, in [20], the authors investigate the possibility of utilizing Artificial Intelligence (AI) and natural language processing in a teaching assistant role to produce an experience of interaction that is close to that with a human, and to also offer a way for teachers to understand the learning progress of each student.. In [21], the authors introduce an intelligent tutor that utilizes AI to respond to programming inquiries from students.

By incorporating a feedback evaluation component, the proposed approach can provide a more comprehensive and accurate assessment of the quality of the feedback generated by the system. This can help to ensure that the feedback provided is not only personalized and relevant but also effective in improving the learner's understanding of Java programming. This is an important addition to the existing literature and can contribute significantly to the development of more effective and efficient ITS.

## 3 Methods and Architecture

The proposed system is designed for an intelligent tutoring system for Java programming, but the approach can be adapted for other programming languages or domains, if adequately altered. The system is intended to help learners improve their programming skills by providing them with personalized feedback that addresses their specific weakness. The system includes three main components, i.e. the Natural Language Parser, the Feedback Generator and the Feedback Evaluator.

### 3.1 Natural Language Parser

The natural language parser component can use various NLP techniques such as tokenization, part-of-speech tagging, and dependency parsing to analyze the learners' comments.

It can identify the main topic of the comment, extract key concepts, and understand the relationships between different words in the comment. This information can then be used to generate personalized feedback for the learner. The natural language parser component can also take into account the learners' background knowledge, learning history, and other contextual information to provide more relevant and effective feedback.

The natural language parser component is crucial in understanding the learner's input and extracting the relevant information needed for generating personalized feedback. In the case of Java programming, the parser needs to identify the code snippet or concept being referred to by the learner's input and interpret it in a way that the feedback generator can use to provide accurate and helpful feedback. For example, the learner may ask "What is the purpose of the 'public static void main (String[] args)' method in Java?" or "Why is my code not compiling?". The natural language parser component can then analyze these comments and extract information, such as the concept or code snippet being referred to, and use this information to generate appropriate feedback.

The user interface is an important component of the ITS, as it is the primary point of interaction between the learner and the system. The text box provided for the learner to input their comments or questions is prominently displayed and labeled clearly..

The architecture of this component typically involves several sub-components that work together to analyze the text and identify important features. The natural language parser is responsible for analyzing the natural language input provided by the student and extracting relevant information that can be used by the tutoring system to provide feedback. The process typically involves several steps:

- Tokenization: A tokenizer is the first step in the process, which breaks the text into individual tokens or words. The input text is broken down into individual words or tokens, which are held for further processing.
- Part-of-speech tagging: The part-of-speech (POS) tagger then labels each token with its corresponding part of speech, such as noun, verb, adjective, etc. This helps the system identify the function of each word in the input text.
- Named-entity recognition: The named-entity recognizer (NER) is used to identify and classify named entities, such as person names, organization names, and location names.
- Dependency parsing: The dependency parser analyzes the relationships between words in the sentence and creates a tree structure that represents the syntactic structure of the sentence. This can help the system understand the meaning of the sentence and extract relevant information.

These sub-components work together to enable the natural language parser component to understand the meaning of the text and extract important information from it.

Once the natural language parser has analyzed the input text and extracted relevant information, the information can be passed on to the feedback generator component for further processing. In view of the above, the series of steps made by the Natural Language Parser is as follows:

1. Initialization
2. Tokenize the input text into a list of words
3. Apply part-of-speech tagging to the list of words
4. Apply named-entity recognition to the list of words
5. Apply dependency parsing to the list of words
6. Extract relevant information, such as the main topic and key concepts, of the response
7. Return the extracted information

### 3.2 Feedback Generator

The Feedback Generator component is responsible for generating feedback based on the output of the natural language parser. It uses this output to provide additional resources or examples to help the student better understand that concept.

The process typically involves the following steps:

- Analysis of the input information: The feedback generator analyzes the information extracted by the natural language parser to identify the key concepts and ideas inserted by the learner.
- Identification of strengths and weaknesses: Based on the analysis of the input information, the feedback generator identifies areas where the student has demonstrated a strong understanding of the material, as well as areas where the student may need further clarification or explanation. For example, if the learner has demonstrated a good understanding of object-oriented programming concepts but struggles with Java syntax, the feedback generator might provide additional resources or examples to help the learner improve their syntax skills. Similarly, if the learner has demonstrated good problem-solving skills but struggles with debugging code, the feedback generator might provide additional examples or resources on debugging techniques.
- Generation of feedback: Using the information gathered in the previous steps, the feedback generator generates feedback that is tailored to the learner's specific needs and level of understanding. The feedback may include suggestions for improvement, examples to illustrate key concepts, or explanations of difficult concepts. The feedback can take various forms, such as textual explanations, examples, or links to additional resources. The feedback generator may also provide positive reinforcement to the learner, highlighting areas where they are doing well and encouraging them to continue to build on their strengths.

Once the feedback has been generated, it can be evaluated by the feedback evaluator component. In view of the above, the steps that can be made by the Feedback Generator is as follows:

1. Initialization

2. Receive the extracted information from the natural language parser
3. Use a rule-based system to generate personalized feedback for the student
4. Return the generated feedback

### 3.3 Feedback Evaluator

The feedback evaluator component is crucial for improving the overall effectiveness of the intelligent tutoring system. By collecting feedback from the learner and analyzing the quality of the feedback provided, the system can make adjustments to its approach to better meet the needs of the student. This helps to ensure that the feedback provided is effective and helpful, ultimately leading to improved learning outcomes. For example, in case of an unhelpful feedback, the system can adjust its approach for providing better feedback on that topic in the future.

The process involves several steps:

- Collection of feedback: This process involves collecting feedback from the student on the feedback provided by the system. This can be done through various methods such as surveys, questionnaires, or direct feedback in the user interface of the intelligent tutoring system. The feedback collected can include the learner's opinion on the usefulness of the feedback, whether it addressed their needs, and suggestions for improvement.
- Evaluation of feedback: The feedback evaluator evaluates the quality of the feedback using a predefined metric. This helps the system identify areas where the feedback could be improved.
- Incorporation of feedback: Based on the results of the evaluation, the feedback evaluator makes improvements to the system, such as adding new examples, clarifying explanations, or modifying the feedback generation algorithm.
- Iteration: It refers to the process of revising and improving the feedback evaluator's recommendations. This process typically involves re-analyzing the input information, identifying any additional strengths and weaknesses, and generating new feedback that addresses the student's individual needs more effectively.

By continually evaluating and improving the feedback provided by the system, the feedback evaluator helps to ensure that the system is effective in helping students learn and master the material. In view of the above, the steps that can be made by the Feedback Evaluator is as follows:

1. Initialization
2. Collect feedback from the student on the effectiveness of the feedback provided by the system
3. Evaluate the quality of the feedback using a predefined metric
4. Use the feedback and evaluation results to identify areas for improvement
5. Incorporate the feedback and evaluation results into the system to improve the quality of future feedback
6. Return the updated system

## 4 Example of Operation

Suppose that the system receives the following response from a student who is asked to explain the difference between an interface and a class in Java:

“An interface is like a blueprint for a class, while a class is an actual implementation. Interfaces define a set of methods that must be implemented by a class that implements the interface, while classes can have their own methods and properties.”

The natural language parser would make the following steps:

1. Initialization
2. Tokenize the input text into a list of words: [“An”, “interface”, “is”, “like”, “a”, “blueprint”, “for”, “a”, “class”, “;”, “while”, “a”, “class”, “is”, “an”, “actual”, “implementation”, “.”, “Interfaces”, “define”, “a”, “set”, “of”, “methods”, “that”, “must”, “be”, “implemented”, “by”, “a”, “class”, “that”, “implements”, “the”, “interface”, “;”, “while”, “classes”, “can”, “have”, “their”, “own”, “methods”, “and”, “properties”, “.”]
3. Apply part-of-speech tagging to the list of words: [(“An”, ‘DT’ (Determiner)), (“interface”, ‘N’ (Noun)), (“is”, ‘VBZ’ (Verb, 3<sup>rd</sup> person singular present)), (“like”, IN’ (Preposition)), (“a”, ‘DT’), (“blueprint”, ‘NN’), (“for”, ‘IN’), (“a”, ‘DT’), (“class”, ‘NN’), (“;”, ‘;’), (“while”, ‘IN’), (“a”, ‘DT’), (“class”, ‘NN’), etc.]

Using the extracted information from the natural language parser, the feedback generator component might generate the following feedback:

“Great explanation! You’ve correctly identified that an interface is like a blueprint for a class and that classes are actual implementations. You also correctly noted that interfaces define a set of methods that must be implemented by a class that implements the interface, while classes can have their own methods and properties. Keep up the good work!”

Suppose the student responds to the feedback with the following comment:

“Thanks for the feedback! I think it would be helpful if you could provide more examples of interfaces and classes in Java”. The feedback evaluator component can process this feedback, following the steps in Subsect. 3.3.

Analyzing the above example of operation, it can be inferred that the natural language parser component was able to extract the relevant information from the student’s response, such as the difference between an interface and a class in Java. The feedback generator component then used this information to generate personalized feedback that addressed the student’s needs. Finally, the feedback evaluator component assessed the quality of the feedback and determined whether it was helpful or not to the student.

## 5 Evaluation

The evaluation process took place for a whole academic semester during the tutoring of the undergraduate course of “Java Programming” in the school of engineering of a public university of the capital city of the country. In particular, three educators, and 110 undergraduate students, participated in the evaluation process. All the measurements

of gender and age were derived from a randomly selected sample and did not have an impact on our research findings.

The population was equally divided by the instructors in two groups, each of which had equal number of students. The first group, namely the experimental group, were asked to use the ITS using NLP techniques, while the second group, namely the control group, did not take advantage of the feedback generation module.

User satisfaction is an important metric for evaluating the effectiveness of an intelligent tutoring system that incorporates personalized feedback using natural language processing. To this direction, a questionnaire was used to measure users' perceptions of the system's effectiveness, ease of use, and overall satisfaction.

A *t*-test analysis was used to compare the mean scores of the two groups of students in order to determine if there is a statistically significant difference in user satisfaction between learners who have used the ITS that incorporates personalized feedback using NLP and those who have not. The *t*-test analysis involved the following steps:

1. Selection of the participants: the two groups were similar in terms of their demographic characteristics and academic background.
2. Survey administration: we administered a questionnaire to both groups of learners to measure their perceptions of the system's effectiveness, ease of use, and overall satisfaction.
3. Mean scores computation: we calculated the mean scores for each group on each of the survey questions.
4. Conduction of the *t*-test: we conducted a two-sample *t*-test to determine if there is a statistically significant difference in the mean scores between the two groups. The *t*-test provided a *p*-value that indicated the probability of observing the difference in mean scores by chance alone.
5. Interpretation of the results: If the *p*-value is less than the significance level (set at .05), then there is evidence to suggest that the difference in mean scores between the two groups is statistically significant. This would indicate that the ITS is more effective at improving user satisfaction than traditional learning methods.

Overall, the *t*-test analysis provides valuable insights into the effectiveness of an ITS that incorporates personalized feedback using NLP by comparing user satisfaction between learners who have used the system and those who have not. However, it's important to note that the results of the *t*-test should be interpreted in conjunction with other metrics such as learning outcomes and engagement metrics to provide a more comprehensive evaluation of the system's effectiveness.

After the completion of the course at the end of the semester, the two groups, experimental and control group, were asked to answer a questionnaire, based on a 7-point Likert scale ranging from (1) strongly disagree to (7) strongly agree (Table 1).

The responses to the aforementioned questions provide valuable insights into the user experience of the ITS and help identify areas for improvement. User satisfaction provides valuable feedback on the usability, effectiveness, and overall satisfaction of the system, which can be used to improve the system's design and functionality.

A statistical hypothesis test was used to assess the proposed system more thoroughly. The 2-tailed *t*-test results are presented in Table 2.

**Table 1.** Evaluation questions of user satisfaction.

Measurement	Question
System's effectiveness	1. Does the system provide feedback that is relevant and useful to your needs?
	2. Does the feedback help you improve your understanding and performance?
Ease of use	3. Is the user interface of the system easy to use?
	4. Are the instructions clear and easy to understand?
Overall satisfaction	5. How satisfied are you with the system overall?
	6. Would you recommend the system to others?

**Table 2.** *t*-Test results.

	Question 1		Question 2		Question 3		
	Experimental group	Control group	Experimental Group	Control group	Experimental Group	Control group	
Mean	6.08	3.49	6.48	3.66	6.35	3.18	
Variance	0.78	0.36	0.39	0.45	0.54	0.59	
t-Stat	2.71		4.19		3.99		
P two-tail	0.0014		0.00055		0.00037		
t Critical two-tail	2.03		1.83		1.90		
		Question 4		Question 5		Question 6	
	Experimental group	Control group	Experimental Group	Control group	Experimental Group	Control group	
	5.99	3.58	6.31	3.55	6.05	3.13	
Variance	0.81	0.39	0.37	0.49	0.50	0.67	
t-Stat	2.65		4.22		3.76		
P two-tail	0.0019		0.00054		0.00041		
t Critical two-tail	2.01		1.88		1.94		

It is recorded a significant difference between the mean values of all the six questions. The results are somehow expected as the ITS incorporates NLP techniques, and students realized improved learning outcomes. Since *t*-Test values for the six questions are greater than the critical *t*, the results suggest that our system had a significant positive effect on the students' satisfaction.

## 6 Conclusions

This paper proposes a natural language processing-based approach to enhance feedback in intelligent tutoring systems for Java programming education. The proposed approach consists of three main components: a natural language parser, a feedback generator, and a feedback evaluator. The natural language parser converts learner comments into structured data that can be analyzed, while the feedback generator uses this data to generate personalized feedback for the learner. The feedback evaluator component then assesses the quality of the generated feedback and determines its helpfulness to the learner. The proposed approach was demonstrated through example of operation and was assessed by real students.

Future research can further explore the use of natural language processing techniques in intelligent tutoring systems for programming education. One potential area of exploration is the use of more advanced natural language processing technique to further enhance the feedback generation process. Additionally, the proposed feedback evaluation component could be further refined to better assess the quality of the generated feedback.

## References

1. Yang, C.C.Y., Ogata, H.: Personalized learning analytics intervention approach for enhancing student learning achievement and behavioral engagement in blended learning. *Educ. Inf. Technol.* (2022). <https://doi.org/10.1007/s10639-022-11291-2>
2. Krouská, A., Troussas, C., Sgouropoulou, C.: A personalized brain-based quiz game for improving students' cognitive functions. In: Frasson, C., Bamidis, P., Vlamos, P. (eds.) *BFAL 2020. LNCS (LNAI)*, vol. 12462, pp. 102–106. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-60735-7\\_11](https://doi.org/10.1007/978-3-030-60735-7_11)
3. Krouská, A., Troussas, C., Sgouropoulou, C.: Applying genetic algorithms for recommending adequate competitors in mobile game-based learning environments. In: Kumar, V., Troussas, C. (eds.) *ITS 2020. LNCS*, vol. 12149, pp. 196–204. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-49663-0\\_23](https://doi.org/10.1007/978-3-030-49663-0_23)
4. Troussas, C., Chrysafiadi, K., Virvou, M.: Machine learning and fuzzy logic techniques for personalized tutoring of foreign languages. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 358–362. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93846-2\\_67](https://doi.org/10.1007/978-3-319-93846-2_67)
5. Kanetaki, Z., Stergiou, C., Bekas, G., Troussas, C., Sgouropoulou, C.: Analysis of engineering student data in online higher education during the COVID-19 pandemic. *Int. J. Eng. Pedag. (iJEP)* **11**(6), 27–49 (2021). <https://doi.org/10.3991/ijep.v11i6.23259>
6. Bellarhmouch, Y., Jeghal, A., Tairi, H., et al.: A proposed architectural learner model for a personalized learning environment. *Educ. Inf. Technol.* (2022). <https://doi.org/10.1007/s10639-022-11392-y>
7. Troussas, C., Krouská, A., Sgouropoulou, C.: Enriching mobile learning software with interactive activities and motivational feedback for advancing users' high-level cognitive skills. *Computers* **11**(2), 18 (2022). <https://doi.org/10.3390/computers11020018>
8. Troussas, C., Krouská, A., Sgouropoulou, C.: Improving learner-computer interaction through intelligent learning material delivery using instructional design modeling. *Entropy* **23**(6), 688 (2021). <https://doi.org/10.3390/e23060668>

9. Troussas, C., Krouska, A., Sgouropoulou, C.: Impact of social networking for advancing learners' knowledge in E-learning environments. *Educ. Inf. Technol.* **26**(4), 4285–4305 (2021). <https://doi.org/10.1007/s10639-021-10483-6>
10. Katz, A., Norris, M., Alsharif, A.M., Klopfer, M.D., Knight, D.B., Grohs, J.R.: Using natural language processing to facilitate student feedback analysis. In: 2021 ASEE Virtual Annual Conference (2021)
11. Kastrati, Z., Dalipi, F., Imran, A.S., Pireva Nuci, K., Wani, M.A.: Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study. *Appl. Sci.* **11**(9), 3986 (2021). <https://doi.org/10.3390/app11093986>
12. Dalipi, F., Zdravkova, K., Ahlgren, F.: Sentiment analysis of students' feedback in MOOCs: a systematic literature review. *Front. Artif. Intell.* **4**, 728708 (2021). <https://doi.org/10.3389/frai.2021.728708>
13. Sangeetha, D.R., Hegde, P.V., Prerana, N.G., Bh, M.K.: Feedback and recommendation system using natural language processing. *SSAHE J. Interdisc. Res.* **17**(1), 17–27 (2020)
14. Albreiki, B., Habuza, T., Shuqfa, Z., Serhani, M.A., Zaki, N., Harous, S.: Customized rule-based model to identify at-risk students and propose rational remedial actions. *Big Data Cogn. Comput.* **5**(4), 71 (2021). <https://doi.org/10.3390/bdcc5040071>
15. Chan, H.C., Wei, K.K., Siau, K.L.: A rule-based system for query feedback. In: Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences, Wailea, HI, USA, pp. 53–61 (1993). <https://doi.org/10.1109/HICSS.1993.284286>
16. Kuo, J.-Y., Lin, H.-C., Wang, P.-F., Nie, Z.-G.: A feedback system supporting students approaching a high-level programming course. *Appl. Sci.* **12**(14), 7064 (2022). <https://doi.org/10.3390/app12147064>
17. Xiao, C., Zhang, Y., Liu, B., Liu, H.: Deep learning for natural language processing in intelligent tutoring systems: a systematic review. *IEEE Trans. Educ.* **64**(1), 50–61 (2021)
18. Huang, Y., Chen, M.C., Sun, Y.S.: Development and evaluation of a personalized computer-aided question generation for english learners to improve proficiency and correct mistakes. ArXiv, abs/1808.09732 (2018). <https://doi.org/10.48550/arXiv.1808.09732>
19. Singh, R., Gulwani, S., Solar-Lezama, A.: Automated feedback generation for introductory programming assignments. In: Proceedings of the 34<sup>th</sup> ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 213), pp. 15–26. ACM, New York (2013). <https://doi.org/10.1145/2491956.2462195>
20. Mathew, A.N., Rohini, V., Paulose, J.: NLP-based personal learning assistant for school education. *Int. J. Electr. Comput. Eng.* **11**(5), 4522–4530 (2021). <https://doi.org/10.11591/ijece.v11i5.pp4522-4530>
21. Roldán-Álvarez, D.: Smart tutor to provide feedback in programming courses. ArXiv, abs/1808.09732, abs/2301.09918 (2023). DOI: <https://doi.org/10.48550/arXiv.2301.09918>



# SC-QUES: A Sentence Completion Question Dataset for English as a Second Language Learners

Qiongqiong Liu<sup>1</sup> , Yaying Huang<sup>1</sup> , Zitao Liu<sup>2</sup> , Shuyan Huang<sup>1</sup> , Jiahao Chen<sup>1</sup> , Xiangyu Zhao<sup>3</sup> , Guimin Lin<sup>4</sup> , Yuyu Zhou<sup>2</sup> , and Weiqi Luo<sup>2</sup>

<sup>1</sup> TAL Education Group, Beijing, China

{liuqiongqiong1,huangyaying1,huangshuyan,chenjiahao}@tal.com

<sup>2</sup> Guangdong Institute of Smart Education, Jinan University, Guangzhou, China  
{liuzitao,zyy,1wq}@jnu.edu.cn

<sup>3</sup> City University of Hong Kong, Hong Kong, China  
xianzhao@cityu.edu.hk

<sup>4</sup> Shenzhen Everants Technology, Shenzhen, China  
lincank@everants.com

**Abstract.** Sentence completion (SC) questions present a sentence with one or more blanks that need to be filled in, three to five possible words or phrases as options. SC questions are widely used for students learning English as a Second Language (ESL). In this paper, we present a large-scale SC dataset, SC-QUES, which is made up of 289,148 ESL SC questions from real-world standardized English examinations. Furthermore, we build a comprehensive benchmark of automatically solving the SC questions by training the large-scale pre-trained language models on the proposed SC-QUES dataset. We conduct detailed analysis of the baseline models performance, limitations and trade-offs. The data and our code are available for research purposes from: <https://github.com/ai4ed/SC-Ques>.

## 1 Introduction

Standardized examination is one of the crucial elements in worldwide education systems of teaching English as a Second Language (ESL) [3, 53]. They have proved a necessary source of evaluation data for investigating and diagnosing the situations that ESL learners grasp the essential language knowledge [10, 13, 31]. The standardized ESL examinations are usually organized in various formats of well-designed language processing tasks to evaluate specific capabilities.

Previous researchers have spent lots of efforts in designing such language proficiency evaluation tasks which can be summarized into three categories: narrow, intermediate, and general [4, 48, 53]. These tasks are designed to assess different ESL language understanding capabilities from word, sentence, paragraph/document levels. For examples, among the narrow tasks, the identification of synonyms and antonyms has been widely used in the Test of English

as a Foreign Language (TOEFL), Graduate Record Exams (GRE) and many other ESL standardized exams [33, 44, 45]. General tasks involve more logical and comprehensive abilities such as logic puzzles in the Law School Administration Test (LSAT) [9, 23, 24], reading comprehension questions from GRE [34, 39], etc. The intermediate tasks stand between the word-level narrow tasks and the paragraph-level general tasks and focus on the sentence level language understanding. Among all the intermediate tasks, sentence completion (SC) questions are well-studied as one of the classic representatives [22, 42, 49]. SC questions present a sentence with one or more blanks that need to be filled in. Three to five possible words (or short phrases) are given as options for each blank and only one of the options yields to a reasonable sentence. An example of SC question is shown in Table 1.

**Table 1.** An illustrative example of SC questions.

- 
- That T-shirt with Yao Ming’s picture on it \_\_\_\_ belong to John.  
He likes him a lot.  
— No, it \_\_\_\_ be his. He hates black color.  
(A) can; can’t (B) may; needn’t (C) must; mustn’t (D) must; can’t
- 

An intelligent computational approach to SC in ESL is able to provide instant feedback to students and help students learn and practice ESL questions anytime anywhere. Besides, it provides feasible solutions to evaluate distractors in SC questions and helps teachers revise and improve the overall qualities of SC questions. Furthermore, the questions can also be used to generate more personalized questions that match the students’ knowledge mastery, the mastery of students can be obtained by the knowledge tracing task [5, 28–30]. Hence, constructing a SC dataset is essential to building an intelligent SC approach.

Although there are a few publicly available SC datasets such as MSR SC dataset<sup>1</sup> [42, 48, 52], their sample sizes are very limited and only contain a few thousand SC questions. Such small datasets are not able to align with the power of the state-of-the-art (SOTA) pre-trained language models (LMs). Furthermore, the number of missing blanks and the length of candidate tokens are fixed in existing open-sourced SC datasets [52]. However, in the real-world English learning scenario, SC questions are usually presented in diverse forms with irregular numbers of blanks and various lengths of to-be-filled tokens.

To tackle the above limitations in existing SC datasets, we introduce SC-QUES, a large-scale SC question dataset for ESL learners. The proposed SC-QUES dataset contains 289,148 questions with one or more missing blanks. To the best of our knowledge, SC-QUES is one of the largest SC question dataset for ESL learners. Meanwhile, we fine-tune the existing SOTA pre-trained LMs

---

<sup>1</sup> <https://www.microsoft.com/en-us/research/project/msr-sentence-completion-challenge/>.

on the SC-QUES dataset and present a systematic benchmark of automatically solving the ESL SC questions in students' real-life scenarios. We conduct comprehensive experimental evaluations to understand the performance impacts of different lengths of SC questions contexts and different numbers of words in candidate options. We conduct the precision-recall trade-off analysis and discuss the practical issues when deploying the SC question AI solver in real-world educational contexts.

## 2 The SC-Ques Dataset

### 2.1 Data Collection

SC questions in SC-QUES are real English assignment questions used for K-12 students and they are developed by teaching content professionals from one of the largest educational technology companies in China. Each SC question in SC-QUES is represented as the single select multiple choice question format that consists of three parts: (1) question stem that denotes the main text content of the target ESL question with one or more blanks; (2) candidate options that are represented as a predetermined set of responses of at least three or more options; and (3) answer that indicates the correct one to be filled in the blanks.

Even though all the SC questions are manually developed by teaching professionals, there exists some data duplication, missing data problems or data corruption issues. Therefore, we conduct a series of steps to filter out or clean up the duplication and ill-formed SC questions. Specifically, we remove the duplicate question with the same options. And we filter out SC questions that if the question stem, candidate options, or answer is missing. Furthermore, we only remain SC questions whose number of candidate options is equal to 3 or 4.

### 2.2 SC Question Categorization

Due to the arbitrary number of blanks and tokens in the ESL SC tasks, questions in SC-QUES are categorized into the following categories according to the blank numbers and tokens numbers and the examples of each category are illustrated in Table 2.

- **C1: One-blank and one-token.** Questions have one to-be-filled blank and the longest candidate option has one token.
- **C2: One-blank and many-token.** Questions have one to-be-filled blank and the longest candidate option has multiple tokens.
- **C3: Many-blank and one-token.** Questions have more than one to-be-filled blanks and the longest candidate option has one token.
- **C4: Many-blank and many-token.** Questions have more than one to-be-filled blanks and the longest candidate option has multiple tokens.

**Table 2.** Illustrative examples of SC questions from different categories.

C1	Jack is five years old now. He can dress ____. (A) herself (B) himself (C) yourself
	Sam and Mike aren't at school. ____ are they? (A) How (B) Who (C) Where
C2	— I'm very tall. Where can I sit? — ____ the classroom. (A) In the front of (B) At the back of (C) In the middle of
	— I find that I have been unlucky these days. — ____ Everything will become better soon. (A) Keep silent! (B) Why not cry? (C) Cheer up! (D) How about you?
C3	— Whose are those jeans? — This pair of jeans ____ mine and that two pairs ____ my brother's. (A) are; are (B) are; is (C) is; are
	— That T-shirt with Yao Ming's picture on it ____ belong to John. He likes him a lot. — No, it ____ be his. He hates black color. (A) can; can't (B) may; needn't (C) must; mustn't (D) must; can't
C4	We have ____ homework to do every day so we can't play ____ computer games. (A) too many; too many (B) too much; too many (C) too many; too much (D) too much; too much
	____ of their bad habits, the boys changed from being dependent on their parents to ____ for others. (A) Having rid; sacrificing (B) To rid; sacrifice (C) Rid; sacrificing (D) Having been rid; sacrifice

## 2.3 Data Statistics

After data cleaning and preprocessing, in total, we end up with 289,148 SC questions in SC-QUES. Specifically, we have 110,645 133,249, 27,886, and 17,368 SC questions in categories C1, C2, C3 and C4 respectively. 84.35% of SC questions have one blank missing. The length distribution of the question stems in SC-QUES is shown in Fig. 1. As we can see that most of the questions contain 5 to 15 words and sentences of 9 words have the largest portion. There are only 5.88% questions that have more than 20 words. This indicates SC questions usually have very limited contexts and imply subtle linguistic knowledge such as grammar, syntax, and semantics, the performance of automatically solving the ESL SC questions may vary a lot.

## 3 Benchmark

### 3.1 Problem Formulation

Let  $\mathbf{q}$  be the SC question stem with one or more blanks. Let  $\mathbf{o}_1, \dots, \mathbf{o}_m$  be the candidate options associated with  $\mathbf{q}$ . Solving the SC question is to find the option that leads to the highest correct probability after completing the to-be-filled sentence with the selected option, i.e.,  $\arg \max_{i=1, \dots, m} \Pr(\mathbf{o}_i | \mathbf{q})$ .

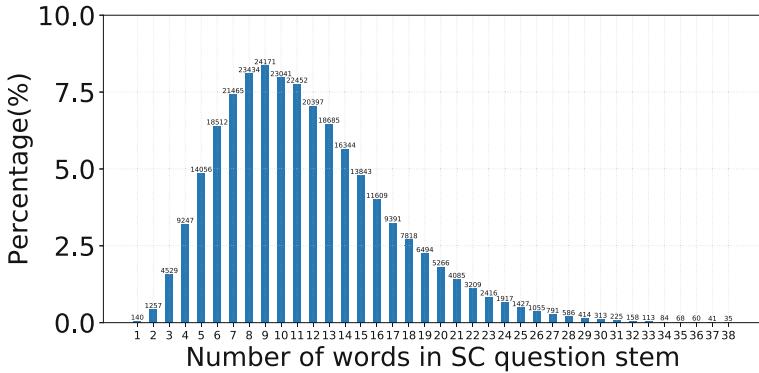


Fig. 1. Length distribution of SC questions in SC-QUES.

### 3.2 Benchmark Workflow

In this work, we view the above problem as a predictive task and we aim to train a classifier to find the correct answer from the option pool of confusing candidates. Specifically, we first fill candidate options into the corresponding blanks to get complete sentences. Then we treat sentences that contain the correct options as positive examples and the rest as negative examples. After that, we utilize a neural LM model to extract the semantically meaningful information within each sentence in SC questions and make final SC question predictions via a multilayer perceptron (MLP). The overall end-to-end benchmark workflow is shown in Fig. 2.

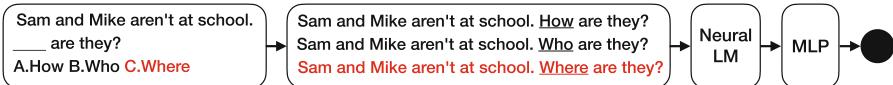


Fig. 2. The overall benchmark workflow. The sentence in red color denotes the correct option and is denoted as the positive example in the training process. (Color figure online)

**The Neural Language Model.** Large-scale pre-trained LMs such as BERT [19], RoBERTa [27], XLNet [50], BART [25] and DeBERTa [17], benefit from self-supervised learning on a large amount of corpus, and has shown its competent generalization to various downstream natural language understanding tasks. In this work, we conduct a systematic benchmark evaluation by measuring the prediction accuracy with existing prevalent large-scale pre-trained LMs listed as follows. Please note that we choose to use their large model versions for experiments. Since we expect each SC question to be solved correctly, we choose to use accuracy as our evaluation metric.

- BERT [19]. A pre-trained natural language understanding model with transformer encoder blocks. We fine-tune BERT on our SC dataset as a sentence classification task. We use a special start token ([CLS]) as the first token of every text sequence and the final hidden state corresponding to this token is used as the aggregated SC sentence representation.
- RoBERTa [27]. RoBERTa improves BERT by replacing static masking with dynamic masking, pre-training more epochs with larger batch size, and removing the next sentence prediction task. We follow the same fine-tuning protocol as described in the BERT method.
- XLNet [50]. XLNet is an autoregressive based pre-training method with transformer decoder blocks. Similar to BERT, we fine-tune the XLNet and utilize the last hidden state as the SC representation.
- BART [25]. BART adapts standard Transformer [46] as its backbone model and utilizes a denoising autoencoder for pretraining sequence-to-sequence models. It is pre-trained to recover different types of text corruptions to their original versions, such as sentence permutation and text infilling.
- DeBERTaV3 [16]. DeBERTaV3 improves the original DeBERTa [17] model by replacing mask language modeling with replaced token detection. The model is trained as a discriminator to predict whether a token in the corrupted input is either original or replaced by a generator.

**The MLP Prediction.** Once we obtain the final hidden state  $\mathbf{t}_n$  from above neural LM module as the aggregated SC sentence representation. We introduce two additional fully-connected layers to perform the binary classification task, i.e.,  $\mathbf{x} = \text{softmax}(\mathbf{W}_1 \tanh(\mathbf{W}_0 \mathbf{t}_n + \mathbf{b}_0) + \mathbf{b}_1)$ , where  $\mathbf{W}_0$ ,  $\mathbf{W}_1$ ,  $\mathbf{b}_0$  and  $\mathbf{b}_1$  are trainable parameters,  $\mathbf{W}_0 \in \mathbb{R}^{1024 \times d}$ ,  $\mathbf{b}_0 \in \mathbb{R}^{1024}$ ,  $\mathbf{W}_1 \in \mathbb{R}^{2 \times 1024}$  and  $\mathbf{b}_1 \in \mathbb{R}^2$ ,  $\mathbf{t}_n \in \mathbb{R}^d$ . The first entry of  $\mathbf{x}$  gives the probability of wrong option while the second entry gives right option probability. The objective is to minimize the cross entropy of the right or wrong option labels.

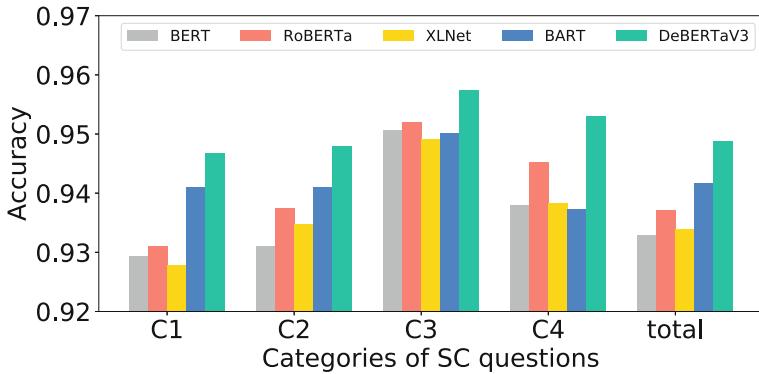
### 3.3 Experimental Setup and Details

We randomly split the entire dataset into training set and testing set with 241,195 and 47,953 SC questions respectively in an approximate ratio of 5 : 1. For BERT, RoBERTa, XLNet and DeBERTaV3, each model has 24 hidden layers, 1024 hidden size and 16 attention heads. For the BART model, it has 12 hidden layers and 16 attention heads in each of the encoder and decoder, and the hidden size is 1024. We employ the AdamW [21] as the optimizer with an initial learning rate of 1e-5 for all models. We fine-tune our model on 4 T V100 GPU devices. Due to the limited memory of GPU devices, the batch size of BERT is set to 32, and the other models set the batch size as 16. The max length of the sentences is set to 128.

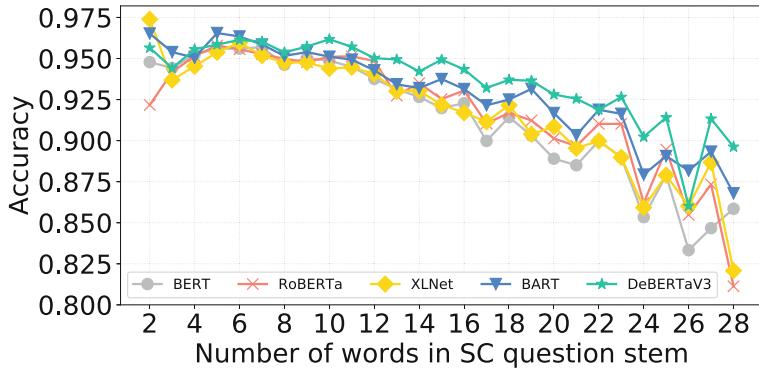
### 3.4 Results

**Overall Performance.** As we can see from Fig. 3, DeBERTaV3 outperforms other models in terms of prediction errors on the entire testing set. This indi-

cates DeBERTaV3 has better generalization performance among all the selected pre-trained LMs which may be due to the effective token detection task to discriminate whether the corrupted token is original or replaced by a generator in DeBERTaV3. When comparing the prediction performance of all the methods on C1, C2, C3 and C4, we can see that C3 and C4 have better performance than C1 and C2, we think the reason is that C1 and C2 have only one blank, the difference of the completed sentences of each option are quite similar which makes the models hard to distinguish this subtle difference. We run the pairwise t-test for each possible pair from these methods and all of them are statistically significantly different at 0.01 level.



**Fig. 3.** Accuracy on different categories of SC-QUES.



**Fig. 4.** Length distribution of SC questions in SC-QUES.

**Impacts of Different Lengths of SC Questions Contexts.** We study the impacts of different lengths of questions and the results are shown in Fig. 4. From

the model performance curve, when the length of the questions is larger than 10, the models' performances drop sharply. This indicates that after exceeding a certain length, questions become more difficult with the increase of the context length. The reason may be that the model is difficult to capture the contextual information in such a long distance. What is interesting is that if the length of the question is less than 3, the models also show slightly poor performance, which indicates that it is hard for the model to distinguish different options once there is not enough context information. More specifically, the performance of DeBERTaV3 is less than the performance of BART when the length of the question is less than 7. With the increasing sequence length, DeBERTaV3 outperforms other models. We believe the reason is that DeBERTaV3 is trained as a discriminator to detect whether the token is original, the longer sequences provide relatively effective information to make DeBERTaV3 discriminate the tokens, so they can get higher prediction results of SC compared to other models.

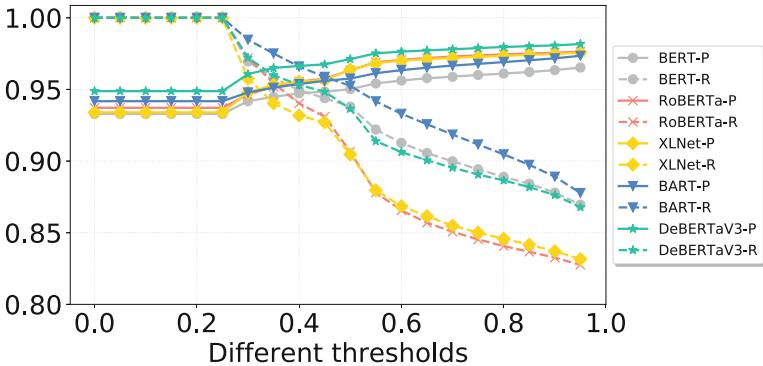
**Impacts of Different Numbers of Words in Candidate Options.** We also study the performance impacts in terms of the length of candidate options. We focus on the questions which have only one blank and two blanks. As shown in Table 3, we can see that the trend is similar to the different lengths of the questions, DeBERTaV3 achieves the best performance in all lengths of candidate options. If the length of the candidate options is too short or too long, the performance would be a little worse.

**Table 3.** Accuracy on different lengths of the longest option in the questions which have *one blank* and *two blanks*.

Methods	one blank				two blanks			
	1	2	3	$\geq 4$	2	3	4	$\geq 5$
BERT	0.9294	0.9433	0.9307	0.8984	0.9496	0.9388	0.9489	0.9161
RoBERTa	0.9311	0.9481	0.9345	0.9130	0.9527	0.9467	0.9506	0.9367
XLNet	0.9278	0.9439	0.9323	0.9136	0.9483	0.9324	0.9489	0.9288
BART	0.9411	0.9505	0.9393	0.9184	0.9503	0.9388	0.9465	0.9272
DeBERTaV3	<b>0.9468</b>	<b>0.9543</b>	<b>0.9450</b>	<b>0.9349</b>	<b>0.9568</b>	<b>0.9571</b>	<b>0.9571</b>	<b>0.9383</b>

**Precision-Recall Trade-off Analysis.** When deploying the model in practice, a wrong answer may give bad guidance to students. In order to reduce such problems, we may refuse to solve some difficult questions and improve the precision of more solvable questions. The models in our benchmark will output the correct probability for each option, for each question, we use a softmax function with temperature 0.1 for the probabilities of all options, the final highest probability can also be viewed as the problem-solvable confidence score. After that, we set a threshold to the correct probability of the model's selected option and accept the above-the-threshold questions as our solvable questions. The recall score is computed as (the number of solvable questions)/(the number of all test questions), and the precision score is calculated as (the number of both solvable and

correct-answered questions)/(the number of solvable questions). The precision and recall curves of all the pre-trained LMs in different thresholds are shown in Fig. 5. We can see that with the threshold growing, the precision is higher and the recall becomes smaller. When the threshold is 0.95, the precision of most of the models are higher than 97.0% and the recall scores keep greater than 80.0%. Especially, regardless of the threshold values, the precision of DeBERTaV3 is always higher than 95.0% and the recall scores never lower than 87.0%.



**Fig. 5.** The precision-recall curve.

## 4 Related Work

### 4.1 Sentence Completion Datasets

SC datasets always include in-completed sentences which require selecting the most appropriate answer from several options. Recently, there are several publicly available SC datasets. MSR is a SC challenge data consisting of 1,040 sentences with four impostor sentences for each original sentence [52]. CODAH involves multiple choice SC questions, each SC question contains four options which are consistent with commonsense [6]. Park and Park collected 1823 cloze-style questions with multiple-choice questions from the Test of Proficiency in Korean (TOPIK) [36]. Nozza et al. provided a manually labeled hurtful SC dataset in six languages [35]. Compared with the above works, our proposed SC-QUES is a relatively large-scale SC dataset with more than 290 thousand ESL SC questions.

### 4.2 Sentence Completion Solvers

Various approaches have been proposed to automatically solve the ESL SC questions. Earlier literature prefer complete the sentence by filling the blank with each

candidate option in turn, and to evaluate its likelihood under an LM [7, 8, 53]. Motivated by the successful applications in language modeling via deep neural networks [14, 32, 40], recent studies tackle the SC challenges based on the recurrent neural networks (RNN) [43, 49]. For examples, Yang and Deng proposed a global context dependent RNN LM to obtain the global semantics of the target in the SC tasks [49]. Since large-scale pre-trained LMs become remarkably powerful models for language understanding and attain unprecedented levels of performance gains on multiple language tasks [18, 19, 37, 38], there are some approaches utilize pre-trained LMs to solve the SC questions [11, 41, 51]. Donahue et al. trained the language model by using the concatenation of artificially-masked texts and the texts which are masked as input [11]. Besides above research works of solving SC questions via language modeling, researchers also make attempts from different machine learning perspectives [1, 2, 12, 15, 26, 48]. Liu et al. developed an gradient search based iterative inference algorithm that can be broadly applied to any neural sequence generative model for the fill-in-the-blank tasks [26]. Researchers also studied automatically solve questions in domains other than ESL, such as history [47], science [20], etc.

## 5 Conclusion

In this paper, we construct a large-scale SC question dataset, SC-QUES, for ESL Learners. The SC-QUES consists of 289,148 SC questions with four categories of questions that contain almost all SC classifications in the English examinations of the real world. Furthermore, we present a benchmark on the proposed SC-QUES for automatically solving the ESL SC questions. The experimental results indicate the effectiveness of our dataset and the proposed benchmark.

**Acknowledgments.** This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500; in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission and in part by NFSC under Grant No. 61877029 and in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

## References

1. Argouarc'h, J.: Dependency, skip-grams, association tensors and machine learning for sentence completion. In: Proceedings of the 14th International Conference on Natural Language Processing (2018)
2. Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., Gelbukh, A.: Multiple choice question (MCQ) answering system for entrance examination. In: Cross-Language Evaluation Forum (2013)
3. Beinborn, L., Zesch, T., Gurevych, I.: Candidate evaluation strategies for improved difficulty prediction of language tests. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (2015)
4. Beinborn, L.M.: Predicting and manipulating the difficulty of text-completion exercises for language learning. Ph.D. thesis, Technische Universität Darmstadt (2016)

5. Chen, J., Liu, Z., Huang, S., Liu, Q., Luo, W.: Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In: AAAI Conference on Artificial Intelligence (2023)
6. Chen, M., D'Arcy, M., Liu, A., Fernandez, J., Downey, D.: CODAH: an adversarially-authored question answering dataset for common sense. In: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP (2019)
7. Chen, S.F.: Performance prediction for exponential language models. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2009)
8. Chen, S.F.: Shrinking exponential language models. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2009)
9. Chesani, F., Mello, P., Milano, M.: Solving mathematical puzzles: a challenging competition for AI. *AI Mag.* **38**(3) (2017)
10. Davey, G., De Lian, C., Higgins, L.: The university entrance examination system in china. *J. Further High. Educ.* **31**(4) (2007)
11. Donahue, C., Lee, M., Liang, P.: Enabling language models to fill in the blanks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
12. Fedus, W., Goodfellow, I., Dai, A.M.: MaskGAN: better text generation via filling in the\_. In: International Conference on Learning Representations (2018)
13. Franke, W.: The reform and abolition of the traditional Chinese examination system, vol. 10. Harvard University Asia Center (1960)
14. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013)
15. Gubbins, J., Vlachos, A.: Dependency language models for sentence completion. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013)
16. He, P., Gao, J., Chen, W.: DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations (2023)
17. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021)
18. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018)
19. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers) (2019)
20. Khot, T., Sabharwal, A., Clark, P.: Scitail: a textual entailment dataset from science question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
22. Lee, K., Lee, G.G.: Sentence completion task using web-scale data. In: 2014 International Conference on Big Data and Smart Computing (2014)

23. Lev, I.: Logic puzzles: a new test-suite for compositional semantics and reasoning (2006)
24. Lev, I., MacCartney, B., Manning, C.D., Levy, R.: Solving logic puzzles: from robust processing to precise semantics. In: Proceedings of the 2nd Workshop on Text Meaning and Interpretation (2004)
25. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
26. Liu, D., Fu, J., Liu, P., Lv, J.: TIGS: an inference algorithm for text infilling with gradient search. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
27. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
28. Liu, Z., et al.: Enhancing deep knowledge tracing with auxiliary tasks. In: The International Conference of World Wide Web (2023)
29. Liu, Z., Liu, Q., Chen, J., Huang, S., Luo, W.: simpleKT: a simple but tough-to-beat baseline for knowledge tracing. In: International Conference on Learning Representations (2023)
30. Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., Luo, W.: pyKT: a python library to benchmark deep learning based knowledge tracing models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
31. Madaus, G.F.: The effects of important tests on students: implications for a national examination system. *Phi Delta Kappan* **73**(3) (1991)
32. Mikolov, T., Kombrink, S., Burget, L., Černocký, J., Khudanpur, S.: Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (2011)
33. Mohammad, S., Dorr, B., Hirst, G.: Computing word-pair antonymy. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (2008)
34. Ng, H.T., Teo, L.H., Kwan, J.L.P.: A machine learning approach to answering questions for reading comprehension tests. In: 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (2000)
35. Nozza, D., Bianchi, F., Hovy, D.: HONEST: measuring hurtful sentence completion in language models. In: The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2021)
36. Park, H., Park, J.: Assessment of word-level neural language models for sentence completion. *Appl. Sci.* **10**(4) (2020)
37. Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (2018)
38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8) (2019)
39. Riloff, E., Thelen, M.: A rule-based question answering system for reading comprehension tests. In: ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems (2000)
40. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11) (1997)

41. Shen, T., Quach, V., Barzilay, R., Jaakkola, T.: Blank language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020)
42. Tang, E.: Assessing the effectiveness of corpus-based methods in solving SAT sentence completion questions. *J. Comput.* **11**(4) (2016)
43. Tran, K.M., Bisazza, A., Monz, C.: Recurrent memory networks for language modeling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
44. Turney, P.: A uniform approach to analogies, synonyms, antonyms, and associations. In: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK (2008)
45. Turney, P.D., Littman, M.L.: Corpus-based learning of analogies and semantic relations. *Mach. Learn.* **60**(1–3) (2005)
46. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
47. Wang, D., et al.: CMU multiple-choice question answering system at NTCIR-11 QA-Lab. In: Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (2014)
48. Woods, A.: Exploiting linguistic features for sentence completion. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2016)
49. Yang, T., Deng, H.: Intelligent sentence completion based on global context dependent recurrent neural network LM. In: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (2019)
50. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
51. Zhu, W., Hu, Z., Xing, E.: Text infilling. arXiv preprint [arXiv:1901.00158](https://arxiv.org/abs/1901.00158) (2019)
52. Zweig, G., Burges, C.J.: The Microsoft research sentence completion challenge. Technical report. Citeseer (2011)
53. Zweig, G., Platt, J.C., Meek, C., Burges, C.J., Yessenalina, A., Liu, Q.: Computational approaches to sentence completion. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2012)



# Conversational Agents and Language Models that Learn from Human Dialogues to Support Design Thinking

Stavros Demetriadis<sup>1</sup> and Yannis Dimitriadis<sup>2</sup>

<sup>1</sup> Aristotle University of Thessaloniki, Thessaloniki, Greece  
sdemetri@csd.auth.gr

<sup>2</sup> Universidad de Valladolid, Valladolid, Spain  
yannis@tel.uva.es

**Abstract.** Design is a highly creative and challenging task and research has already explored possible ways for using conversational agents (CAs) to support humans participating in co-design sessions. However, research reports that a) humans in these sessions expect more essential support from CAs, and b) it is important to develop CAs that continually learn from communication -like humans do- and not simply from labeled datasets. Addressing the above needs, this paper explores the specific question of how to extract useful knowledge from human dialogues observed during co-design sessions and make this knowledge available through a CA supporting humans in similar design activities. In our approach we explore the potential of the GPT-3 Large Language Model (LLM) to provide useful output extracted from unstructured data such as free dialogues. We provide evidence that by implementing an appropriate “extraction task” on the LLM it is possible to efficiently (and without human-in-the-loop) extract knowledge that can then be embedded in the cognitive base of a CA. We identify at least four major steps/assumptions in this process that need to be further researched, namely: A1) Knowledge modeling, A2) Extraction task, A3) LLM-based facilitation, and A4) Humans’ benefit. We provide demonstrations of the extraction and facilitation steps using the GPT-3 model and we also identify and comment on various worth exploring open research questions.

**Keywords:** Conversational agent · Large language model (LLM) · Design thinking

## 1 Introduction

Current advances in the Natural Language Processing (NLP) domain, related to a considerable degree to the rise of Artificial Neural Networks (ANN), have led to the development of Conversational Agents (CAs or ‘chatbots’), a type of software that can offer services to humans through natural language communication [1]. Research has provided encouraging initial evidence that CAs may effectively support humans in domains such as e-business, e-health and e-learning (e.g. [2, 3]) and also in the particularly creative and challenging design domain [4–6].

However, users' expectations regarding CAs supporting humans in design work, prescribe a type of agent capable of delivering almost human-like supportive interventions (e.g. [7]). CA technology needs to advance well beyond its current state-of-the-art (SOTA) to provide for this high level of agent capacity, one important advanced level functionality being agent's learning ability. Similar to humans who learn to a great extent from conversational interactions (e.g. [8]) and "distill" their experiences in some form of knowledge transferable also in other contexts, this option should certainly be explored also for CA learning. Following this line of thinking we focus on the question of how to extract useful knowledge from human dialogues in co-design sessions and embed it next in the cognitive base of a CA to support humans in sessions of similar perspective.

In this paper we explore the hypothesis that current Large Language Models (LLMs) may provide an innovative answer to this question; an answer that may open new perspectives in conceptualizing Human-AI interaction and offer significant research opportunities. We present a relevant conceptual framework and a sequence of four methodological steps labelled as "assumptions" to indicate that their feasibility, efficiency, and exact form of implementation need to be backed up by future research concrete evidence. Currently we demonstrate the feasibility of using GPT-3 in these steps but without any further evaluation of outcomes. We also highlight opportunities for further exploration of key relevant research questions. The rest of the paper is structured as follows:

In Sect. 2 we provide a concise review of SOTA advances in the LLMs and CAs domain, reporting also on the use of CAs to support design thinking. In Sect. 3 we present the four methodological steps demonstrating -where applicable- their feasibility with examples realized in the GPT-3 environment [9]. In Sect. 4 we propose an architecture for a CA implementing the aforementioned steps. We conclude by providing a non-exhaustive list of open research questions.

## 2 Background

### 2.1 Large Language Models and Conversational Agents

Artificial intelligence (AI) is "*the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings*" [10]. Using AI artifacts to perform various tasks can be conceptualized as an "augmentation" of human cognitive abilities [11] and this certainly includes large language models and conversational agents in the natural language processing domain. A language model is typically described as a probability distribution over language tokens (words or parts of words) that provides a basis to computer software in performing text predictive modelling tasks [12]. With the rise of ANNs it is possible to develop complex LLMs making use of transformer architecture enhanced also with attention mechanisms [13] (for example, BERT and GPT families of LLMs). After the initial "pre-training", methods of LLM meta-learning (known mainly as "transfer learning") can be applied to render the model more effective in downstream tasks. Of great importance, because of their effort efficiency, are the *K-shot learning* methods (*ibid.*), where "shot" refers to the training examples offered to the model to adjust it for the new task. Versions of this method include "zero-shot" (no training example at all), "one-shot" (one training example), or "few-shot" (more than one examples) [14].

CAs typically identify the user's intent (and relevant entities) in user's text input and provide an appropriate reply (for details the reader is directed to relevant publications such as [1, 15]). In general, agents employing AI-based methods are either “retrieval-based” or “generative” [1]. Mixed approaches are also possible combining benefits of the two worlds (e.g. [16]). An agent can further “learn” by improving its accuracy in identifying user's intent and by expanding its knowledge base, thus being able to provide ever more accurate and useful replies to user prompts. Research has mostly focused on agent's learning from training sets of labeled data but seldom on learning through conversational interactions with humans (e.g. [17]).

## 2.2 Conversational Agents and Design Thinking

*“Design thinking is a human-centered approach to innovation that draws from the designer's toolkit to integrate the needs of people, the possibilities of technology, and the requirements for business success”* [18]. In this paper, we use the term “co-design” session [19] to refer to the type of sessions being design-focused and full of social interactions (either onsite or online). Along with other researchers we also maintain that CAs can productively participate in co-design sessions and enhance the efficiency of humans to work collaboratively and accomplish the design objective (e.g. [7]). A review by Debowski et al. [5] distinguishes five main orientations relevant to CAs' possible contribution in design thinking sessions (with emphasis on remotely conducted sessions). Overall, the review reports evidence that interacting with a CA may improve a human co-designer's performance and the authors call for further research in the area.

Interestingly, Bittner et al. [7] report that users expect an “almost human agent” that provides “helpful and not frustrating input”, can be used in “combination with human facilitator”, and may also have beneficial “impact on team collaboration”. Research also has nevertheless provided evidence that the effectiveness of digital facilitation can be comparable to that of humans, although it appears to be significantly worse in “subjectively perceived helpfulness” [6]. To further improve CAs performance and acceptance we argue that research should explore the development of a continually learning CA for design thinking. This type of agent would learn in a way similar to humans, by integrating in its knowledge base useful information and knowledge distilled from available dialogue data from previous co-design sessions. We proceed to further explore this perspective in the next sections.

## 3 Learning from Co-designers' Dialogues

### 3.1 Extracting Information and Knowledge from Human Dialogues

The problem of extracting knowledge from human dialogues can be considered as a special case of the general problem of extracting structured information from highly unstructured text [20]. Research so far has explored different solutions to various versions of this problem and methods developed for extracting specific information from dialogues have also been reported (e.g. [21, 22]). Evidence also is available that LLMs can report on the quality of a conversation as there exists a positive correlation between the output of the language models and human-assigned dialogue scores [23].

Based on the above and other relevant evidence we argue that LLMs can offer a viable solution for the combined problem of “extracting from dialogues” and “embedding to CA facilitators” the knowledge that can be helpful for human designers. To further model the problem we state and analyse the four assumptions that follow.

### 3.2 A1: “Knowledge Modelling” Assumption

*Knowledge embedded in dialogues can/should be modelled in a way that facilitates its extraction and further processing.*

“Knowledge is best shared through conversation” [24] and modeling this knowledge is important in developing highly efficient dialogue systems (e.g., [25]). In human dialogues extracted from co-design sessions we expect to find knowledge expressed with various degrees of context-dependency and also using more than one modalities (typically language and visuals). Therefore, we assume that a useful initial modelling of interlocutors’ utterances can be based on these two key features: their “context-dependency” and “modality”.

“Context-dependency” refers to the degree that an utterance explicates a piece of knowledge in a clear and well understood way (for example a clear rule statement) or requires to be connected to other pieces of information (available or not in the dialogue) in order to become meaningful. Please note that this distinction is not necessarily relevant to the commonly discussed “tacit vs. explicit” knowledge distinction (e.g. [26]), since tacit and explicit knowledge can both be expressed with a varied degree of context-dependency in human dialogues. “Modality” refers to discussants using one or more modalities to convey information, such as spoken/written language and pictorial information (images, graphs, etc.) or face expressions, body posture/gestures, etc. We emphasize here the need for efficiently modelling the dialogue content from the context/modality perspective and possibly from other perspectives too.

### 3.3 A2: “Extraction Task” Assumption

*Useful knowledge embedded in co-design dialogues can be extracted by implementing some efficient LLM-based extraction task.*

Humans typically learn by generalizing from their experiences in a form of deeper (abstract) representation that usually can be also expressed in language-based form, for example, a summary or concluding statement. Similarly, we suggest that a LLM can extract useful knowledge from authentic human co-design dialogues in some appropriately structured form.

Figure 1 presents: a) a description of one possible extraction GPT-3 LLM task, b) one-shot training of the GPT-3, also modelling the expected reply and (c) after the one-shot training, the LLM can be used to provide similar “Topic and Conclusion” task extractions from other dialogues (as demonstrated in Fig. 2). Please note that the dialogues in Figs. 1 and 2 are synthesized based on the material (cards) published by the “Build a Bot” design workshop [27]. Specifically, the suggestion that discussant B makes regarding the design issue under consideration is based on material included in the design cards of the aforementioned workshop.

Provide the Topic and a concise Conclusion of the discussion between A and B, like the example.

A: I give you a possible bot response: "Don't worry, it all goes to the cloud (a data storage center that lives in the internet and has a lot of space)." What do you think?

B: What are we supposed to do? Criticise and provide our own alternative?

A: Brainstorming and, yes, tell me what you think and if you find this reply appropriate or you suggest one of your own.

B: I don't think most users would want sensitive data going anywhere someone could potentially see it or do something with it. I value privacy, especially as a Black person whose information has been used against my wishes or to discriminate against me. Not everyone will feel the same way or worry. This rule could give users more freedom to decide what gets saved and make them feel like they have more control. I suggest as alternative reply: "Users can choose to review their data every week before it goes to the cloud and/or have it permanently deleted."

[Topic and Conclusion]: Data storage in Chatbot discussion. Dialogue data when talking to a chatbot is an important ethical issue related to User privacy. One possible selection would be that users are able to control their data and permanently delete data if they want.

**Fig. 1.** GPT-3 one-shot training for the “Topic and Conclusion” extraction task. All GPT-3 performance examples presented in this work have been administered using the text-davinci-003 model and with Temperature parameter set to 0.7.

A: Now see a next topic I want you to consider. How traditional female names (and voices) like Siri, Alexa and Jenny can promote sexism? Would you select such names for your chatbot or suggest otherwise?

B: How much time do we have?

A: Well, let's say 10 minutes.

B: Do we post something onboard or just take a note for discussion?

A: Just few notes and suggestions.

B: Well, my suggestion is Toto. I don't think it promotes gender stereotypes and it's gender neutral. As a black cis woman, my positionality has made me very aware of gender stereotypes which is why I chose a name uncommon to humans. I think users will like it. I also like Sam as it refers to a gender neutral AI assistant and promotes the inclusion of people who are non-binary or gender non-conforming.

[Topic and Conclusion]: The discussion was about selecting genderless names for chatbots to avoid promoting sexism. Suggestions include Toto and Sam, which are gender neutral and promote inclusion of people who are non-binary or gender non-conforming.

**Fig. 2.** GPT-3 extraction from a new dialogue part in the “Topic and Conclusion” task.

### 3.4 A3: “LLM-Based Facilitation” Assumption

*Knowledge extracted from dialogues can be conveyed to co-designers by an LLM-based facilitation agent that generates replies to humans’ questions in a co-design session.*

The knowledge (conclusions) extracted from dialogues can become learning material to further train the LLM for a new downstream task, that we call “LLM-based facilitation”. We identify two procedures relevant to this “facilitation” assumption: A) “LLM training”: how to further train the LLM using the extracted knowledge, and B) “LLM advice generation”: how to have the LLM generate appropriate human-like replies as advice when requested in a co-design session. We demonstrate here one simple solution to the above two procedures starting from the LLM one-shot training (see Fig. 3).

In relation to Fig. 3 please note that:

Provide some advice to designers like the example, considering 1) the topic they discuss and 2) the conclusions from previous sessions.

Previous sessions

[Topic and Conclusion]: Data storage in Chatbot discussion. Dialogue data when talking to a chatbot is an important ethical issue related to User privacy. One possible selection would be that users are able to control their data and permanently delete data if they want.

[Topic and Conclusion]: The discussion topic was about selecting gender-neutral names for chatbot and avoiding traditional female names to promote sexism. Suggestions from A and B included Toto and Sam, both of which could promote inclusivity of non-binary and gender non-conforming users, while not promoting gender stereotypes.

Example

[Human facilitator]: Let's see now the data storage issue. As you understand has ethical consequences and is a delicate matter for users to engage in productive interaction with a bot.

[AI-facilitator]: You may like to think of the following option employed in another similar case: Users are able to control their data and permanently delete data if they want.

*a) Describing the task to the model*

*b) Training data from previous sessions*

*c) Providing an exemplary AI-facilitator reply*

**Fig. 3.** One-shot training of the GPT-3 for the “LLM-based facilitation” task.

1. The “Previous sessions” section (item b in Fig. 3) includes outcomes of the [Topic and Conclusion] extraction task (knowledge extracted from human dialogues) to be used as training material of the LLM. The [Topic and Conclusion] tag is not really necessary but it is used to help the reader’s understanding. “Previous sessions” may include numerous useful outcomes and be integrated to the LLM through API code, so technically its size can be significantly larger than demonstrated here.
2. The “Example” section (item c in Fig. 3) presents a hypothetical dialogue between a Human and an AI-facilitator. The AI-facilitator utterance is modelled to fit a scenario where advice is required on the data storage issue.

Note that “fine tuning” the LLM with a considerable amount of extracted data would also be an option but requires much more invested effort and is beyond the scope of this paper. After the above one-shot training, the GPT-3 is able to function as AI-facilitator providing appropriate replies to relevant prompts of human co-designers. In Fig. 4 we present the advice of the AI-facilitator (green highlight) to the issue of “Selecting the name of the bot” introduced by the Human Facilitator in an also hypothetical scenario. The trained LLM provides an advice based on the knowledge extracted in step A2 and integrated as learning material in the “Previous sessions” section of its one-shot training setting.

[Human facilitator]: Do you think that how the bot is called is completely neutral and not worthing any consideration by designers? See what the AI has to say.

[AI-facilitator]: It is important to consider selecting gender-neutral names for chatbot in order to avoid traditional female names to promote sexism. Suggestions from A and B included Toto and Sam, both of which could promote inclusivity of non-binary and gender non-conforming users, while not promoting gender stereotypes.

*The AI-facilitator provides relevant advice (green highlight)*

**Fig. 4.** Advice by the GPT-3-based AI-facilitator (Color figure online)

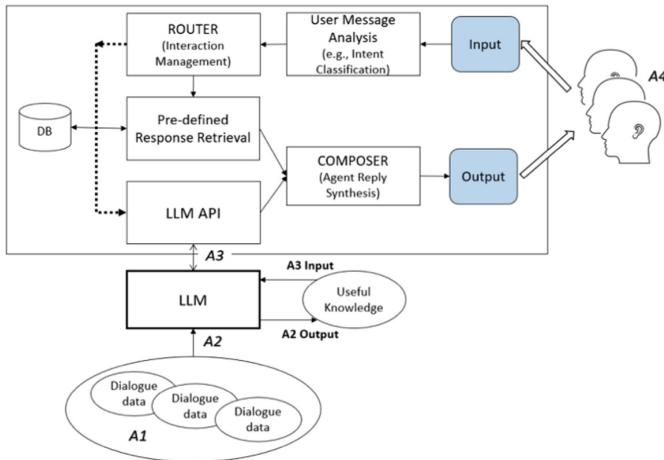
### 3.5 A4: “Humans’ Benefit” Assumption

*Human co-designers may benefit in their work when interacting with an agent that is continually learning by integrating knowledge extracted from available dialogues.*

We have demonstrated so far that by using GPT-3 it is possible to successfully perform the extraction task (step A2) and the facilitation task (step A3). However, it is not self-evident that human co-designers will reap any benefits when interacting with an agent with these functional features. Research has identified several human-agent interaction issues that go beyond the simple agent developmental level and touch the deeper socio-psychological dimensions of the interaction [28]. It is important to moderate and regulate human-agent interaction in order to establish a productive co-operation between humans and the AI-facilitator [7]. Of equal importance are issues such as enhancing an intellectually positive attitude in humans for leveraging the AI-facilitator capacity, empowering humans with control over the AI presence and also, inspiring humans how not to unproductively rely on AI but use it as a tool for deeper productive exploration of design issues (e.g. [29]). Overall, we believe there exists a major research opportunity regarding the multifaceted human-agent interactions in co-design sessions relevant also to the augmentation of human cognitive abilities through the use of AI tools and we call for further research in this area.

## 4 A Proposed Architecture for an LLM-Based AI-Facilitator

In this section we integrate the steps and procedure described so far into a coherent CA architecture able to operate as an LLM-based AI-facilitator that learns form available human designers’ conversations (Fig. 5). As described, human dialogue data can be appropriately modeled and used as training data of the LLM (A1 in Fig. 5) to perform the “extraction” task (A2) with useful knowledge as A2 output. When co-designers (A4) interact with the CA a “User Message Analysis” component provides necessary information to the “Router” component to decide whether the human’s prompt: a) should be answered by retrieving a predefined response (for example, when a user asks for information strictly relevant to the specific activity, e.g. the activity timetable), or b) it is a request for AI-facilitator’s advice which should be directed to the LLM for a reply (dotted arrows) (A3), where the LLM now uses “useful knowledge” as input data to provide relevant advice. Finally a “Composer” component can produce a synthesis of the responses obtained from the two routes (predefined and LLM-based) to be sent back to users (A4).



**Fig. 5.** Proposed CA architecture (A1 to A4 refer to the four steps/assumptions presented)

## 5 Discussion and Conclusions

In this work we have explored the question of how to extract useful knowledge inherent in human dialogues from co-design sessions and subsequently integrate this knowledge in a CA to act as AI-facilitator and support humans in similar sessions. We presented a framework of four steps/assumptions (A1 to A4) to formulate the overall procedure and demonstrated that -in principle- the use of GPT-3 LLM is an effective solution to perform both the extraction (step A2) and the facilitation task (step A3). Additionally we presented a proposed CA architecture that learns from available conversational data and conveys useful knowledge to humans when required. Obviously, there are many interesting research questions and we would like to highlight the following as seen from the broader perspective of the four suggested procedural steps/assumptions:

- A1: What type of text preprocessing and knowledge modelling could provide a basis for improved outcomes in the subsequent extraction and facilitation tasks?
- A2: What type of “extraction task” could achieve better outcomes especially when processing big conversational data?
- A3: What is the accuracy level (precision, recall, F1, etc.) achieved by a CA developed according to the proposed architecture?
- A4: How to manage the complex human-to-agent interactions in co-design sessions so that the presence of an AI-facilitator offers major benefits to humans?

**Acknowledgments.** Many thanks to colleagues and doctoral students at Universidad de Valladolid and Aristotle University of Thessaloniki for their apt comments and suggestions regarding the ideas presented in this work. This research is partially funded by the European Regional Development Fund and the National Research Agency of Spanish Ministry of Science, Innovation and Universities, under project grant PID2020-112584RB-C32. Also, partially funded by the Universidad de Valladolid visiting professorship of Prof. S. Demetriadis during 2022.

## References

1. Adamopoulou, E., Moussiades, L.: Chatbots: history, technology, and applications. *Mach. Learn. Appl.* **2**, 100006 (2020). ISSN 2666-8270. <https://doi.org/10.1016/j.mlwa.2020.100006>
2. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: a systematic review. *Comput. Educ. Artif. Intell.* **2**, 100033 (2021). ISSN 2666-920X. <https://doi.org/10.1016/j.caei.2021.100033>
3. Car, L.T., et al.: Conversational agents in health care: scoping review and conceptual analysis. *J. Med. Internet Res.* **22**(8) (2020). <https://doi.org/10.2196/17158>
4. Cautela C., Mortati M., Dell'Era C., Gastaldi L.: The impact of artificial intelligence on design thinking practice: insights from the ecosystem of start-ups. *Strat. Des. Res. J.* **12**(1), 114–134 (2019). ISSN: 1984–2988. <https://doi.org/10.4013/sdrj.2019.121.08>
5. Debowski, N., Tavanapour, N., Bittner, E.A.C.: Conversational agents in creative work – a systematic literature review and research agenda for remote design thinking. In: PACIS 2022 Proceedings, vol. 261 (2022). <https://aisel.aisnet.org/pacis2022/261>
6. Przybilla, L., Baar, L., Wiesche, M., Krcmar, H.: Machines as teammates in creative teams: digital facilitation of the dual pathways to creativity. In Proceedings of the SIGMIS-CPR 2019 Conference, pp. 94–102. ACM, New York (2019). <https://doi.org/10.1145/3322385.3322402>
7. Bittner, E.A., Mirbabaie, M., Morana, S.: Digital facilitation assistance for collaborative, creative design processes. In: Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS), pp. 370–379 (2021). <http://hdl.handle.net/10125/70655>
8. Tenenbaum, H.R., Winstone, N.E., Leman, P.J., Avery, R.E.: How effective is peer interaction in facilitating learning? a meta-analysis. *J. Educ. Psychol.* **112**(7), 1303–1319 (2020). <https://doi.org/10.1037/edu0000436>
9. OpenAI GPT-3 Playground. <https://gpt3demo.com/apps/openai-gpt-3-playground>
10. Copeland, B.: Artificial intelligence. Encyclopedia Britannica. <https://www.britannica.com/technology/artificial-intelligence>
11. Siemens, G., et al.: Human and artificial cognition. *Comput. Educ. Artif. Intell.* **3**, 100107 (2022). <https://doi.org/10.1016/j.caei.2022.100107>
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019 Conference, Minneapolis, pp. 4171–4186 (2019). <https://doi.org/10.48550/arXiv.1810.04805>
13. Vaswani, A., et al.: Attention is all you need. In: Proceedings of NIPS 2017 Conference Red Hook, USA, pp. 6000–6010 (2017). <https://doi.org/10.48550/arXiv.1706.03762>
14. Brown, T.B., Mann, B., Nick, R., et al.: Language models are few-shot learners. In: arXiv: [2005.14165v4](https://arxiv.org/abs/2005.14165v4) [cs.CL] (2020). <https://doi.org/10.48550/arXiv.2005.14165>
15. McTear, M.: Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. University of Toronto, Graeme Hirst (2020)
16. Song, Y., Li, C.-T., Nie, J.-Y., Zhang, M., Zhao, D., Yan, R.: An ensemble of retrieval-based and generation-based human-computer conversation systems. In: Proceedings of IJCAI-2018, Stockholm, Sweden, pp. 4382–4388. <https://doi.org/10.24963/ijcai.2018/609>
17. Hancock, B., Bordes, A., Mazare, P.-E., Weston, J.: Learning from dialogue after deployment: feed yourself, chatbot! In: Proceedings of the 57th Annual Meeting of ACL, pp. 3667–3684. ACL, Florence (2019). <https://arxiv.org/abs/1901.05415v1>
18. Brown, T.: Design Thinking Defined. IDEO Design Thinking Website. <https://designthinking.ideo.com/>
19. Blakemore, L.: What does co-design really mean? UTS (University of Technology, Sydney) Website. <https://lx.uts.edu.au/blog/2022/01/10/what-does-co-design-really-mean/>
20. Adnan, K., Akbar, R.: An analytical study of information extraction from unstructured and multidimensional big data. *J. Big Data* **6**(1), 1–38 (2019). <https://doi.org/10.1186/s40537-019-0254-8>

21. Kannan, A., Chen, K., Jaunzeikare, D., Rajkomar, A.R.: Semi-supervised learning for information extraction from dialogue. In: Proceedings of Interspeech Conference, Hyderabad (2018). [https://www.isca-speech.org/archive\\_v0/Interspeech\\_2018/pdfs/1318.pdf](https://www.isca-speech.org/archive_v0/Interspeech_2018/pdfs/1318.pdf)
22. Wu, C.-S., Madotto, A., Lin, Z., Xu, P., Fung, P.: Getting to know you: user attribute extraction from dialogues. In: Proceedings of the 12th LREC Conference, Marseille, pp. 581–589 (2020). [arXiv:1908.04621v1 \[cs.CL\]](https://doi.org/10.48550/arXiv.1908.04621). <https://doi.org/10.48550/arXiv.1908.04621>
23. Nedelchev, R.; Lehmann, J., Usbeck, R.: Language model transformers as evaluators for open-domain dialogues. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 6797–6808 (Online) (2020). <https://aclanthology.org/2020.coling-main.599.pdf>
24. Gurteen, D.: Conversational Leadership. Online chapter: “Sharing Knowledge through conversation. <https://conversational-leadership.net/sharing-knowledge-through-face-to-face-conversation/>
25. Varonina, L., Kopp, S.: Knowledge modelling for establishment of common ground in dialogue systems. Italian J. Comput. Linguist. **7**–1(2), 9–26 (2021). <https://doi.org/10.4000/ijcl.797>
26. Virkus, S.: Key Concepts in Information and Knowledge Management. Tallinn University (2014). [https://www.tlu.ee/~sirvir/Information%20and%20Knowledge%20Management/Key\\_Concepts\\_of\\_IKM/index.html](https://www.tlu.ee/~sirvir/Information%20and%20Knowledge%20Management/Key_Concepts_of_IKM/index.html)
27. Mogos, A., McBain, L., Stariha, M., Carter, C.: “Build a Bot” design workshop. Stanford d. School. <https://dschool.stanford.edu/resources/build-your-own-bot>
28. Chaves, A.P., Gerosa, M.A.: How should my chatbot interact? a survey on social characteristics in human-chatbot interaction design. Int. J. Hum.-Comput. Interact. **37**(8), 729–758 (2021). <https://doi.org/10.1080/10447318.2020.1841438>
29. Shneiderman, B.: Human-Centered Artificial Intelligence: Trusted, Reliable & Safe. [arXiv: 2002.04087v1 \[cs.HC\]](https://doi.org/10.48550/arXiv.2002.04087) (2020). <https://doi.org/10.48550/arXiv.2002.04087>

# Author Index

## A

- Abuazeh, Moh'd 171  
Alrajhi, Laila 651  
Antoine, Maxime 381  
Arroyo, Ivon 298  
Assaf, Alyssa Hajj 66

## B

- Badier, Anaëlle 54  
Banjade, Rabin 124  
Bekhter, Danylo 565  
Ben Abdessalem, Hamdi 66, 209, 381, 442, 491, 530  
Ben Soussia, Amal 100  
Benassi, Mariagrazia 16  
Bey, Anis 290  
Blanchard, Emmanuel G. 147  
Bloemen, Steven 257  
Boyer, Anne 100  
Bredeweg, Bert 257  
Brockmann, Annalise 137  
Brzoza, Piotr 42

## C

- Castro, Francisco 298  
Champagnat, Ronan 290  
Chang, Maiga 333, 342, 630  
Chen, Jiahao 678  
Córdova, Paulo Roberto 281  
Courtemanche, Marc-Antoine 29, 183  
Cristea, Alexandra I. 651

## D

- Daniela-Maria, Cristea 518  
de Azevedo Restelli Tedesco, Patricia Cabral 75  
de Barros Costa, Evandro 75  
de Blas, Clara Simon 391  
De Silva, Supun 88  
de Souza Filho, Iderli Pereira 281  
Demetriadis, Stavros 591, 691

- Di Eugenio, Barbara 112  
Dimitriadis, Yannis 691  
Doru-Anastasiu, Popescu 518  
Duong-Trung, Nghia 16

## E

- Edwards, Lucas 137

## F

- Feng, Guang 507  
Frasson, Claude 66, 209, 381, 442, 491, 530  
Freedman, Reva 137

## G

- Gamillscheg, Florian 415  
Gašpar, Angelina 220  
Ghaderi, Maryam 209, 381  
Gharbaoui, Hiba 364  
Grubišić, Ani 220  
Grüttmüller, Martin 16  
Guedes, Gilleanes Thorwald Araujo 281  
Gumerov, Marat 195  
Gütl, Christian 415

## H

- Hanse, Joris 257  
Hargreaves, Carol Anne 467  
Hayashi, Yugo 404  
Hensen, Benedikt 565  
Holt, Joanna 257  
Hood, Colton 415  
Huang, Shuyan 678  
Huang, Xiaoshan 430  
Huang, Yaying 678

## I

- Iriarte, Maria F. 630  
Ismail, Ibrahim 576  
Ivrissimtzis, Ioannis 479

**J**

- Jang, Yeonju 3  
 Jung, Heeseok 3

**K**

- Kaiss, Wijdane 159  
 Kano, Masaji 404  
 Kawulok, Mateusz 42  
 Khalaj, Amin Bonyad 209  
 Kiron, Nafisul 307  
 Kirste, Thomas 171  
 Kragten, Marco 257  
 Krahn, Ted 333  
 Kravcik, Milos 16  
 Kroucka, Akrivi 667  
 Kuo, Rita 333, 630

**L**

- La Barbera, Michele 315  
 LaBarbera, Dean 137  
 Lajoie, Susanne P. 430  
 Lefevre, Marie 54  
 Lefort, Mathieu 54  
 Lemire, Daniel 324  
 Li, Shan 430  
 Li, Zhaoxing 479, 553  
 Lin, Fuhua 88  
 Lin, Guimin 678  
 Liu, Qiongqiong 678  
 Liu, Xin 507  
 Liu, Zitao 678  
 Luo, Weiqi 678

**M**

- Maćkowski, Michał 42  
 Mansouri, Khalifa 159, 364  
 Matias, Ana 604  
 Meitus, Vladymyr 391  
 Miladi, Fatma 324  
 Mohamed, Azza 576  
 Monaikul, Natawut 112  
 Morimoto, Naoyuki 638  
 Morita, Junya 404  
 Moussiades, Lefteris 270  
 Mylonas, Phivos 667

**N**

- Naples, Virginia 137  
 Nastase, Mariana Madalina 352  
 Nicolae, Bold 518  
 Nielsen, Rodney 619  
 Nikou, Stavros A. 342  
 Nkambou, Roger 29, 183, 541  
 Nkulu-Ily, Yves S. 452

**O**

- Ohmoto, Yoshimasa 404  
 Oli, Priti 124  
 Omar, Mehnuma Tabassum 307  
 Orsoni, Matteo 16

**P**

- Pan, Tingfeng 507  
 Papakostas, Christos 667  
 Pietracupa, Massimo 530  
 Pietroszek, Krzysztof 415  
 Pögelt, Alexander 16  
 Poirier, Franck 159, 364  
 Popescu, Elvira 352  
 Psyché, Valéry 183, 324

**R**

- Ranniger, Claudia 415  
 Rasul, Injila 298  
 Rebol, Manuel 415  
 Rocha, Hemilis Joyce Barbosa 75  
 Rocha, Julios Suruagi 75  
 Roussanaly, Azim 100  
 Rus, Vasile 124  
 Rutenberg, Adam 415

**S**

- Sciarrone, Filippo 315  
 Sgouropoulou, Cleo 667  
 Shankar, Bharath 467  
 Shi, Lei 479, 553  
 Shiga, Kentaro 638  
 Shimojo, Shigen 404  
 Sikka, Neal 415  
 Sosnowski, Tomasz 171  
 Spinczyk, Dominik 42  
 Steinmauerer, Alexander 415  
 Sullivan, Ian 137  
 Sychev, Oleg 195, 233

**T**

- Tato, Ange Adrienne Nyamen 183  
Tato, Ange 29, 541  
Tato, Gabrielle 541  
Tchio, Guy Carlos Tamkodjou 183  
Temperini, Marco 315  
Troussas, Christos 667  
Tzimas, Dimitrios 591

**U**

- Ubani, Solomon 619  
Uglev, Viktor 233, 371

**V**

- Vaendel, Dennis 257  
Vasić, Daniel 220  
Vassileva, Julita 307  
Vicari, Rosa Maria 281

**W**

- Wang, Jindi 479, 553  
Wu, Wenyuan 507

**Y**

- Yessad, Amel 247  
Yoo, Jaesang 3  
Yoon, Yohaan 3  
Yordanova, Kristina 171

**Z**

- Zarour, Mahdi 491  
Zhao, Xiangyu 678  
Zhou, Yunzhan 479, 553  
Zhou, Yuyu 678  
Zipitria, Iraide 604  
Žitko, Branko 220  
Zografos, George 270  
Zohiar, Mai 576  
Zou, Di 630