論文 / 著書情報
Article / Book Information

題目(和文)	パーソナライズされたオンライン適応学習システム
Title(English)	Personalized Online Adaptive Learning System
著者(和文)	CARLON May Kristine Jonson
Author(English)	May Kristine Jonson Carlon
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第12136号, 授与年月日:2021年9月24日, 学位の種別:課程博士, 審査員:CROSS JEFFREY SCOTT,髙田 潤一,野原 佳代子,山下 幸彦,岡崎 直観,BERRAR DANIEL PETER
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第12136号, Conferred date:2021/9/24, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

PERSONALIZED ONLINE ADAPTIVE LEARNING SYSTEM

パーソナライズされたオンライン適応学習システム

MAY KRISTINE JONSON CARLON

SUPERVISED BY:
PROFESSOR JEFFREY SCOTT CROSS

Doctor of Philosophy
Global Engineering for Development, Environment, and Society
Dissertation

Department of Transdisciplinary Science and Engineering
School of Environment and Society
Tokyo Institute of Technology

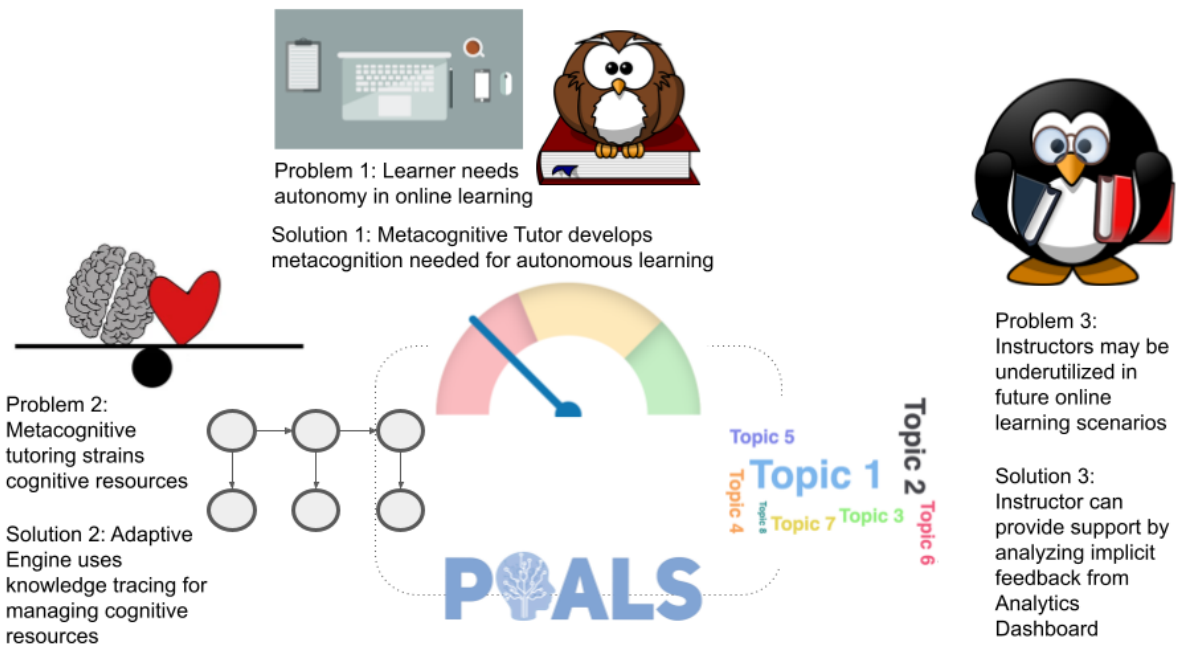
September 2021

May Kristine Jonson Carlon.
Personalized Online Adaptive Learning System, パーソナライズされたオンライン適応学習システム
Doctor of Philosophy, Global Engineering for Development, Environment, and Society
SUPERVISED BY Professor Jeffrey Scott Cross

ABSTRACT

Studies have shown that metacognition contributes to learners' academic performance and improves their growth mindset. However, training for metacognition, a domain-independent skill, usually involves learning a domain-specific skill alongside. This puts strain on the learners' cognitive resources. In this research, we used learner input-based prompts in the open edX learning management system to help learners develop metacognitive skills in online learning platforms. We then explored machine learning algorithms using metacognitive measures for adaptive learning to lessen fatigue. Finally, we used natural language processing techniques to obtain feedback from the learners' interaction that can be used by instructors to provide necessary learning interventions. Remaining areas to be explored include deployment of new adaptive learning algorithms, metacognitive tutoring qualitative studies, and further refinements on the instructor feedback.

Keywords: *metacognition, self-regulated learning, online learning, learning management system, learning tools integration, technology acceptance, adaptive learning, offline measures, online measures, mixed methods, knowledge tracing, latent variable models, hidden Markov models, artificial neural networks, deep learning, learning analytics, learning dashboard, feedback, sentiment analysis, topic modeling, textual similarity, Japanese natural language processing, algorithmic fairness*



PUBLICATIONS AND PRESENTATIONS

REFEREED

M. K. J. Carlon and J. S. Cross, "A review of quantitative offline measurement tools for computer-based metacognitive tutoring effectiveness assessment," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Dec. 2020, pp. 258–264. DOI: 10.1109/TALE48869.2020.9368470, Details in Sections 1.2, 2.1.

——, "Learning analytics dashboard prototype for implicit feedback from metacognitive prompt responses," in *29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education*, 2021, Details in Section 4.8.

——, "Development of open-response prompt-based metacognitive tutor for online classrooms," in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Submitted, Details in Sections 2.2, 2.3.3.

——, "Knowledge tracing for adaptive learning in a metacognitive tutor," *Open Education Studies*, Submitted, Details in Chapter 3.

UNREFEREED

M. K. J. Carlon and J. S. Cross, "Challenges of developing a metacognitive tutor on Open edX," in *Fourteenth Asia-Oceania Top University League on Engineering Student Conference*, 2019, Details in Section A.1.

——, "Developing learner metacognitive skills on an online environment," in *The 2020 Annual Spring Conference of Japan Society for Educational Technology*, 2020, Details in Section 2.3.1.

——, "Open response prompts in an online metacognitive tutor," in *The 2021 Annual Spring Conference of Japan Society for Educational Technology*, 2021, Details in Section 2.3.2.

——, "POALS Analytics Engine: A student affect dashboard," in *The Eighth UK Japan Engineering Education League Workshop*, 2021, Details in Chapter 4.

M. K. J. Carlon, C. Seng, and J. S. Cross, "Countering negative Matthew effect in undergraduate research with metacognition and digital learning," in *The 2020 Annual Fall Conference of Japan Society for Educational Technology*, 2020, Mentioned in Chapter 5.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

AI artificial intelligence

ANN artificial neural network

BERT Bidirectional Encoder Representations from Transformers

BKT Bayesian Knowledge Tracing

CMOOC connectivist MOOC

COVID-19 Corona Virus Disease 2019

DKT Deep Knowledge Tracing

DL deep learning

DNN Deep Neural Network

EM expectation maximization

FC Fully Correct

FI Fully Incorrect

FOB Fully Optimistic Bias

FPB Fully Pessimistic Bias

GAMES Goal-oriented studying, Active studying, Meaningful and memorable studying, Explain to understand, and Self-monitor

GPT-3 Generative Pre-trained Transformer 3

HMM hidden Markov model

HITL human-in-the-loop

ITS intelligent tutoring system

IRT Item Response Theory

JMOOC Japan Massive Online Course Consortium

JSPS Japan Society for Promotion of Science

KMA Knowledge Monitoring Accuracy

KMB Knowledge Monitoring Bias

KOC knowledge of cognition

LAD learning analytics dashboard

LDA latent Dirichlet allocation

LMS learning management system

LOESS locally estimated scatterplot smoothing

LTI Learning Tools Interoperability

MAI Metacognitive Awareness Inventory

MAPS Motivation, Agency, Possible Selves

MOOC massive open online course

MVP minimum viable product

NB No Bias

NLP natural language processing

NLTK Natural Language Toolkit

PC Partially Correct

POALS Personalized Online Adaptive Learning System

POB Partially Optimistic Bias

PPB Partial Pessimistic Bias

RA Reflection Assistant

RA-ANN Reflection Assistant with artificial neural network

RA-BKT Reflection Assistant with Bayesian Knowledge Tracing

RNN recurrent neural network

ROC regulation of cognition

SDG Sustainable Development Goal

SPOC small private online course

SSL Secured Sockets Layer

TAM Technology Acceptance Model

TF-IDF term frequency-inverse document frequency

TOKYOTECH Tokyo Institute of Technology

UN United Nations

URL uniform resource locator

VADER Valence Aware Dictionary and sEntiment Reasoner

1 INTRODUCTION

The White Rabbit put on his spectacles. "Where shall I begin, please your Majesty?" he asked. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
–Lewis Caroll, *Alice in Wonderland*

1.1 A FEW THEORIES ON LEARNING

When speaking about the beginnings of the academe, it would not be surprising if people think of the ancient Greek philosophers in the Hellenistic period. After all, the word *academia* itself traces its roots to *Akademia*, Plato's school of philosophy. Additionally, a learning method that is still being used up to today can be attributed to the great philosopher, Socrates[1]. Socratic method is a form of inquiry where individuals exchange questions and answers to understand concepts. The idea is that humans learn by scrutinizing their theories and addressing holes with reasoning and logic. Plato, Socrates' student, refined the Socratic method by writing down the dialogues or exchanges between several individuals.

[1] *H. Delić and S. Bećirović, "Socratic method as an approach to teaching,"* European Researcher. Series A, *no. 10, pp. 511–517, 2016.*

To have a dialogue, at least an individual must be knowledgeable. But what does it take to know or to think? Aristotle, Plato's student, analyzed this question by differentiating sensing, imagining, and thinking. Aristotle postulated that thinking starts from recognizing previous knowledge acquired, then this knowledge is combined to form new unexplored ideas. Thus, for Aristotle, learning is all in the mind and separate from outside circumstances. More importantly, Aristotle sees learning as two-fold: awareness of what is already known and creating new knowledge by synthesizing what is already known.

[2] *J. Piaget, "Cognitive development in children: Piaget,"* Journal of Research in Science Teaching, *vol. 2, no. 3, pp. 176–186, 1964.*

Jean Piaget formalized the idea of forming knowledge from existing knowledge as **constructivism**[2]. Constructivism, or the theory of cognitive development, cemented the notion of learning as a developmental process. This developmental point-of-view was supported by Piaget's contemporaries such as Lev Vygotsky, who defined the **zone of proximal development**. The zone of proximal development is the "goldilocks" zone, where an individual learns best. In this zone, the learner can only complete a task with sufficient guidance; thus, it is beyond where the learner can accomplish independently but not too far out that it is discouraging for the learner.

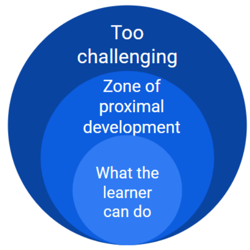


Figure 1: Zone of Proximal Development. The "goldilocks" zone is just enough to challenge the learner, but not too much to discourage them.

Another way to look at Aristotle's idea of learning is knowing what is already known. John Flavell formalized this as **metacognition**, or one's ability to be aware of what they already know and take necessary actions to influence their learning[3]. This drives home the idea that learning is active: learners can exercise control over how they learn if they have sufficient awareness. Thus, while a guide may be critical for learning as advanced by Vygotsky, much of learning remains a personal experience.

[3] *J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry,"* American Psychologist, *vol. 34, no. 10, p. 906, 1979.*

1.2 CHANGING EDUCATIONAL LANDSCAPES

Portions published as: M. K. J. Carlon and J. S. Cross, "A review of quantitative offline measurement tools for computer-based metacognitive tutoring effectiveness assessment," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Dec. 2020, pp. 258–264. DOI: 10.1109/TALE48869.2020.9368470

Two of the 21st-century salient points learners face right now include our society's transition to the knowledge economy and the growing importance of computers in education.

1.2.1 *Transformation to Knowledge Economy*

We are transitioning into a knowledge-based economy, also known as Society 5.0[4]. This shift is characterized by the exponential growth in knowledge resources and digital transformation. Significant societal transformation such as Society 5.0 is not unique in our history, the most recent of which is the Industrial Revolution. With these significant transformations come paradigm shifts in labor and, consequently, in education.

[4] *Japan Cabinet Office. (nd). Society 5.0, [Online]. Available:* https://www8.cao.go.jp/cstp/english/society5_0/index.html.

Significant technological advances caused highly specialized work done by artisans to be doable by unskilled labor using machines during the Industrial Revolution. Tasks usually done by a single artisan became decomposed into assembly line work where pieces of machinery do not require the same acuity as an artisan to complete the job. Because of the assembly line setup, more products can be produced in a shorter amount of time. This made the production process less labor-intensive per product, offsetting the high capital cost modernization entailed at that time[5].

[5] *C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?"* Technological Forecasting and Social Change, *vol. 114, pp. 254–280, 2017.*

With Society 5.0, modernization is geared towards automation of tasks that do not require highly skilled labor. The general aim is to enhance human well-being by freeing them from mundane labor and seeking more fulfilling work. Thus, unlike the industrial revolution where highly skilled artisans are negatively affected and replaced by unskilled labor, Society 5.0 favors highly skilled laborers. It also puts pressure on the labor market to continuously upskill. Ongoing computerization through robotics and automation can lead to the obsoletion of some jobs and the creation of new ones. Several companies have embraced computer-assisted learning to update their learning culture and acknowledge the growing presence of future of work and digital transformation[6].

In 2015, the United Nations (UN) released the Sustainable Development Goals (SDGs) which are targeted to be accomplished by 2030; one of the SDGs is on Quality Education (SDG4). As of UN 's report in 2017, 60% of schools in primary and secondary education in developing countries have access to the internet. Thus, online education has a great potential in achieving SDG4[7]. But quality education does not stop at accessibility; it is also important to equip learners with the necessary skills that can make them successful in the face of challenges posed by a knowledge economy.

In the past, literacy and numeracy skills were essential as these are key to enabling learners to accumulate content and knowledge. However, with the knowledge economy, knowledge has become more accessible. Hence, the mere accumulation of knowledge has become insufficient. Recent pedagogical movements focused on skills that foster engagement and deeper learning. This is to better prepare learners with new workplace demands for flexibility and adaptability. One of these so-called 21st-century skills is metacognition[8], the core of this research. Metacognition enables autonomous learning, or learners having the capacity to learn at their own pace.

1.2.2 *Changes in Online Learning*

Most computer-based instruction enables self-paced learning to various extents. For instance, most massive open online courses (MOOCs) are primarily offered in an asynchronous format, even for courses with schedules defined by the teachers. Furthermore, the proliferation of mobile devices enables learners to participate in microlearning –- a pedagogy where lessons are packed in few minute bursts –- on small pockets of time, such as during commute[9]. However, this freedom afforded to learners has its disadvantages; one of the most salient is that the lack of guidance and immediate feedback can result in high dropout rates[10].

The need to address the lack of guidance in computer-based instruction is becoming more imminent as the role of computer-based instruction in education has increased significantly in the past decade. In 2012, MOOCs had gained a surge of interest, and the year was named the "Year of the MOOC"[11]. Shortly after, the Georgia Institute of Technology offered its first MOOC-based graduate program. Several universities quickly followed: edX is expected to host 16 such programs in 2020, while Coursera intends to host 13 in the same year[12]. Thus, not only are digital platforms becoming more critical

[6] *Web Courseworks. (2021). eLearning hype curve predictions, [Online]. Available:* https://webcourseworks.com/elearning-predictions-hype-curve/.

[7] *United Nations Sustainable Development Goals. (2017). Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all, [Online]. Available:* https://unstats.un.org/sdgs/report/2017/goal-04/.

[8] *E. R. Lai and M. Viering, "Assessing 21st century skills: Integrating research findings,"* Pearson, *2012.*

[9] *I. Jahnke, Y.-M. Lee, M. Pham,* et al.*, "Unpacking the inherent design principles of mobile microlearning,"* Technology, Knowledge and Learning, *vol. 25, no. 3, pp. 585–619, 2020.*

[10] *R. F. Kizilcec and D. Goldfarb, "Growth mindset predicts student achievement and behavior in mobile learning," in* Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, *2019, pp. 1–10.*

[11] *L. Pappano. (Nov. 2012). The year of the MOOC. T. N. Y. Times, Ed., [Online]. Available:* https://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html.

to education, but they are slowly becoming the center of the learning experience.

With the recent pandemic caused by Corona Virus Disease 2019 (COVID-19), various computer-based instruction and collaboration software have gained attention as schools and universities are forced to lockdown. In addition, self-learning, to which computer-based instruction can be a considerable boost, has been promoted as a coping mechanism for extended periods of social distancing[13]. However, the transition from brick-and-mortar to entirely online or even blended learning has been fraught with challenges, with some students threatening lawsuits for tuition refunds[14]. The most salient problem is the lack of infrastructure (e.g., internet connection) from both the teacher side and the learner side; equally important is the lack of skillset also from both teachers and learners to survive in the online platform[15].

Unfortunately, current trends brought about by COVID-19 may persist for a while. Transitioning back to the old normal may be difficult or even no longer possible. Vaccine availability may still be a problem in some places, the virus may become endemic, or its eventual severity may remain unknown[16].

1.3 RESEARCH AGENDA

Japan itself is no stranger to the MOOC trend: the Japan Massive Online Course Consortium (JMOOC) was formed in 2013 and currently has four online education platforms (gacco, Fisdom, OpenLearning Japan, and OUJ MOOC). Additionally, several Japanese universities are developing online courses for the big global players in online education (Coursera, edX, and FutureLearn, among others). As such, MOOC development is a viable option for achieving Japan's Society 5.0 goals.

Unfortunately, high dropout numbers are common in MOOCs[17]. For instance, some learners are ill-prepared for the courses they signed up for, making it hard for them to catch up. Some are above the teacher's skill level during course creation, making the coursework unnecessarily tedious and disengaging. High dropout rates are not necessarily a problem for MOOCs as learner motivation might have been different from the start. If learners enroll in a course just to check it out – similar to a person shopping for a new fiction book – dropping out is not a serious issue. However, it can be problematic for higher education classes that shift to the MOOC format where learners might be forced to sign-up regardless of motive.

There are several ways to assist learners in online learning environments. One is to evaluate and design the online classroom based on general principles in human-computer interaction to make learning experience as smooth as possible[18]. This can include facilitating discussion and collaborations on online learning platforms where such

[12] *D. A. Joyner and C. Isbell, "Master's at scale: Five years in a scalable online graduate degree," in* Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, *2019, pp. 1–10.*

[13] *United Nations Educational, Scientific and Cultural Organization. (Jun. 2020). Education: From disruption to recovery, [Online]. Available:* https://en.unesco.org/covid19/educationresponse.

[14] *A. Kamenetz. (May 2020). Colleges faces student lawsuits seeking refunds after coronavirus closures. NPR, Ed., [Online]. Available:* https://www.npr.org/2020/05/29/863804342/colleges-face-student-lawsuits-seeking-refunds-after-coronavirus-closures.

[15] *J. Crawford, K. Butler-Henderson, J. Rudolph,* et al.*, "COVID-19: 20 countries' higher education intra-period digital pedagogy responses,"* Journal of Applied Learning & Teaching, *vol. 3, no. 1, pp. 1–20, 2020.*

[16] *W. Wan and C. Y. Johnson. (May 2020). Are we near to having a vaccine for Covid-19? W. Post, Ed., [Online]. Available:* https://www.washingtonpost.com/health/2020/05/27/coronavirus-endemic/.

activities may not be as intuitive. Another way is to design the instruction to develop the learners on several knowledge dimensions.

One of the most popular frameworks used for instructional design is Bloom's taxonomy for cognitive learning. In a recent update to Bloom's taxonomy, the old knowledge dimension was divided into factual, conceptual, and procedural dimensions[19]. In addition, the metacognitive dimension was added to the original knowledge dimension. As such, providing metacognitive instruction on top of the usual cognitive dimensions is an instructional design way to support learners in online learning environments.

There are two levers that teachers can use to improve learning outcomes if we are to draw from the self-regulated learning concept: motivation and metacognition. Time and time again, educators are pushed to aim at increasing learner motivation to, in return, increase learning outcomes. One way to increase learners' motivation, even for uninteresting tasks, is to promote the task's value to the learners[20]. However, the same lack of face-to-face interaction that results in high dropout rates in computer-based interaction can hinder the teacher in providing this source of motivation. It is thus essential for the learners to realize the value of learning on their own.

The overall lack of preparedness to succeed online warrants an alternative to MOOCs' typical instruction to ensure learning. Traditionally, tutoring, or a system where a subject-matter expert works with a single learner, has helped the learner cope with difficulties in a typical classroom setting[21]. Tutoring focuses more on individual learner needs than the learning outcomes of a larger group of learners. However, providing individualized attention can be difficult for teachers who handle several learners in traditional classrooms. This is even more so during the COVID-19 pandemic where administrative tasks related to ensuring learners' health and safety are piled on top of existing teaching tasks.

For learners to succeed in online learning environments, interventions centering on improving metacognition are essential. These include teachers providing training and support to learners as they develop their metacognitive skills, which can be challenging considering that teacher resources are not infinite. Computer tutoring systems, which were shown to have the potential to teach as effectively as humans, can help reduce the burden on teachers.

1.4 THEORETICAL FRAMEWORK

Researchers in the past have shown the effectiveness of using prompts in developing metacognitive skills[22]. Prompts, or triggers to actions, can be domain-independent. The actions triggered from the learners can be something that they can apply in other domains (e.g., breaking down an enormous task into smaller steps can be helpful both in

[17] *Y. Zheng, Z. Gao, Y. Wang,* et al., *"MOOC dropout prediction using FWTS-CNN Model based on fused feature weighting and time series,"* IEEE Access, *vol. 8, pp. 225 324–225 335, 2020.*

[18] *D. Joyner, "The CHI of teaching online: Blurring the lines between user interfaces and learner interfaces," in* Designing for the User Experience in Learning Systems, *Springer, 2018, pp. 81–102.*

[19] *L. W. Anderson, B. S. Bloom,* et al., A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. *Longman, 2001.*

[20] *H. Jang, "Supporting students' motivation, engagement, and learning during an uninteresting activity.,"* Journal of Educational Psychology, *vol. 100, no. 4, p. 798, 2008.*

[21] *K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems,"* Educational Psychologist, *vol. 46, no. 4, pp. 197–221, 2011.*

[22] *C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in* International Conference on Intelligent Tutoring Systems, *Springer, 2004, pp. 668–677.*

mathematics and English composition) as opposed to a specific domain (e.g., recalling a mnemonic that was provided by the teacher and not created by the learners themselves).

A metacognitive skill set is critical in learning environments where the teacher is not physically available to provide needed guidance, as in online learning. Even advanced learners can benefit from metacognitive skill development as it is a learning skill transferable across different subject domains. Using domain-independent prompts also has the additional benefit of the teacher being able to use the same tools in several other classes.

Another option is to personalize content presented to learners so that less prepared learners can cope while preventing advanced learners from being disengaged. This can be done through adaptive learning or educational technologies that monitor learner progress and modify instruction accordingly. This has the benefit of reducing the work required and not boring the learners who are ahead of the class, and giving more opportunities for learners who are behind to catch up. A straightforward approach to this is to provide diagnostic tests at the start of the class and allow learners to skip the modules they performed well in. Another common adaptive learning method used is providing exercises for learners depending on their topic mastery. Knowledge tracing algorithms are most commonly used in predicting the learner's mastery level[23].

Using knowledge tracing algorithms is not the only way to personalize learning. This can also be done by enabling the learners to provide feedback, thus allowing the teacher to adjust and give the learners the best learning experience. However, unlike in an in-person class where the learner can approach the teacher when needed or the teacher can gauge the learner's engagement through observation, online learners have fewer opportunities to provide feedback. They are limited to forum discussions or periodic course surveys. Research on MOOCs though has shown that only a minority of learners take advantage of forums[24]. The other common source of feedback, course surveys, on the other hand, are sparse; they are usually delivered only at the beginning and at the end of the course.

An interesting point is how metacognitive tutoring can work hand-in-hand with adaptive learning along with learner-provided feedback. For one, metacognitive tutoring will result in added workload, which could be better managed by adaptive learning. Also, as introducing metacognitive tutoring may lead to different learner interactions with the course material, it is possible to get feedback that may not be evident through the usual means.

Given that the above techniques (metacognitive tutoring, adaptive learning, and feedback) are effective in other learning platforms, we are interested in knowing how these translate to online learning platforms. How can we develop metacognitive skills among online learn-

[23] *M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," in* International Conference on Artificial Intelligence in Education, *Springer, 2013, pp. 171–180.*

[24] *E. Huang, H. Valdiviejas, and N. Bosch, "I'm sure! Automatic detection of metacognition in online course discussion forums," in* 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), *IEEE, 2019, pp. 1–7.*

ers while personalizing their experiences and getting relevant feedback that would otherwise be unheard of? We specifically aim to answer the following questions:

- Are open-response prompts effective in developing metacognitive skills on an online learning platform?
- Can we use innovative ways to improve knowledge tracing algorithms for adaptive learning?
- Can we find alternative sources of feedback to assess the effectiveness of an online class?

This research aims to work at the intersection of education accessibility and developing 21st-century skills. There are many risks, though, for the experiments to fail to achieve the expected results. Whether the results end up being positive, negative, or inconclusive, this research can still extend the current body of knowledge in the following ways:

- It reassesses the effectiveness of metacognitive prompting on an online learning environment with learners of vastly different demographics.
- It uses an affective measure (learner's assessment) and not just performance in modeling the knowledge tracing algorithms.
- It explores alternative sources of course quality feedback.
- It creates a direct link between metacognitive skills development, learning analytics, and adaptive learning.

1.5 METHODOLOGY OVERVIEW

This research was conducted by creating Personalized Online Adaptive Learning System (POALS). It is a tool that extends Open edX: a learning management system (LMS) being used by the Tokyo Institute of Technology (TokyoTech) for developing MOOCs and small private online courses (SPOCs). An overview of POALS is shown in Figure 2. POALS is a modified version of a previously existing metacognitive tutor. Modifications include the addition of an Adaptive Engine and an Analytics Dashboard. The Adaptive Engine uses metacognitive measurements for knowledge tracing. The Analytics Dashboard, on the other hand, uses natural language processing (NLP) techniques to extract latent learner feedback from various text supplied by learners.

The overall methodology used is design-based research: a common approach in learning sciences. In design-based research, solutions (or more commonly called interventions) are created to solve problems. These solutions are put to test to evaluate their effectiveness, which

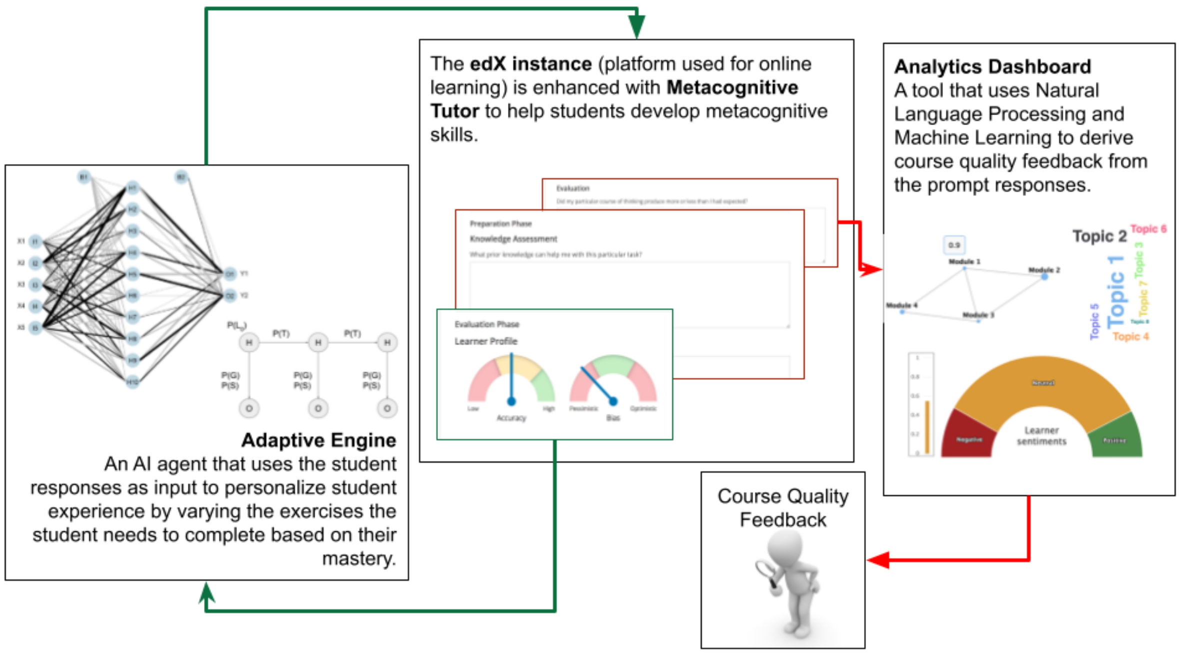


Figure 2: POALS Overview. POALS is made up of three major components: the Metacognitive Tutor, the Adaptive Engine, and the Analytics Dashboard.

then informs whether adjustments and retests are necessary. Design-based research is fit for learning sciences since learning environments are complex systems where not all possibilities might be evident to researchers at the onset. Using design-based research allows researchers to accommodate newly found phenomena not previously anticipated[25].

[25] A. L. Brown, *"Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings,"* The Journal of the Learning Sciences, *vol. 2, no. 2, pp. 141–178, 1992.*

A mixed-methods approach was used to assess the open-response prompts effectiveness in developing metacognitive skills. That is, the Metacognitive Tutor was evaluated from both quantitative and qualitative points of view. Quantitative results were derived from metacognition metrics POALS is tracking and from responses in closed response questionnaires administered at the start and the end of the experiments. The reliability of the selected questionnaire in our new context (different population and setting) was checked using Cronbach's alpha. Qualitative results were derived from the learner inputs on the open-response prompts displayed by POALS to understand the engagement levels of the learners with POALS. By looking at different measures for similar concepts, we enable our experiments to make their validity evident or even potentially arrive at divergent results that could lead to other important research questions not previously raised.

Machine learning research best practices were used in developing knowledge tracing algorithms. This includes preparing a test set to check the resulting models separate from the training set for model creation. Data distribution in both the training and test sets was examined for distribution similarity. Multifold cross-validations were done during training to increase the generalizability or applicability of the resulting models on scenarios not seen during training. Multiple algorithms, particularly interpretable and explainable algorithms, were compared against baseline values. The new models were also compared against previously investigated algorithms.

Wireframing and proof of concept development were done for exploring alternative feedback sources. The goal is to provide an interface for teachers to use their expertise in introducing interventions in their classes based on feedback that the learners may not provide straightforwardly. Various NLP approaches were used. Since POALS was used in a Japanese classroom, NLP techniques on Japanese, which are not as well studied as to their English counterparts, had been the primary engineering challenge.

1.6 OUTLINE

The outline of this document is shown in Figure 3.

This dissertation titled Personalized Online Adaptive Learning System describes a web-based system designed to help learners succeed in online learning environments. The learners must be trained to be autonomous by equipping them with metacognitive skills. Teaching metacognition inevitably introduces cognitive strain, which can vary among individuals. Thus, we introduce adaptive learning to personalize each learning experience. We tap into the teachers as learning facilitators by creating an analytics dashboard to give implicit feedback to teachers that they can use to provide interventions if necessary.

This chapter, Chapter 1 INTRODUCTION, provides a short historical overview to have a better understanding of views about an individual's learning and how these views have changed through time. In addition, this recall was used to discuss the motivation behind this research and introduce the methodologies used in answering the derived research questions.

Chapter 2 THE METACOGNITIVE TUTOR defines what metacognition is and describes how it can be taught and measured. Metacognition, which is essential to succeed in online learning environments, spans three distinct phases: planning, monitoring, and evaluating. An existing metacognitive tutor targeting knowledge of cognition and regulation of cognition at different phases previously shown to be effective in an experimental setting was considered. This was adapted to be more optimized and usable for online use, which is now POALS' Metacognitive Tutor. This chapter answers the question: are open-response prompts effective in developing metacognitive skills on an online learning platform? Indeed, POALS' Metacognitive Tutor was

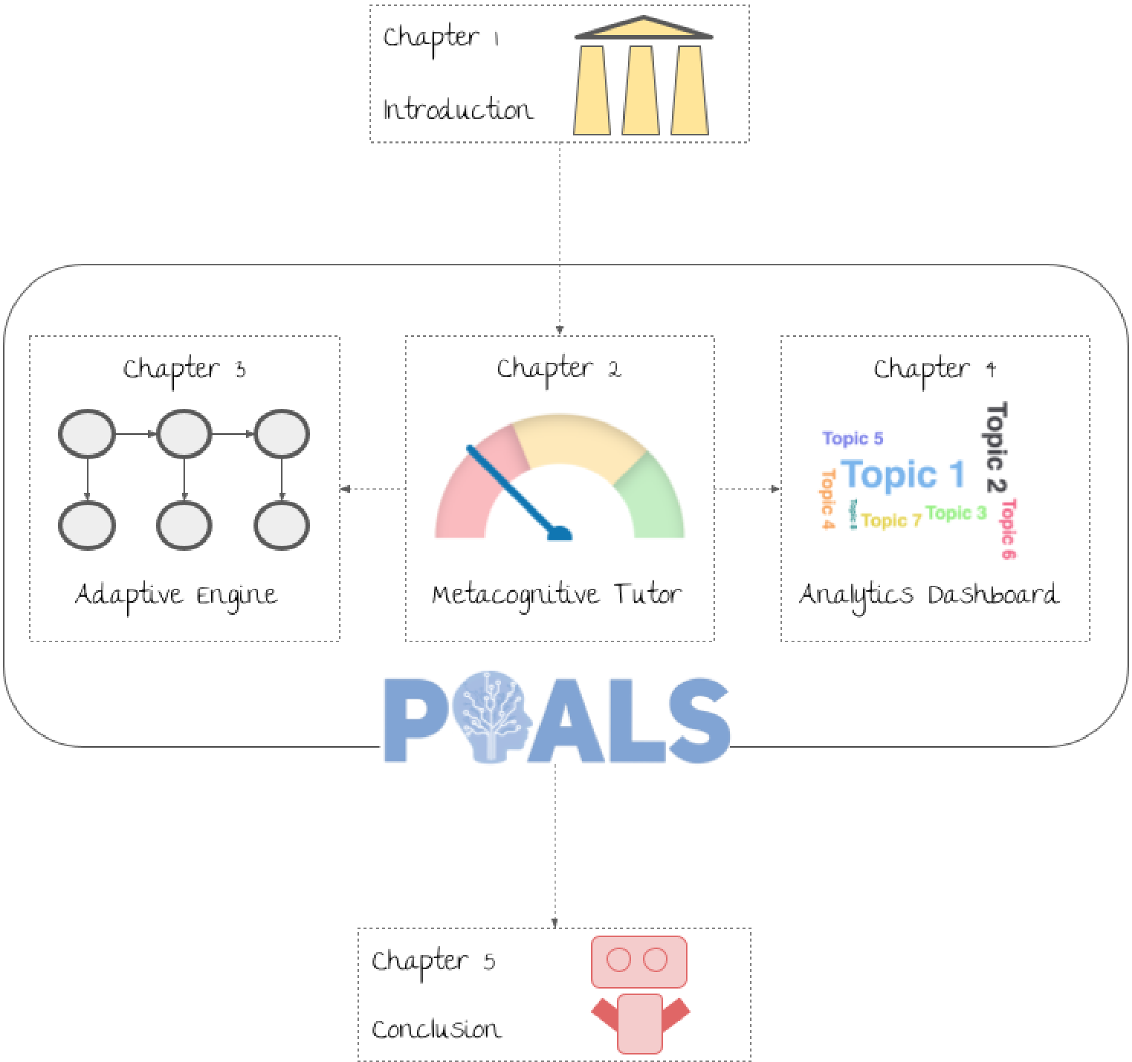


Figure 3: Dissertation Outline. Each POALS component has a dedicated chapter, preceded by this chapter (Introduction) and closed by a Conclusion chapter.

shown to be effective in improving learner metacognition to varying extents through a series of experiments. The tool is cognitive domain agnostic; thus, it can be a convenient means of tutoring metacognition in online learning environments.

Previous studies on metacognition show that metacognitive training on top of cognitive learning can strain learners' cognitive resources. Adaptive learning techniques such as knowledge tracing are an active research area for managing cognitive resources in online learning environments. Chapter 3 THE ADAPTIVE ENGINE investigates adaptive learning as a latent variable modeling problem that can be solved with machine learning. This chapter answers the question: can we use innovative ways to improve knowledge tracing algorithms for adaptive learning? Various algorithms were used to train models using a synthetic dataset created from predetermined learner personas. The models using metacognitive inputs performed better than the standard models while still following learning intuitions. This indicates that combining knowledge tracing and metacognitive tutoring is a viable option for improving learning outcomes. This serves as the backbone for POALS' Adaptive Engine.

Chapter 4 THE ANALYTICS DASHBOARD introduces POALS' Analytics Dashboard which serves as the teacher's window to their learners' implicit feedback. The proof-of-concept shows an aggregate of tools the teacher can use to understand learner sentiment, diagnose possible misconceptions, and check learning retention. Because the Analytics Dashboard utilizes the metacognitive prompt responses, problems with other sources of feedback (e.g., discussion forums participated by only a few, course surveys which are very sparse) can be resolved by providing a private and consistent channel between learners and teachers.

Important results, POALS' limitations, its potential societal impact, and possible future work are laid out in Chapter 5 CONCLUSION. This includes exploring how technology can make education more equitable and checking that algorithms intended to foster learning are fair. Educational technology hype trends from 2018 to present and its consequences to online learning environments are presented. Metacognition is also viewed as part of self-regulation, a concept that contributes to lifelong learning and an individual's growth. This opens future work extending the current study of technology-enhanced learning to the related areas of motivation, self-efficacy, learner behaviors, and performance.

2

THE METACOGNITIVE TUTOR

"Come, we shall have some fun now!" thought Alice. "I'm glad they've begun asking riddles. I believe I can guess that," she added aloud.
"Do you mean that you think you can find out the answer to it?" said the March Hare.
"Exactly so," said Alice.
"Then you should say what you mean," the March Hare went on.
"I do," Alice hastily replied; "at least–at least I mean what I say -– that's the same thing, you know."
"Not the same thing a bit!" said the Hatter. "You might just as well say that 'I see what I eat' is the same thing as 'I eat what I see'!"
"You might just as well say," added the March Hare, "that 'I like what I get' is the same thing as 'I get what I like'!"
"You might just as well say," added the Dormouse, who seemed to be talking in his sleep, "that 'I breathe when I sleep' is the same thing as 'I sleep when I breathe'!"

–Lewis Caroll, *Alice in Wonderland*

In common use, the prefix **meta** is added to a word to make the new word self-referential. To illustrate, **metadata** is *data about data*: information like who created the data, how big is the data, what is the data format, and so forth. Another example is **metamorphosis**, or the biological process that accounts for the changes (read: *morphosis*) that occur as a living creature changes (e.g., from egg to larva to pupa). **Metacognition**, therefore, is *cognition about cognition*: or put simply, thinking about thinking.

Then again, what does it mean to think or have cognition? In cognitive psychology, cognition is typically assumed to be information processing in the brain that can be extended to its environment[26]. When thinking exclusively of the brain, this processing can include various functions such as producing and understanding language, reasoning, decision-making, problem-solving, applying knowledge and paying attention. When extended to the environment, this includes being aware of tools that can be used for information processing (e.g., navigating with a map when driving instead of memorizing the path before starting to travel). It is important to note that cognition is a *process* that an individual continuously bears in mind.

[26] *O. Blomberg, "Conceptions of cognition for cognitive engineering,"* The International Journal of Aviation Psychology, *vol. 21, no. 1, pp. 85–104, 2011.*

Metacognition encompasses several skills, such as goal setting and knowledge monitoring, among others. A related concept to metacognition is self-regulation. This is the learners' ability to take control of their learning by tapping not just metacognitive processes but also motivational processes[27]. Both metacognition and self-regulation

[27] *B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview,"* Educational Psychologist, *vol. 25, no. 1, pp. 3–17, 1990.*

were shown to contribute to learners' academic performance regardless of their intelligence and age group[28].

The most straightforward definition of metacognition is in the ability to reflect and adapt accordingly. Autonomous learning, or learning independently, has always been associated with metacognition[29]. Recently, lifelong learning is becoming more critical as the workplace changes due to the proliferation of robotics and automation[30]. Metacognition is also seen as a unique human quality that allows us to deal with modern life (e.g., the exponential growth of knowledge and technological advances, among others) through reflection and adaptation[31]. As such, metacognition is typically seen as integral to 21st-century education.

Metacognition is not unique to humans: it is exhibited by some animals[32] and deliberately implemented into artificial systems[33]. Nevertheless, these abilities in non-humans are still seen as limited by human standards. Metacognition is an integral part of the development of the human mind[34], lending it to be utilized in non-binary decisions ("to eat or not to eat" versus "I will read more about cyberspace") that non-humans are not as capable of. Thus, metacognition is an important consideration in developing technologies, especially those targeted in developing the human mind.

Recently, online learning platforms are becoming more popular partly due to massive open online courses (MOOCs). Online learning platforms give the learners the advantage of learning at their own pace, so they do not have the same restrictions as regular classrooms. For instance, while learners might have limited chances to digest the information provided by teachers in a face-to-face class, they can take pauses as necessary in online learning platforms. Furthermore, because learners can dedicate more time to learning through online learning platforms, they are ideal venues for metacognitive instruction. With online learning platforms, they can allocate more time to learn both cognitive and metacognitive components at their own pace.

2.1 WHAT IS METACOGNITION?

Portions published as: M. K. J. Carlon and J. S. Cross, "A review of quantitative offline measurement tools for computer-based metacognitive tutoring effectiveness assessment," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Dec. 2020, pp. 258–264. DOI: 10.1109/TALE48869.2020.9368470

2.1.1 *An Overview of Metacognition Theory*

Metacognition involves knowing how much one knows about a specific topic, regulating how one learns, and making adjustments to im-

[28] *K. Ohtani and T. Hisasaka, "Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance,"* Metacognition and Learning, *vol. 13, no. 2, pp. 179–212, 2018.*

[29] *M. Victori and W. Lockhart, "Enhancing metacognition in self-directed language learning,"* System, *vol. 23, no. 2, pp. 223–234, 1995.*

[30] *C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?"* Technological Forecasting and Social Change, *vol. 114, pp. 254–280, 2017.*

[31] *D. Sale, "Metacognitive capability: The superordinate competence for the twenty-first century," in* Creative Teachers, *Springer, 2020, pp. 77–129.*

[32] *N. Kornell, "Metacognition in humans and animals,"* Current Directions in Psychological Science, *vol. 18, no. 1, pp. 11–15, 2009.*

[33] *A. Sloman,* Varieties of metacognition in natural and artificial systems, *2011.*

prove learning. By definition, people with high metacognitive ability can plan how they will study, reflect on their learning progress, and have different strategies to help their learning. Thus, if a person has high metacognitive ability, they can succeed in school and educate themselves. This also applies to adults who are engaged in lifelong learning.

Empirical studies have shown that learners with high metacognitive ability have performance-enhancing behavioral characteristics. For instance, in a particular study that uses a metacognitive tutor that requires reflection activities, the learners exposed to the metacognitive tutor were found to be less likely to give up on solving complex math problems[35]. If they do complete answering challenging math problems, they tend to give more accurate results. This means that metacognition has benefits for enhancing one's thinking and attitude as well.

A concrete definition for metacognition is yet to be established. This is due to the different foci of researchers working on the said topic[36]. One of the most popular definitions is dividing the taxonomy of metacognition to the knowledge of cognition, regulation of cognition, and other metacognition[37]. Figure 4 summarizes this definition of metacognition.

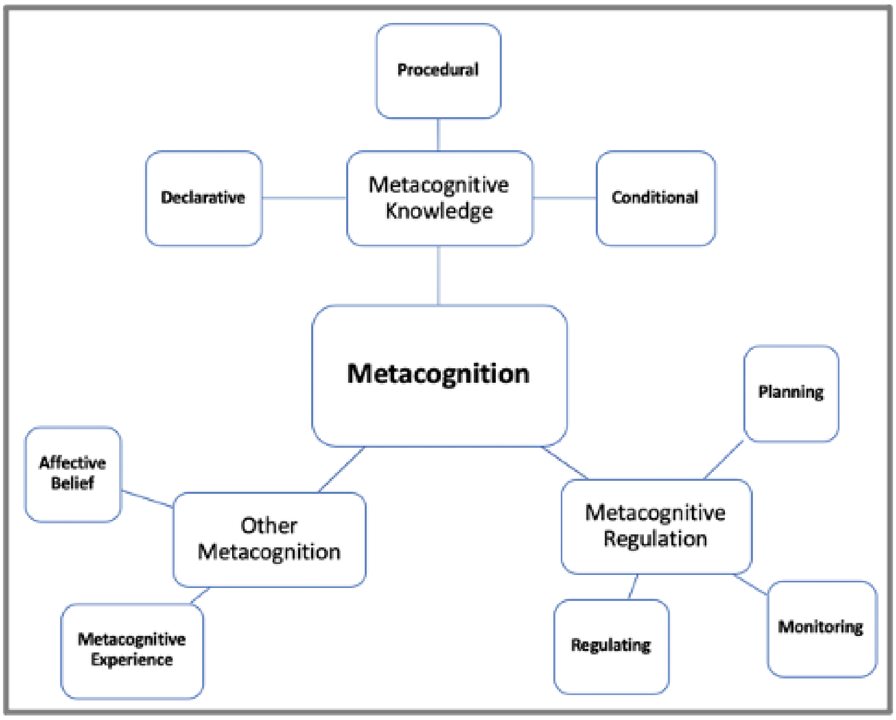


Figure 4: Metacognition Overview. Metacognition is a multi-faceted concept.

Knowledge of cognition, or metacognitive knowledge, is an individual's awareness of their knowledge levels (e.g., to what extent do

[34] A. Demetriou, A. Efklides, M. Platsidou, et al., *"The architecture and dynamics of developing mind: Experiential structuralism as a frame for unifying cognitive developmental theories,"* Monographs of the Society for Research in Child Development, *pp. i–202, 1993.*

[35] *C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in* International Conference on Intelligent Tutoring Systems, *Springer, 2004, pp. 668–677.*

[36] *R. Azevedo, "Reflections on the field of metacognition: Issues, challenges, and opportunities,"* Metacognition and Learning, *Jun. 2020.*

[37] *P. Tarricone,* The taxonomy of metacognition. *Psychology Press, 2011.*

they understand a topic). Metacognitive knowledge can further be classified as either declarative, procedural, or conditional. Declarative knowledge refers to knowledge about oneself (e.g., what they currently know) and factors that may affect their performance (e.g., difficulty of the task at hand). On the other hand, procedural knowledge refers to knowing how to execute (e.g., strategies). Finally, conditional knowledge refers to knowing when and why should declarative and procedural knowledge should be used. For instance, acknowledging that a help-seeking strategy (procedural knowledge) can be used when you do not have sufficient information to complete a task (declarative knowledge) is an example of conditional knowledge.

Regulation of cognition, or metacognitive regulation, is ones' ability to take control of their learning. Like metacognitive knowledge, regulation of cognition can be divided into several components: planning, monitoring, and evaluating. Planning can be tantamount to goal-setting, where one sets a target and a series of actions to achieve the set target. Monitoring is keeping track of ones' progress. Finally, evaluating is reflecting on the task that was performed.

Other metacognition refers to all other metacognitive components: metacognitive experience, affective beliefs, and social metacognition. Metacognitive experience refers to experiences relating to cognitive endeavors (e.g., having the chance to create their mnemonic to help them remember long strings of information). Affective beliefs relate to other beliefs connected to metacognition (e.g., self-efficacy or belief in one's ability to complete a task as viewed from a metacognitive perspective). Finally, social metacognition refers to an individual's awareness of others' mental processes and the corresponding effect on their personal beliefs (e.g., hearing someone you know smarter than you doubt their knowledge of their topic may cast doubts on your knowledge). Hence, while metacognition is a very personal quality, it can also be affected by social dynamics.

Aside from intrapersonal and interpersonal qualities, metacognitive levels were also seen to be related with developmental stages, where adults exhibit higher levels of metacognitive abilities compared to children and those still in school[38]. Likewise, not all tasks require the same amount of metacognitive engagement, or even the same set of metacognitive skills. The more information is associated with a task, the more metacognitively involved it will be[39]. This is explained by dual-process theory, which divides tasks into two types: Type 1 - Automaticity and Type 2 - Cognitive Decoupling. Type 2 activities requiring higher order thinking skills will require more metacognition than Type 1 activities. For example, health care personnel need to have automaticity for scenarios requiring fast decisions such as in life-threatening emergencies where slower, more deliberate thinking brought about by metacognition can be debilitating[40]. On the other

[38] *K. Ohtani and T. Hisasaka, "Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance,"* Metacognition and Learning, *vol. 13, no. 2, pp. 179–212, 2018.*

[39] *L. Vangsness and M. E. Young, "More isn't always better: When metacognitive prompts are misleading,"* Metacognition and Learning, *pp. 1–22, 2020.*

[40] *K. S. Chew, S. J. Durning, and J. J. Van Merriënboer, "Teaching metacognition in clinical decision-making using a novel mnemonic checklist: An exploratory study,"* Singapore Medical Journal, *vol. 57, no. 12, p. 694, 2016.*

hand, they need to take more time for situations calling for cognitive decoupling where weighing all possible alternatives is necessary (e.g., deciding elective surgical procedures).

Finally, there are cases where metacognition can be seriously detrimental. Negative metacognitive beliefs happen when one's reflection leads them to believe that their line of thought could cause them danger ("All my decisions had been wrong") instead of bringing positive result ("Worry is a good sign: I can be prepared.")[41]. Research has shown that negative metacognitive beliefs can lead to longer times dwelling in metacognitive activities (i.e., decreased metacognitive efficiency) which can potentially be indicative of depression[42].

[41] *R. Sellers, A. Wells, and A. P. Morrison, "An experimental manipulation of negative metacognitive beliefs in non-clinical paranoia: Effects on intrusions and state anxiety,"* Journal of Experimental Psychopathology, *vol. 9, no. 3, jep–062 117, 2018.*

[42] *C. Papageorgiou and A. Wells, "An empirical test of a clinical metacognitive model of rumination and depression,"* Cognitive therapy and research, *vol. 27, no. 3, pp. 261–273, 2003.*

2.1.2 *Measuring Metacognition*

Metacognition can be measured from at least two perspectives: *online or offline* measurements and *quantitative or qualitative* methods. These methods can further be combined; that is, metacognition may be measured by a combination of offline and online measurements and quantitative and qualitative methods. This concept is not limited to metacognition and applies to other constructs (e.g., motivation), especially in social sciences. Each of these methods has its advantages and disadvantages.

In the context of learning analytics, online measurement measures the target construct – in our case, metacognition – while the learners are undergoing a learning activity. Examples include computer logs gathered while the learner is interfacing with a computer-based instruction or think-aloud protocols where learners describe their thinking process to an observer while undergoing a learning activity. Because the online measurement is done while the learner uses the construct, it is usually deemed more accurate than offline measurements. However, online measurements may fail to account for the complexity of the system measured in (e.g., some relevant interactions may not be logged). They may also lead to poor explainability as it is hard to separate possible distractions (e.g., tool malfunctions) from the learning activity[43].

[43] *D. Tempelaar, B. Rienties, and Q. Nguyen, "Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application,"* PLOS One, *vol. 15, no. 6, e0233977, 2020.*

Offline measurement, on the other hand, is measuring the target construct outside the learning activity. This is usually done before and after a learning intervention to see if the intervention affected the target construct. Offline measurements typically are self-reports that come in the form of questionnaires or interviews. Offline measurements are often cheaper than online measurements since they do not require a complicated setup to collect logs or tedious administration, as with think-alouds. However, being self-reports, offline measurements are prone to recall problems, prompting effects, individual reference points, and other social biases.

Quantitative methods involve the use of numbers to test or confirm a hypothesis. These can be in the form of questionnaires or usage logs. These are usually seen as more objective than qualitative methods and allow for generalizable results if the research is designed to account for causality. However, since most quantitative data stem from closed responses from subjects, it may be challenging to derive deeper insights[44].

[44] S. Baškarada and A. Koronios, "A philosophical discussion of qualitative, quantitative, and mixed methods research in social science," Qualitative Research Journal, 2018.

Qualitative methods involve the collection of thoughts and observations followed by summarizing, categorizing, and interpreting. Examples of this include interviews and think-alouds. Qualitative methods enable researchers to develop an in-depth understanding of a phenomenon since subjects are typically given a chance to elaborate on their experiences. However, qualitative methods may not be replicable; most of the data gathered and the way they are interpreted can be subjective especially if the protocol used lacks rigor.

It is best to combine online measurement with offline measurement and quantitative methods with qualitative methods. This mixed approach can enable a researcher to address the gap of each measurement and method. However, having all these measurements and methods together in a research project may not always be feasible; hence a researcher or teacher may be compelled to choose the best option or options available to them. Arguably, the quantitative offline method is the cheapest approach as quantitative is easier to interpret and offline is quick to administer. However, quantitative offline measurement is only cheap when the tool to be used was already previously vouched. Furthermore, creating self-report tools requires testing for validity and reliability[45].

[45] C. Demetriou, B. U. Ozer, and C. A. Essau, "Self-report questionnaires," The Encyclopedia of Clinical Psychology, pp. 1–6, 2014.

In this research, we used a pre-validated questionnaire delivered before and after the experiments as offline measurement. This was complemented by an online measurement called the Learner Profile that was implemented into Personalized Online Adaptive Learning System (POALS) Metacognitive Tutor. Both the questionnaire and the Learner Profile are quantitative. For qualitative analysis, we looked into the responses provided by the learners in POALS Metacognitive Tutor's open response prompts.

2.2 DOMAIN-AGNOSTIC METACOGNITIVE TUTORS

Portions submitted as: M. K. J. Carlon and J. S. Cross, "Development of open-response prompt-based metacognitive tutor for online classrooms," in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Submitted

We can see here that since metacognition operates at a higher level than cognition, metacognition does not have to be tied to a particular domain. For example, suppose a learner can detect that they are having difficulty understanding a math lesson. In that case, it is not hard to imagine that the same learner would recognize if they also have difficulty understanding their English lessons. Nevertheless, domain knowledge may still matter. For example, a metacognitive learner may honestly believe that they understood a physics lab lesson, but only because they do not have sufficient procedural knowledge to detect that they are not knowledgeable in proper lab equipment usage. That brings us to another quirk about metacognition: it is not a single knowledge or skill. Instead, it can manifest itself through different means such as good reflection on one's skills or managing their learning resources, just to name two.

Several studies have created tools for developing metacognition tightly coupled with their respective domains. Some examples include works in engineering[46], physical education[47], language learning[48], nursing[49], and teacher development[50], among others. These approaches may not be easily portable to any other domain they were not previously investigated on. Thus, these domain-specific approaches can be costly to deploy.

Since metacognition is domain-independent in the first place, having domain-independent metacognitive development tools is plausible. To build context despite not targeting a cognitive domain, several existing domain-independent tools are targeted at specific metacognitive skills instead. Some examples include those tackling with goal setting[51], control[52], awareness[53], and help-seeking[54], among others. In this research, we are looking at a metacognitive model that is domain-independent and, at the same time, targets a more holistic metacognitive development.

2.2.1 *The Reflection Assistant*

The Reflection Assistant (RA) is a generic metacognition model designed to explore metacognitive instruction on problem-solving interactive learning environments[55]. Its instructional framework is based on the hierarchical model of metacognition. In this model, metacognition is seen as a group of skills including planning, selecting strategies, evaluating learning, knowledge monitoring, and controlling[56]. In particular, RA focused on developing selecting strategies and evaluating learning metacognitive skills.

A critical concept in instructional design is cognitive load or the amount of working memory resources needed to complete a task[57]. Cognitive load can be classified as intrinsic, extraneous, and germane. Intrinsic cognitive load refers to the inherent difficulty of the task at hand. On the other hand, extraneous cognitive load refers to the effort required to process the information but may not be necessary to complete a task. For example, a mathematics problem can be more difficult by being poorly worded. Likewise, ineffective presentation adds extraneous cognitive load to the intrinsic cognitive load needed

[46] S.-H. Chang, M.-L. Chen, Y.-K. Kuo, et al., *"A simulation-based LED design project in photonics instruction based on industry–university collaboration,"* IEEE Transactions on Education, *vol. 54, no. 4, pp. 582–589, 2011.*

[47] *L. Cid, A. Pires, C. Borrego*, et al., *"Motivational determinants of physical education grades and the intention to practice sport in the future,"* PLoS One, *vol. 14, no. 5, e0217218, 2019.*

[48] *M. Hariri-Akbari, B. Shokrvash, F. Mahmoodi*, et al., *"Conversion of extrinsic into intrinsic motivation and computer based testing (CBT),"* BMC Medical Education, *vol. 18, no. 1, pp. 1–8, 2018.*

[49] *L.-L. Hsu and S.-I. Hsieh, "Factors affecting metacognition of undergraduate nursing students in a blended learning environment,"* International Journal of Nursing Practice, *vol. 20, no. 3, pp. 233–241, 2014.*

[50] *P. Virtanen, H. M. Niemi, and A. Nevgi, "Active learning and self-regulation enhance student teachers' professional competences,"* Australian Journal of Teacher Education, *vol. 42, no. 12, p. 1, 2017.*

to solve the problem. Finally, germane load refers to processing patterns of thoughts that can be used to support learning. While the intrinsic cognitive load is always thought to be immutable, instructional designers can manipulate extraneous (e.g., by making problems more understandable) and germane (e.g., by introducing steps learners can follow) cognitive loads to support better learning.

In its implementation, RA took into account the cognitive load that the learner might experience during metacognitive instruction combined with domain instruction by looking at the conceptual stages of problem-solving. The conceptual stages of problem-solving suppose that problem-solving occurs in three stages: the preparation phase, the problem-solving phase, and the verification evaluation phase[58]. Learners experience the highest intrinsic cognitive load during the problem-solving phase, while the preparation and evaluation phases can be promising avenues for tapping into the germane cognitive load. RA does its metacognitive instruction during the preparation and evaluation phases to allow the learners to fully dedicate their intrinsic cognitive resources to cognitive learning during the problem-solving phase. Structures are introduced in the preparation and evaluation phases to use germane cognitive load in reinforcing learning.

Figure 5 shows an overview of the metacognitive instruction in RA. The entire instruction can be seen as a cycle of six steps, where each step is a separate screen in the intelligent tutoring system (ITS). During the preparation phase, the learners are asked to reflect on their previous performance, read the upcoming problem to be solved, assess the problem's difficulty, and plan the strategy to be used during problem-solving. During the evaluation phase, the learners check the teacher's solution and reflect on their problem-solving process.

To enable the learners to reflect on their metacognitive performance, RA tracks the learners' Knowledge Monitoring Accuracy (KMA) which provides a measure of the learners' awareness of their knowledge[59]. Before selecting the metacognitive strategies they will use, the learners are asked to predict whether they will be able to answer each question correctly (C), partially correctly (P), or incorrectly (I). The intermediate measures Fully Correct (FC), Partially Correct (PC), and Fully Incorrect (FI) are updated for every problem solved based on the learners' difficulty assessment and actual problem-solving performance as summarized in Table 1.

The KMA is computed in Equation 1 using the cumulative intermediate measures. The resulting score ranges from -1 to 1, where the score is better when higher. The learners are classified to have low accuracy when their KMA is below -0.25, high when their KMA is 0.5 and above, and average otherwise. The original researchers set these values according to their intuition. These can arguably be fur-

[51] *E. Chang and S. M. Lee, "Mediating effect of goal adjustment on the relationship between socially prescribed perfectionism and academic burnout,"* Psychology in the Schools, *vol. 57, no. 2, pp. 284–295, 2020.*

[52] *P. D. Converse, E. Steinhauser, and J. Pathak, "Individual differences in reactions to goal-performance discrepancies over time,"* Personality and Individual Differences, *vol. 48, no. 2, pp. 138–143, 2010.*

[53] *B. A. Fernie, U. Y. Kopar, P. L. Fisher,* et al., *"Further development and testing of the metacognitive model of procrastination: Self-reported academic performance,"* Journal of Affective Disorders, *vol. 240, pp. 1–5, 2018.*

[54] *S. Won, L. C. Hensley, and C. A. Wolters, "Brief research report: Sense of belonging and academic help-seeking as self-regulated learning,"* The Journal of Experimental Education, *vol. 89, no. 1, pp. 112–124, 2021.*

[55] *C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in* International Conference on Intelligent Tutoring Systems, *Springer, 2004, pp. 668–677.*

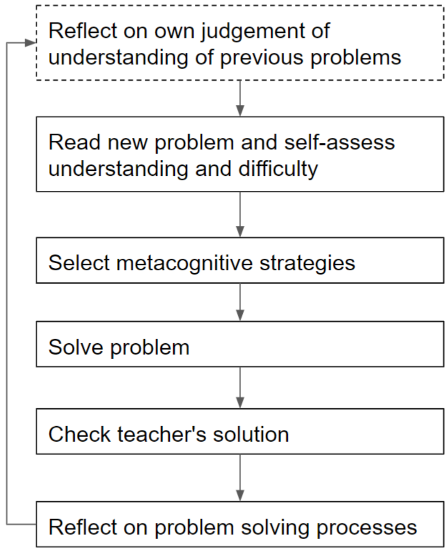


Figure 5: RA Tutoring Workflow. The RA tutoring workflow is a cycle made up of six steps.

Table 1: KMA Increment Matrix. The FC, PC, and FI values are incremented after each exercise using the following judgment matrix.

Score	Confidence Self-Report		
	C	P	I
Correct	FC	PC	FI
Partially Correct	PC	FC	PC
Wrong	FI	PC	FC

ther validated, but we are directly adapting the values originally set for simplicity.

$$\text{KMA} = \frac{\text{FC} - 0.5 * \text{PC} - \text{FI}}{\text{FC} + \text{PC} + \text{FI}} \tag{1}$$

However, the KMA alone does not give us a hint on the sentiment of the learners regarding their metacognitive skills. The Knowledge Monitoring Bias (KMB) was added to the RA model to show whether the learners have an optimistic, pessimistic, or unbiased view of their skills. Just like the KMA, the KMB has intermediate measures that are incremented after each problem-solving based on learners' problem difficulty prediction and actual problem-solving performance. The measures No Bias (NB), Partially Optimistic Bias (POB), Fully Optimistic Bias (FOB), Partial Pessimistic Bias (PPB), and Fully Pessimistic Bias (FPB) are updated based on Table 2.

Table 2: KMB Increment Matrix. The NB, POB, PPB, FOB, and FPB values are incremented after each exercise using the following judgment matrix.

Performance	Confidence Self-Report		
	C	P	I
Correct	NB	PPB	FPB
Partially Correct	POB	NB	PPB
Wrong	FOB	POB	NB

The cumulative measures are then used to compute the KMB with Equation 2. Like KMA, the KMB values range from -1 to 1. This time, higher scores do not necessarily mean better outcomes. The closer the KMB to 0 is, the better is the score. When learners have low accuracy, they are classified to be pessimistic when their KMB is below -0.25, optimistic when their KMB is at least 0.25, or random in their judgments otherwise.

$$\text{KMB} = \frac{\text{FOB} + 0.5 * (\text{POB} - \text{PPB}) - \text{FPB}}{\text{FOB} + \text{POB} + \text{NB} + \text{PPB} + \text{FPB}} \tag{2}$$

As such, RA teaches selecting strategies by requiring the learners to plan their problem-solving strategy during the preparation phase and evaluating learning by prompting the learners to reflect on their problem-solving process during the evaluation phase. They are also trained on knowledge monitoring by showing the learners their KMA and KMB scores during the preparation and evaluation phases. The KMA and the KMB are collectively called the Learner Profile. Table 3 shows the learner classification applied by RA according to the

[56] *S. Tobias and H. T. Everson, "Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring. Research Report No. 2002-3,"* College Entrance Examination Board, *2002.*

[57] *J. Sweller, J. J. van Merriënboer, and F. Paas, "Cognitive architecture and instructional design: 20 years later,"* Educational Psychology Review, *vol. 31, no. 2, pp. 261–292, 2019.*

[58] *D. F. Halpern,* Thought and knowledge: An introduction to critical thinking. *Psychology Press, 2013.*

[59] *S. Tobias and H. T. Everson, "Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring. Research Report No. 2002-3,"* College Entrance Examination Board, *2002.*

Learner Profile values. The ideal score range is between 0.5 to 1 for KMA and between -0.25 to 0.25 for KMB. For both measures, the ideal ranges are constrained to just a quarter of the possible ranges. Thus, only the learners with consistently accurate behaviors are put in the optimal classification.

Table 3: Learner Profile Classification. The learners are classified based on their KMA and KMB scores.

Score Range	Classification	
	KMA	KMB
[-1, -0.25)	Low	Pessimistic
[-0.25, 0.25)	Average	Random or Ideal
[0.25, 0.5)	Average	Optimistic
[0.5, 1]	High	Optimistic

RA was tested as part of an interactive learning environment for an undergraduate algebra class. The empirical study was done in three one-hour sessions participated in by 25 undergraduate students. The study had positive results: those who used RA had more correct answers per total problems attempted showing better problem-solving performance. They also attempted a smaller number of problems given the same duration, which is indicative of their persistence in solving the problems (i.e., not giving up quickly) and their effort to tackle the problem while using metacognitive skills.

2.2.2 *Reflecting on RA for MOOC Use*

Considerable research works on MOOCs have investigated a few aspects of RA. For instance, providing self-regulated learning suggestions was found to be perceptually helpful[60]. A shortcoming, though, is that suggestion prompting does not necessarily increase learner performance. A key factor here is that the learners are only given suggestions, and taking action was left up to them. Systems that require learners to explain themselves[61] or self-reflect critically[62] on at-scale platforms are also already existing. However, these are tightly coupled to a subject matter (i.e., not domain-independent). A work that can be used in a domain-independent setting is on allowing learners to self-evaluate their work based on a rubric[63]. Still, the focus of the said research is the scalability of evaluating complex projects instead of effectiveness in developing metacognition.

As we have already underscored, metacognitive tutoring can be challenging, but it can help learners succeed in today's online learning environments. Despite the problems experienced with the sud-

[60] *R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado, "Recommending self-regulated learning strategies does not improve performance in a MOOC," in* Proceedings of the third (2016) ACM Conference on Learning@ Scale, *2016, pp. 101–104.*

[61] *E. Farrow and J. D. Moore, "Beetle-Grow: An effective intelligent tutoring system to support conceptual change," in* Proceedings of the Third (2016) ACM Conference on Learning@ Scale, *2016, pp. 331–332.*

[62] *P. Ortiz and D. F. Harrell, "Chimeria: Grayscale MOOC: Towards critical self-reflection at scale," in* Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, *2019, pp. 1–4.*

[63] *J. Wilkowski, D. M. Russell, and A. Deutsch, "Self-evaluation in advanced power searching and mapping with Google MOOCs," in* Proceedings of the first ACM Conference on Learning@ Scale, *2014, pp. 109–116.*

den switch to online learning due to the COVID-19 pandemic, online learning demand can only be expected to continue for the near future given the modern challenges we face[64]. These include the anticipated shift to Society 5.0 (living in an environment where cyberspace merges with the physical space or the so-called knowledge-based economy)[65] and the pandemic situations' lingering uncertainty.

[64] M. K. J. Carlon and J. S. Cross, "A review of quantitative offline measurement tools for computer-based metacognitive tutoring effectiveness assessment," in 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), IEEE, Dec. 2020, pp. 258–264. DOI: 10.1109/TALE48869.2020.9368470.

[65] Japan Cabinet Office. (nd). Society 5.0, [Online]. Available: https://www8.cao.go.jp/cstp/english/society5_0/index.html.

When the original empirical study using RA is viewed with current trends, a few problems have been identified. First and foremost, the study was conducted in a laboratory setting, and thus its effect on actual classroom settings has not been verified. Moreover, the study measured the performance improvement when using RA but did not measure metacognitive development. Additionally, the way learners interact with digital instruction may have changed significantly with the rising popularity of online learning, computerization, and other learning technologies a decade since the RA research was conducted.

From a usability perspective, the original RA was split into six different screens: one for each cyclical step. When the same format is applied to learning management system (LMS) plugins, which typically are viewable as web applications on browsers, having too many screens may affect usability negatively. The original format may require the learners to move across multiple screens, and doing so on a web platform may introduce lag times.

2.2.3 *Personalized Online Adaptive Learning System (POALS) – Metacognitive Tutor*

We first analyzed how the RA model may be used in an online learning environment, typically those using LMS which is commonly used for MOOCs or small private online courses (SPOCs). We then created POALS which includes a metacognitive tutor that is intended to replicate the RA model conceptually. Figure 6 shows the differences between the original RA model and POALS Metacognitive Tutor.

The three problem-solving stages were made more well defined in POALS Metacognitive Tutor. Each problem is presented to the learners in three sub-screens, and each of the sub-screens corresponds to a problem-solving stage. A sub-screen is not displayed until the learners reach the corresponding problem-solving stage. Once a sub-screen is displayed, learners can scroll back to previous sub-screens to check their previous inputs if they need to. The look-and-feel is similar to Open edX, the LMS created by the Massachusetts Institute of Technology and Harvard University currently being used for edX, one of the leading MOOC providers. Note that Tokyo Institute of Technologys (TokyoTechs) provides MOOCs on edX and uses edX Edge (an implementation of Open edX) for its SPOCs.

Figure 7 shows the sub-screen for the preparation phase. The learners are shown the problem overview to help them reflect during the

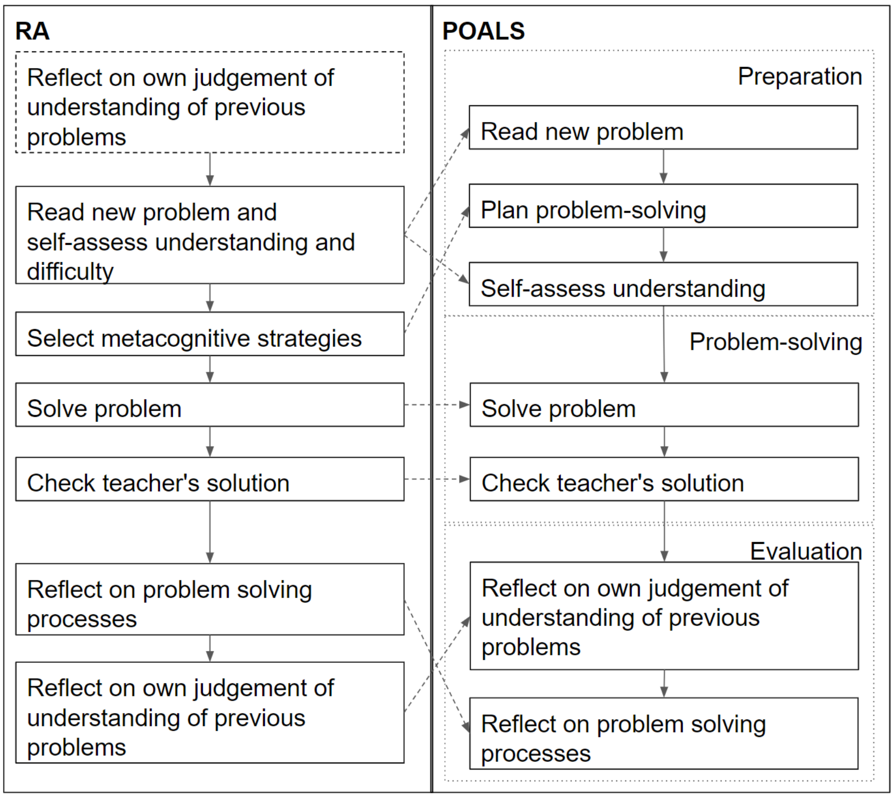


Figure 6: RA and POALS Metacognitive Tutor Comparison. The steps for RA (left) were grouped into Halpern's problem-solving phases in POALS (right).

preparation phase. Originally, RA required the learners to reflect on previous problems' performance to trigger reflection earlier on. However, we think this might be counterproductive as it may preempt the learners from exploring multiple possibilities. Learners who performed well in the past might be compelled to use the strategies that have worked for them previously, even though those might no longer apply to the current problem. Inversely, learners who did not perform well in the past may be quick to dismiss certain strategies without deeper thought.

Instead of immediately showing the Learner Profile, we used open-response prompts for the preparation phase for the learners to reflect on their preparedness to solve the problems. The open-response prompt replaces the old step – selecting metacognitive strategies – which is not ideal from maintainability perspective. What could be suitable metacognitive strategies may differ across cognitive domains or even across tasks within a cognitive domain. Thus, having the previous selecting metacognitive strategies step would require modifications for each course or even each activity where POALS Metacognitive Tutor is used.

Finally, the learners are asked to assess their capability to answer the problem after the other reflection activities for the preparation phase. This was placed at the end of the preparation phase again to prevent learners from building premature judgments. Should the learners decide that they are not ready to answer the problem, they

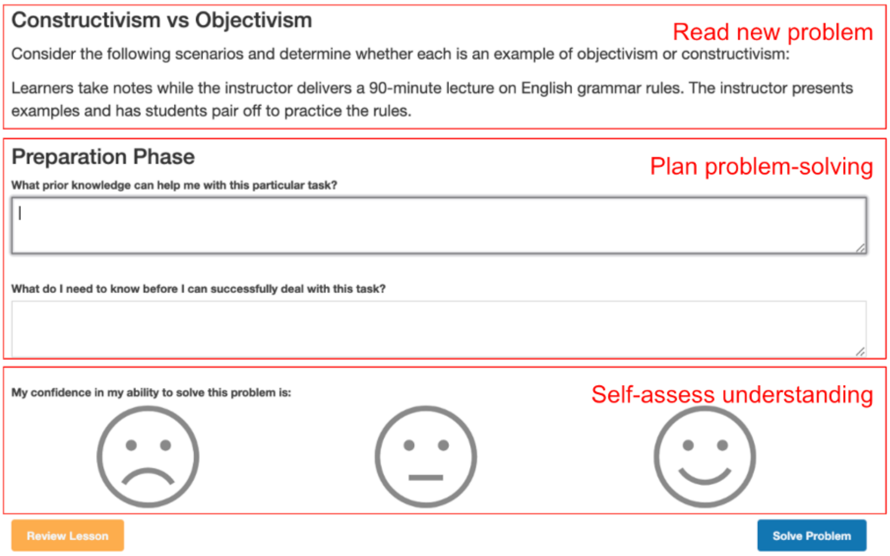


Figure 7: Preparation Phase. Learners are shown an overview of the problem, then are asked to plan for problem-solving and evaluate their readiness.

can choose to save their reflection inputs and review their lessons instead of pushing through the problem-solving. This was allowed since POALS is created for formative assessment where problems are used to teach the learners, not for summative assessment where problems are used for grading.

Figure 8 shows the sub-screen for the problem-solving stage. This sub-screen is displayed when the learner chooses to solve the problem from the preparation sub-screen. The correct answer to the problem is displayed once the Submit button is clicked. The learner can also choose to check the teacher's solution without submitting their answer by clicking the Show Answer button. Since POALS is for formative assessment, we also allow the learners to save their answers and go back to the lesson's material to review by clicking the Save button. Clicking buttons other than Submit button does not affect the Learner Profile.

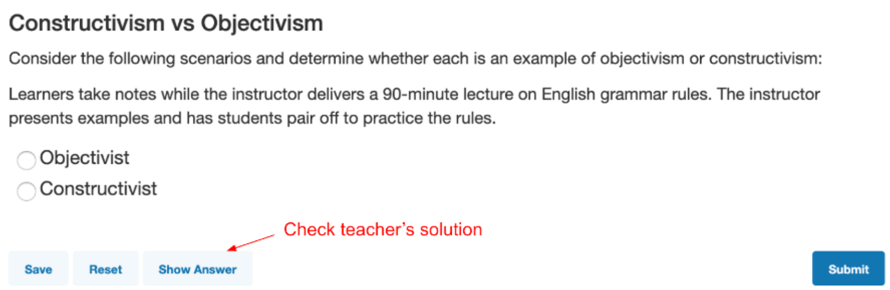


Figure 8: Problem-Solving Phase. This screen follows closely the problem screens in edX.

Figure 9 shows the sub-screen for the evaluation stage. This sub-screen is displayed when the learner chooses to submit their answer from the problem-solving sub-screen. This includes the Learner Profile, which displays the learners' KMA (labeled as Awareness) and KMB (labeled as Outlook) scores before the reflection prompts. This method of display is different from the original RA where the Learner

Profile is displayed after the evaluation reflection activities. The intention for placing the Learner Profile before the evaluation phase reflection prompts as opposed to after as what is done in the original RA is for the learners to account for both their metacognitive and cognitive performance during reflection. Some pointers based on their KMA and KMB scores are also displayed to the learners.

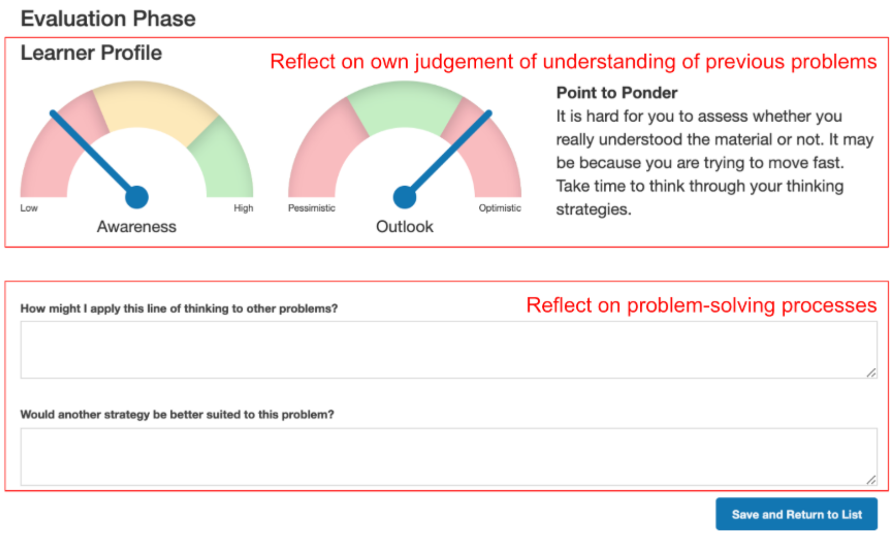


Figure 9: Evaluation Phase. Learners reflect on their problem-solving experience using the Awareness and Outlook measures.

Figure 10 shows the Learner Profile details. The KMA and KMB measures are shown as analog-style gauges with colors reminiscent of road traffic signals. Green areas indicate that the learner performs ideally, while red areas indicate that the learner still has to work on their metacognitive skills. Table 4 shows the pointers displayed to the learners based on the Learner Profile.

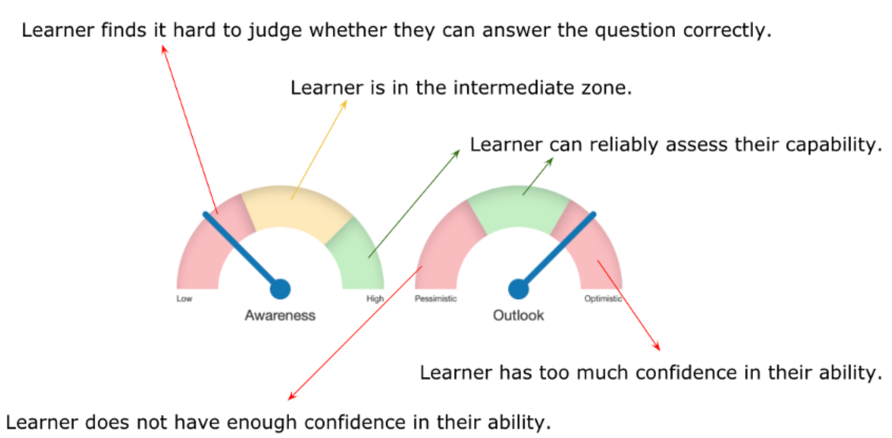


Figure 10: Learner Profile. Color-coding reminiscent of road traffic lights is used to signal metacognitive development progress.

Flavell's model of metacognition identifies two elements: knowledge of cognition (KOC) and regulation of cognition (ROC)[66]. Attention must be given to the learners' ability to plan, monitor, and evaluate their thinking to develop metacognition[67]. POALS' open response prompts are targeted at developing KOC by asking them about their

[66] J. H. Flavell, *"Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry,"* American Psychologist, *vol. 34, no. 10, p. 906, 1979.*

[67] *R. Fogarty,* The mindful school: how to teach for metacognitive reflection. *ERIC, 1994.*

Table 4: Learner Profile Pointers. The following pointers are displayed to the learners depending on their Learner Profile Classification. The Japanese text was introduced after revision.

Profile	Pointers
Low / Pessimistic	You tend to think your understanding is not enough even if you perform relatively well. Reflect on why you feel unsure about your answers. あなたは相対的にうまくやっていても、理解が十分ではないと思う傾向があります。また、なぜ答えがわからないのかを熟考します。
Low / Random	It is hard for you to assess whether you understood the material or not. It may be because you are trying to move fast. Take time to think through your thinking strategies. トピックを本当に理解しているのかどうかを評価することは、あなたにとって難しいです。次の内容に移るのが早いからかもしれません。時間をかけて自身の思考戦略を構築してください。
Low / Optimistic	There are times when you think you have already understood the material, but your actual performance tells otherwise. It would be helpful to reflect on what you thought was right that turned out to be wrong. トピックを既に理解していると思っていたが、実際にはそこまで理解していなかったという時があります。正しいと思ったことが間違っていると判明した時は、そのことを省みてみましょう。自身にとって役に立つはずです。
Average / Pessimistic	You are exhibiting more pessimism than necessary. Being a skeptic can be good since it can push you to work harder until you are very sure of your understanding, but it can also keep you from moving forward. あなたは必要以上に悲観的です。懐疑的になることは、自身の理解が完璧になるまで努力し続けることにつながるため良いことですが、一方で、与えられたタスクの進捗を妨げることにもつながります。
Average / Random	While you are neither optimistic nor pessimistic, you still need to better grasp your level of understanding of the material. As a reminder, there is no need for you to rush. あなたは楽観的でも悲観的でもありませんが、トピックに対する自身の理解度をよりよく把握する必要があります。念のため言いますが、急ぐ必要はありません。
Average / Optimistic	You are exhibiting more optimism than necessary. It is good to have optimism to keep your motivation level high, but some skepticism can be helpful too. あなたは必要以上に楽観的です。あなたのモチベーションを高く保つためにも楽観的になることは良いことですが、懐疑的になることも時には役立つでしょう。
High	You are doing great metacognitively. Keep it up! あなたはよく頑張っています。引き続き頑張ってください！

declarative and procedural knowledge. At the same time, the Learner Profile is intended to develop ROC through monitoring.

2.3 STUDIES ON POALS METACOGNITIVE TUTOR

The POALS Metacognitive Tutor was evaluated at different stages. A pilot study was first conducted for a relatively small class before conducting a full-scale experiment. The results of the pilot study were used to inform the updates needed before the full experiment. A spot-check was also conducted in the middle of the full-scale experiment to ensure that POALS Metacognitive Tutor updates did not introduce inadvertent adverse effects.

2.3.1 *Pilot Study*

Presented as: M. K. J. Carlon and J. S. Cross, "Developing learner metacognitive skills on an online environment," in *The 2020 Annual Spring Conference of Japan Society for Educational Technology*, 2020

POALS was pilot tested on an undergraduate class on educational technology delivered in a blended learning format. In this format, online educational materials are made available to the learners while still giving them the opportunity of face-to-face interaction through traditional classrooms. POALS Metacognitive Tutor was deployed to a unit of the class, and the experiment was conducted with 17 learners. The LMS edX Edge, the SPOC counterpart of edX, was used.

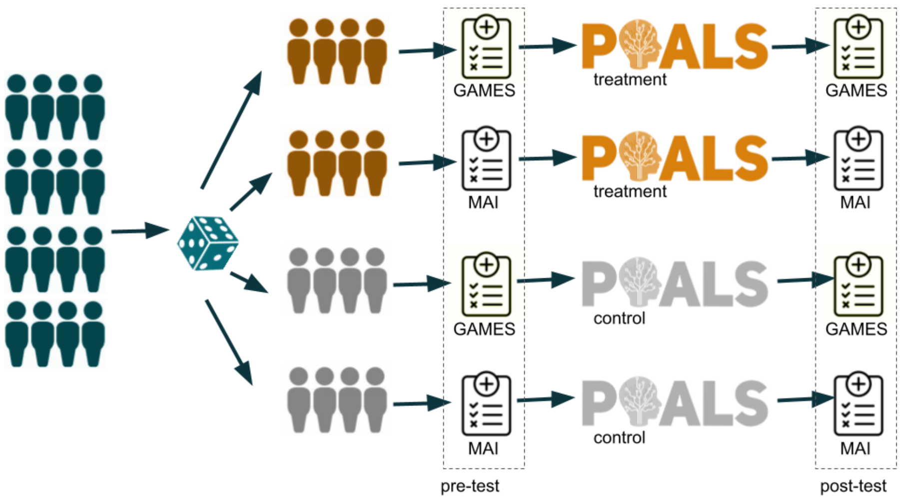


Figure 11: Randomized Control Trial. For the pilot study, four conditions were set to test both the software and the questionnaires.

A randomized controlled trial experimental design as illustrated in Figure 11 was employed for the pilot study. Two sets of control and treatment groups were created, where those in the treatment groups see all the sub-screens of POALS Metacognitive Tutor while those in the control groups only see the problem-solving sub-screens. The problem-solving sub-screen of POALS Metacognitive Tutor is identical to the typical problem component of edX. The experiment has multiple goals:

- To gauge to what extent can the results of the original empirical study on RA be conceptually replicated on POALS Metacognitive Tutor,
- To see how POALS Metacognitive Tutor works in its intended use case (beta test),
- To choose an offline measurement that can be used to validate the effectiveness of POALS Metacognitive Tutor when a larger scale study is conducted, and
- To identify usability issues with POALS.

Again, there are two ways for us to measure metacognitive development: through online measurements and offline measurements[68]. Online measurement is where we measure the metacognitive development while the learners are doing our exercises. For this, we used our Learner Profile. Ideally, Awareness should be as close as possible to 1, while Outlook should be as close as possible to 0. Figure 12 (a) shows the Awareness and Outlook scores of the learners in the treatment groups at the start and the end of the experiment. We can see a noticeable improvement in the awareness score, so we suspect that our tool is effective to some extent. There is a noticeable difference in outlook, but the result is not much more desirable than the learner outlook at the start.

[68] G. Schraw, *"Measuring self-regulation in computer-based learning environments,"* Educational Psychologist, *vol. 45, no. 4, pp. 258–266, 2010.*

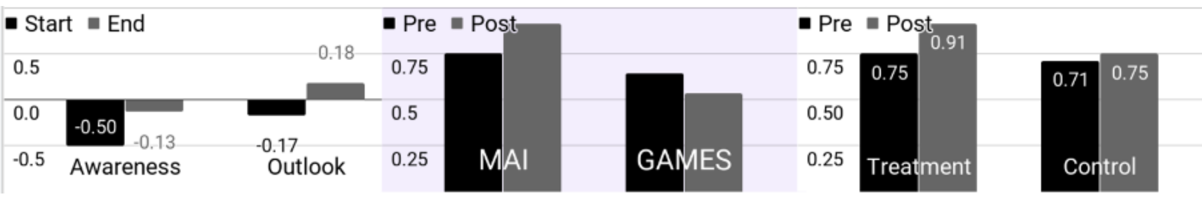


Figure 12: Pilot Study Results. Metacognitive measurement results using online measures (a, left), offline measures (b, center), and MAI across groups (c, right).

Another way of measuring is through offline measurements, where we measure the capabilities when the learners are not using the skill in question. The Metacognitive Awareness Inventory (MAI) and the Goal-oriented studying, Active studying, Meaningful and memorable studying, Explain to understand, and Self-monitor (GAMES) were the candidates for offline measurements of the learners' metacognition. The MAI was created to inquire about a person's knowledge of cognition and regulation of cognition[69]. The GAMES questionnaire, on the other hand, was created to evoke awareness of their self-regulating learning behaviors[70].

[69] G. Schraw and R. S. Dennison, *"Assessing metacognitive awareness,"* Contemporary Educational Psychology, *vol. 19, no. 4, pp. 460–475, 1994.*

The correlation between the starting and final metacognitive ability measure values are compared with the starting and final values in the Learner Profile. Figure 12 (b) shows the normalized MAI and GAMES scores of the learners in the treatment groups before and after exposure to POALS. We can see that MAI better reflects our observation from the online measurements. Hence, we will be using MAI for future studies involving larger samples. This result was not surprising considering that MAI is the one targeted for metacognition while GAMES was created for an adjacent concept.

Figure 12 (c) shows the MAI scores of the learners before and after the experiment divided into treatment and control groups. This reinforces our result from the online measurements that POALS can be effective in developing metacognition among learners in an online learning environment. However, as previously noted, the sample size is small. From the 17 learners who participated, only eight learners were made to answer the MAI. Thus, the treatment and control groups only have four learners each. As such, we have sufficient motivation to continue with our experiments but not strong enough results to arrive at conclusions.

Table 5: Learner Usability Feedback. These open responses from the pilot study learners were used as inputs to POALS software updates.

Positive	Negative
It was easier to answer the questions.	When we open a question, it shows the icon for correct answer. If you click that, it shows the right answer but records that you haven't answered to this question. Instead of this, if it showed the icon for correct answer after answering the question, that would be better.
I can do homework even through the phone.	I don't like doing multiple choice quizzes.
Easy to control.	Hard to click.

For usability, learners assigned to the treatment group answered a questionnaire that was modified from the Technology Acceptance Model (TAM). The original TAM was composed of six questions probing on the perceived usefulness of a tool and another six questions for evaluating usability, all questions presented in a Likert scale with seven answer options[71]. The modifications include reducing the answer options to five (1 being the worst and 5 the best) to prevent decision paralysis and reducing the Likert questions to nine to exclude questions that are not directly applicable for assessing our tool. In addition, two open-ended questions were also added to enable the subjects to express their opinions on the tool.

[71] *F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," PhD thesis, Massachusetts Institute of Technology, 1985.*

The average overall rating provided by the learners is 2.7, which is below the median point of 3. This indicates the reception to POALS Metacognitive Tutor is not very positive. Table 5 shows some free-form responses from the learners, which gives us more details as to why POALS Metacognitive Tutor was received negatively. Some software defects were encountered by the learners, which are since then fixed. Another contributing factor is the dislike for the assessment types used, which is not due to POALS itself. The other responses seem to indicate that POALS will just be as acceptable to the learners as the LMS it was used on.

However, a severe usability flaw was observed not through the usability questionnaire responses but from the actual usage by the learners. In several cases, student responses to metacognitive prompts consisted of one-liners such as "Yes," "No," "Maybe," and "I don't know," among others. This could be a sign of fatigue or lack of motivation to participate, indicating that more work is needed to help learners manage their cognitive resources.

2.3.2 *Revision and Spotcheck*

> Presented as: M. K. J. Carlon and J. S. Cross, "Open response prompts in an online metacognitive tutor," in *The 2021 Annual Spring Conference of Japan Society for Educational Technology*, 2021

POALS Metacognitive Tutor showed positive results during the pilot study in terms of metacognitive development, but the learners gave terse responses (sometimes, single words) for the prompts. These prompt responses were mandatory, but no minimum input length was required. While there are works that show that prompts are beneficial[72], there are also those that say that prompts can be detrimental if excessively used[73]. Unfortunately, we could not find research that has explored the optimal number of metacognitive prompts. We already attempted to optimize POALS Metacognitive Tutor used for the pilot study by having only two prompts in each phase while still keeping the metacognitive measures.

The prompts' under-utilization may be due to fatigue since learners need to answer the prompts every time they tackle a quiz question. It is also possible that the learners fail to see their significance. To address these concerns, POALS Metacognitive Tutor was modified so that the learners will only see one prompt in each phase. The prompts were edited to allow the learners to think more concretely, and the evaluation prompt was modified depending on the learners' answers during the problem-solving phase. This way, the learners can see that their actions matter and they should reflect on them accordingly. Minimum inputs were still not set to maintain learner autonomy. Table 6 shows the open response questions before and after the updates.

[72] *K. Berthold, M. Nückles, and A. Renkl, "Do learning protocols support learning strategies and outcomes? the role of cognitive and metacognitive prompts,"* Learning and Instruction, *vol. 17, no. 5, pp. 564–577, 2007.*

[73] *L. Vangsness and M. E. Young, "More isn't always better: When metacognitive prompts are misleading,"* Metacognition and Learning, *pp. 1–22, 2020.*

Table 6: POALS Metacognitive Tutor Update. Open response prompts before and after the POALS Metacognitive Tutor update.

Phase	Before	After
Preparation	What prior knowledge can help me with this particular task?	What do I need to know before I can successfully deal with this task? このタスクを正常に処理するには、何を知る必要がありますか？
	What do I need to know before I can successfully deal with this task?	
Evaluation	How might I apply this line of thinking to other problems?	*When answer is correct–* What worked out when I was solving this problem? この問題に取り組んでいたときに何がうまくいきましたか？
	Would another strategy be better suited to this problem?	*When answer is wrong–* What went wrong when I was solving this problem? この問題に取り組んでいたときに何がうまくいかなかったのですか？

The new POALS Metacognitive Tutor was deployed in an electrical engineering blended class taught during Summer 2020, which is offered in Japanese. As such, all English text used in POALS were augmented with their Japanese translations. In addition, data from this new deployment was gathered before the full experiment was completed to check for adverse effects due to revisions.

First, we needed to verify that the reduction of prompts did not negatively affect the metacognitive tutoring. Figure 13 shows the learners' metacognitive scores against the number of opportunities plotted with regression lines. The lines' positive slopes indicate that POALS is still able to assist in metacognitive development. The p-value resulting from the linear modeling used for trend analysis was also calculated. The null hypothesis which is being tested by the p-value for linear modeling is that the independent variable does not have any significant correlation with the dependent variable. When using the typical cut-off of 0.05 for rejecting the null hypothesis, we can say that statistical significance is observed. As such, we have some confidence that the trend shown by the linear modeling is reflective of the actual relationship between number of opportunities and metacognitive scores.

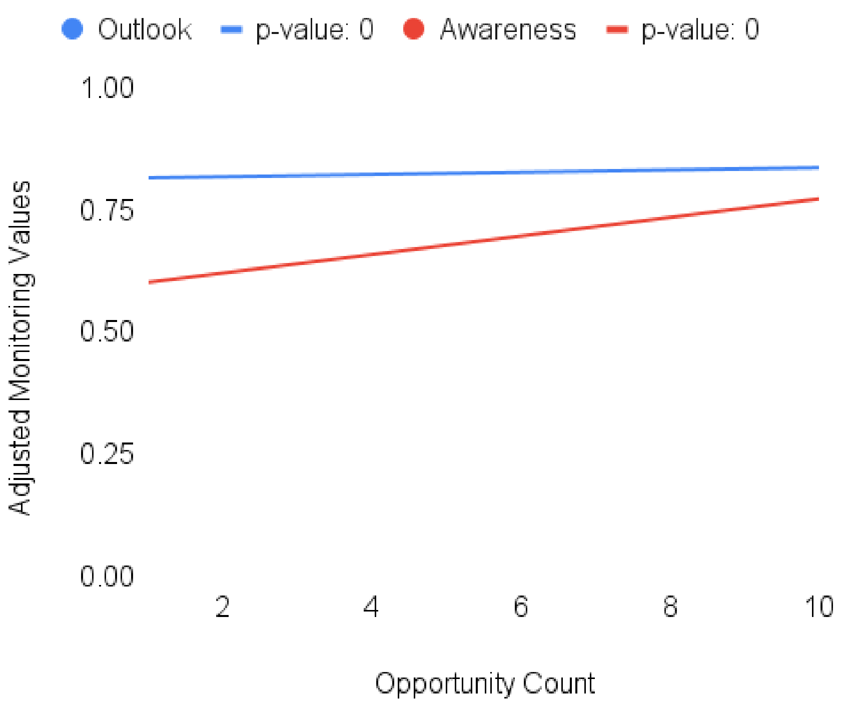


Figure 13: Learner Profile Tracking (Spotcheck). Adjusted Learner Profile values through opportunity count during spotcheck.

POALS Metacognitive Tutor conducts metacognitive tutoring in two ways. One is by active tutoring through the prompts. The other is through passive tutoring, where the learners are asked about their confidence ratings in answering problems. The ratings, together with their actual performance, are used to compute the Awareness (KMA)

Figure 14: Gains and Preparation Phase Correct Count. Difference between Learner Profile values at the start and at the end of the experiment compared with correct usage of the preparation phase prompts.

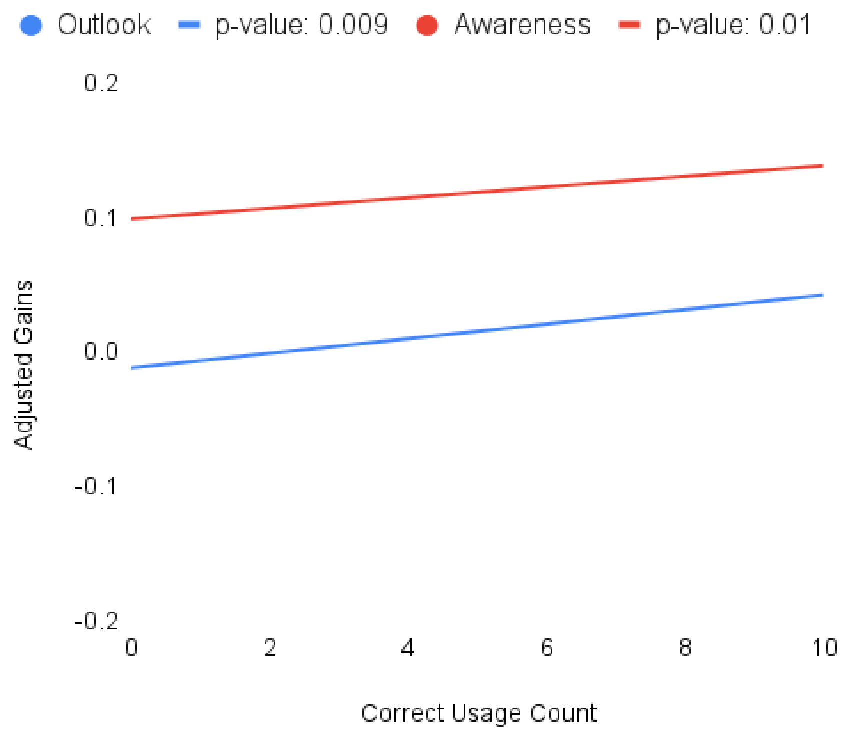
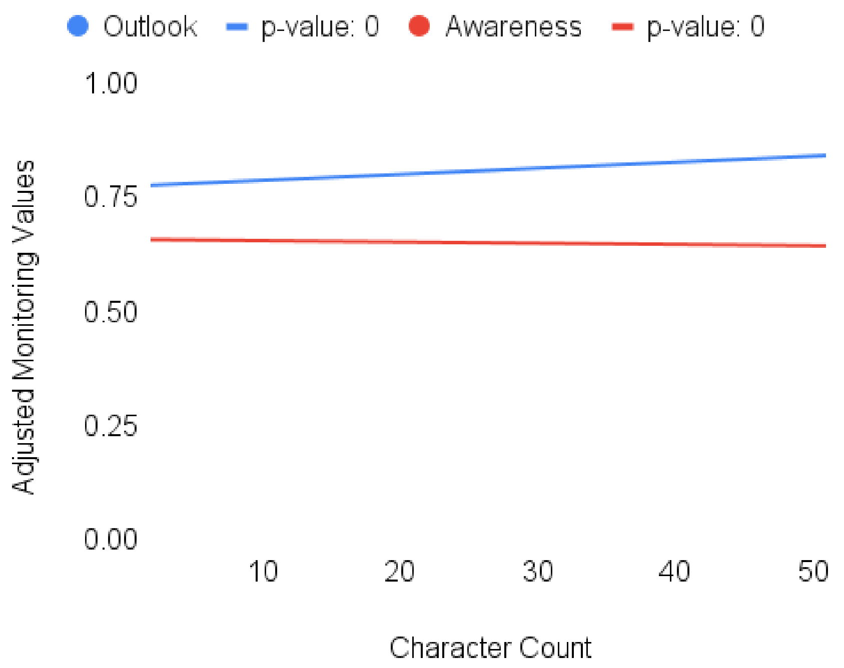


Figure 15: Learner Profile and Evaluation Phase Input. Adjusted Learner Profile values and lengths of evaluation phase responses.



and Outlook (KMB) metacognitive scores. The learners are then given hints on how they could improve their learning based on these scores.

This raises the question: given that they can be taxing, are the prompts essential? Figure 14 shows the number of times the learners correctly used (i.e., the learners inputted meaningful answers) the preparation prompts to review their lesson plotted against the difference between the metacognitive scores at the start and the end of the experiment. Just like in Figure 13, regression lines were used for analysis. We can see from here that if used correctly, preparation prompts can lead to higher Outlook and Awareness scores. Likewise, the p-value is also less than 0.05, indicating statistical significance.

The evaluation prompt is equally valuable. Figure 15 is similar to Figure 13, except that the horizontal axis pertains to the number of characters inputted by the learners in the evaluation prompts. The more inputs the learner makes, the better are their outlook scores. Just like the other cases investigated, there is statistical significance with p-value being less than 0.05.

In summary, outlook and awareness improvements are still evident even after the prompts were significantly reduced. Proper usage of the preparation prompt can lead to better metacognitive outcomes. Encouraging learners to reflect more through the evaluation prompts can be beneficial to outlook.

2.3.3 *Full Experiment*

> Portions submitted as: M. K. J. Carlon and J. S. Cross, "Development of open-response prompt-based metacognitive tutor for online classrooms," in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Submitted

Finally, a full experiment was conducted to check the effectiveness of the updated POALS. This study aims to answer the following research questions:

- **RQ1:** Can we see improved performance on POALS Metacognitive Tutor similar to what was observed in RA?
- **RQ2:** Can POALS Metacognitive Tutor improve metacognition?
- **RQ3:** Is POALS Metacognitive Tutor usable in an online learning environment?

2.3.3.1 *Experimental Design and Participants*

POALS was thoroughly tested with a seven-week undergraduate electrical engineering class offered in a Japanese technical university. Initially, the course is delivered in a blended format. Learners and teachers meet face-to-face during the first week while the middle four weeks are delivered online through the edX Edge platform. The learners can work on the online component of the class at their own pace. Each week of the online course is divided into multiple parts, where only the first parts are required, and the rest are optional enrichment materials. The final two weeks are in-person laboratory activities and report writing. Due to the COVID-19 pandemic, the class offered in Fall 2020 was delivered in a hybrid online format. No changes were made to the original online portion of the course except for the introduction of POALS. The previously face-to-face meetings were conducted online synchronously using a video conferencing tool. In addition, POALS was deployed to the mandatory first quiz each week of the online course. There are ten exercises in total spread across the four weeks.

Participation in the study was voluntary. The learners were informed that their teachers would not know whether they participated in the research or not. Thus, non-participation will not affect the learners negatively. The research protocol has passed the university's human subject ethics review. Table 7 shows the participant details.

Table 7: Full Experiment Participants. Gender, group breakdown, and age of the participants.

Variables	Levels	Unit	Value
Gender	Male	Count (Percentage)	22 (75.86)
	Female		7 (24.14)
Group	Control		15 (51.72)
	Treatment		14 (48.28)
Age		Average	19

The learners were asked to answer the MAI with Japanese translation after providing their consent. The same questionnaire was answered at the end of the experiment. A randomized controlled trial experimental design was employed for the study. Those in the treatment group see all the sub-screens of POALS Metacognitive Tutor, while those in the control group only see the problem-solving sub-screens. In addition, both groups were made aware that they are accessing the experimental software without informing them what the treatment looks like. This is to control for placebo effect in the post-experiment questionnaire. The placebo effect is when research subjects perceive phantom improvements just because they expect to by mere exposure to a treatment[74]. Since both groups knew they are in some form of experiment, both groups may experience the placebo effect making it a common factor and not something isolated to the treatment group.

[74] W. R. Boot, D. J. Simons, C. Stothart, et al., *"The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects,"* Perspectives on Psychological Science, *vol. 8, no. 4, pp. 445–454, 2013.*

Aside from the MAI, the learners in the treatment group were also asked to answer the modified TAM with Japanese translation. The learners were informed that their details would remain confidential, and they are encouraged to answer all the questionnaires as honestly as possible. This is to control social desirability bias[75] where the research subjects respond according to what they perceive would make them look better (in our case, the learners having good study habits) or satisfy the researcher (the tool is working smoothly).

Thematic analysis, a qualitative research method where texts are coded according to themes[76], was used for the preparation open-response prompt. The learner responses are coded as follows:

- **indifference**: Response is contrite and uninformative.
- **strategy**: A strategy (e.g., *watching videos, reading the question carefully*) was provided.
- **general**: The learner alluded to course material without providing information that can be traced back to a specific module (e.g., *information about electrical engineering*).
- **specific**: A specific course content related to the question was specified.

We also noted the response lengths as a pseudo-measure of the reflection depth during the evaluation phase.

[76] *V. Braun and V. Clarke, "Using thematic analysis in psychology,"* Qualitative Research in Psychology, *vol. 3, no. 2, pp. 77–101, 2006.*

2.3.3.2 *RQ1: Can we see improved performance on POALS Metacognitive Tutor similar to what was observed in RA?*

The overall performance between the control and treatment groups is compared using the Mann-Whitney test (or two-sided unpaired Wilcoxon signed-rank test), a non-parametric hypothesis test for independent groups[77]. Effect sizes are typically interpreted as a small effect for values between 0.1 and 0.3, medium from 0.3 to 0.5 for moderate effect, and greater than 0.5 for large effect[78]. Despite the seemingly positive statistics shown in Table 8, there is no statistical significance between groups ($W = 71$, $p = 0.127 > 0.05$) and the effect is small (effect size = $0.287 < 0.3$). As such, while exposure to POALS Metacognitive Tutor may lead to improved performance, our data lacked the statistical significance to support said conclusion.

[77] *H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other,"* The Annals of Mathematical Statistics, *pp. 50–60, 1947.*

[78] *A. Kassambara,* rstatix: Pipe-friendly framework for basic statistical tests, *R package version 0.6.0, 2020. [Online]. Available: https://CRAN.R-project.org/package=rstatix.*

Table 8: Learner Total Scores. Descriptive statistics of the learners' total scores.

Statistic	Control	Treatment
Minimum	6	6
Maximum	9	10
Median	7	8
Mean	7.4	8.07
Standard Deviation	0.91	1.21

Unlike the RA experiment where the learners were constrained to just an hour of use of the system, POALS Metacognitive Tutor users were not given time restrictions. This aligns well with the typical use case for online learners, where they can work on the course at their convenience. For RA, fewer exercises attempted by the learners in the treatment group were taken as evidence of metacognitive ability. This is presumably because the more metacognitive learners take more time to think through their answers. In our case, because learners are not given time limits, they are expected to complete all exercises. Therefore, we attempted to see the correlations between problem-solving time and performance. However, since the activities are self-paced, there were instances where learners did not finish an exercise and resumed working the next day. Therefore, we deem that our time logs will not accurately picture the dedicated problem-solving effort.

2.3.3.3 *RQ2: Can POALS Metacognitive Tutor improve metacognition?*

As a recap, there are two ways to measure metacognitive development: through online measurements and offline measurements. Online measurements tracks the metacognitive development while doing our exercises. For this, we used our Learner Profile. Ideally, KMA should be reaching 1, while KMB should be reaching 0. For RA, the Learner Profile was tracked for both the control and treatment groups. For POALS, Learner Profile tracking was done for the treatment group only since we deem the self-assessment to be a reflective activity. Doing the same for the control group may induce reflection on the subjects, thus defeating having a control group.

Figure 16 shows the locally estimated scatterplot smoothing (LOESS) curves for KMA and KMB scores throughout the experiment. Localized smoothing where weighted regression is done for few nearby data points[79] was used instead of other regression methods since the data points are relatively sparse. There is a noticeable positive difference in Outlook, though the result is not considerably better than the learner Outlook at the start (the perceived slope is gentle). While KMB is behaving the way we hoped it to be to some extent, the same cannot be said about KMA, which is contradictory to what we had seen during the pilot study. With the contradictory result and the absence of corresponding control group measurement, it will be hard to derive conclusions from the Learner Profile.

[79] *W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots,"* Journal of the American Statistical Association, *vol. 74, no. 368, pp. 829–836, 1979.*

Again, another way of measuring is through offline measurements. We measure the capabilities when the learners are not using the skill in question with offline measurements. The MAI was created to inquire about a person's KOC and ROC. The MAI is comprised of 52 items: 17 items measuring KOC and 35 measuring ROC. Learners in both control and treatment groups answered the MAI at the beginning (pre-test) and the end (post-test) of the experiment. Thus, the MAI may be a better measurement for ROC development than the

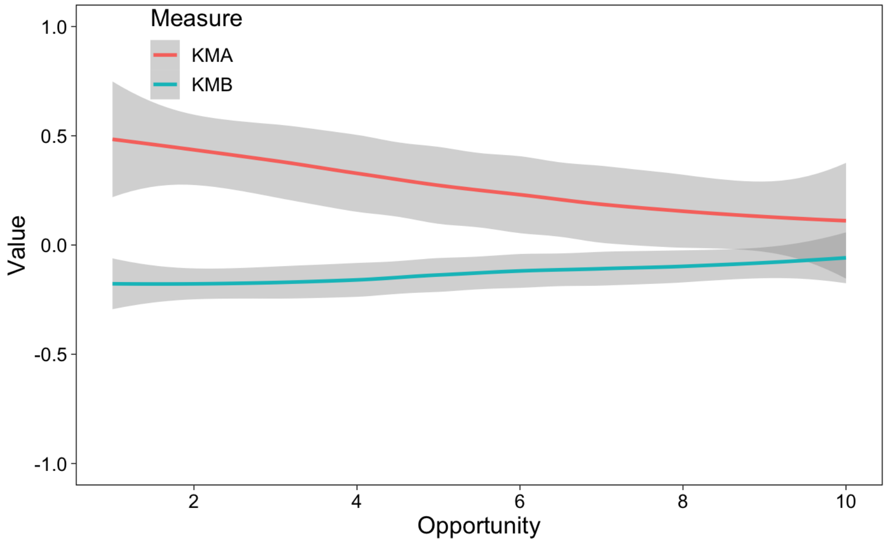


Figure 16: Learner Profile Tracking (Full Experiment). Learner Profile values through opportunity during full experiment.

Learner Profile since it is possible to compare the treatment group with the control group.

The original MAI uses 100-millimeter bands with the ends indicate complete disagreement and agreement. The respondent can then place tick marks within the band according to their level of agreement with a given statement. We modified the response type to binary (true or false) instead for ease. Additionally, the questions were presented to the students in English and Japanese. Because of the MAI changes, we checked its reliability using Cronbach alpha. The computed Cronbach alpha for MAI and its subscales are all more than 0.75 for both pre-test and post-test, as shown in Table 9, suggesting good internal consistency.

Scale	Pre-test	Post-test
MAI	0.885	0.893
ROC	0.838	0.866
KOC	0.801	0.757
Standard Deviation	0.91	1.21

Table 9: MAI Cronbach alpha Scores. Values during pre-test and post-test are compared.

To check that the learner assignment to groups resulted in the same distribution, we conducted a two-sided unpaired Wilcoxon signed-rank test to compare the control and treatment groups during the pre-test. There were no statistical significance found for both KOC ($W = 98$, $p = 0.774 > 0.05$) and ROC ($W = 125$, $p = 0.393$), indicating that we were not able to reject the hypothesis of the test that the two groups are equal.

The W-statistic is based on the sum of the ranks of the first sample with the minimum value subtracted. This value is used for computing the p-value.

The V-statistic is based on the pairwise difference between the individuals in two groups. This value is used for computing the p-value.

The one-sided Wilcoxon signed-rank test was used to compare the pre-test and post-test values of KOC and ROC for both control and treatment groups. The effect size was likewise calculated. The results are summarized in Table 10. Only the ROC for the treatment group exhibited statistical significance, indicating that POALS may have improved the students' ROC. The effect size is large; hence we see this result as very promising. However, the same cannot be said for KOC. As such, it is possible that the metacognitive development had not been significant enough to affect performance significantly as we had seen in Section 2.3.3.2 RQ1: Can we see improved performance on POALS Metacognitive Tutor similar to what was observed in RA?.

Table 10: Pre-test and Post-test Comparisons. Inferential statistics were used for comparison.

Group	Scale	Statistic	Value
Control	KOC	V (p)	14.5 (0.054)
		effect size (magnitude)	0.483 (moderate)
	ROC	V (p)	31.5 (0.291)
		effect size (magnitude)	0.11 (small)
Treatment	KOC	V (p)	29.5 (0.076)
		effect size (magnitude)	0.39 (moderate)
	ROC	V (p)	0 (< 0.001)
		effect size (magnitude)	0.874 (large)

For POALS Metacognitive Tutor, the Learner Profile was intended to trigger ROC while the open response prompt questions are targeted at probing KOC. From the poor KOC performance and our experience of students not engaging with prompts during pilot study, we reviewed how the students utilized the prompts during the full experiment. Upon closer inspection of the prompt responses, we noticed that several learners noted generic actions such as reviewing the video lectures or reading the question more carefully. While those reflections might indeed be accurate, they are not mainly targeted at improving the specific skill being tested by the exercise. The highest percentage of preparation phase response types where the students answer wrong is a strategy (32%), followed by indifference (11%).

The learner responses in the evaluation prompt also mattered. In cases where the learners answered an exercise correctly followed by an incorrect answer, the learners provided evaluation responses that are less than ten bytes 56% of the time. On the other hand, learners who previously got it wrong could transition to correct if they input more than 10 bytes 75% of the time. Thus, it is important to stress to the learners to take the metacognitive prompts seriously to see the benefits to their performance. This challenge is also recognized by other researchers studying metacognition. For metacognitive instruction to be effective, it must be well-rooted in a context[80].

[80] *I. Roll, V. Aleven, B. M. McLaren,* et al., *"Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking,"* Metacognition and Learning, *vol. 2, no. 2, pp. 125–140, 2007.*

Byte count instead of word count was used to account for the response sizes. A single Sino-Japanese character is typically two bytes, while English-Japanese translators set 2.5 Sino-Japanese characters per English word as a rule-of-thumb[81]. English words, on the other hand, have around 4.79 characters[82] which are encoded as single bytes. Thus, we would like the learners to have specified at most minuscule two medium-length words for the response to count as significant. Of course, it is possible to use natural language processing (NLP) parsers to more accurately count the number of words or simply count the words manually. It appeared from our spot-check, though, that the above strategy provided an excellent word count approximation, less prone to human error present in manual counting, and more straightforward than NLP parsers whose implementations may not be consistent. For example, 「勉強しませんでした」can be interpreted as a single token (or word), two tokens (勉強+ しませんでした), or even four tokens (勉強+ しま+ せん+ でした) depending on the tokenization mechanism of the NLP parser.

[81] *ProZ.com Translation Services*, Word count of translation into English from Japanese, *Retrieved February 22, 2021 from* `https://www.proz.com/forum/japanese_%E6%97%A5%E6%9C%AC%E8%AA%9E/14521-word_count_of_translation_into_english_from_japanese.html`, *nd.*

[82] *P. Norvig*, English letter frequency counts: Mayzner revisited or ETAOIN SRHLDCU, *Retrieved February 23, 2021 from* `http://norvig.com/mayzner.html`, *nd.*

2.3.3.4 *RQ3: Is POALS usable in an online learning environment?*

For usability, learners assigned to the treatment group answered a modified questionnaire based on TAM. Those assigned in the control group were not asked to answer the modified TAM since they were not exposed to POALS Metacognitive Tutor. Again, the original TAM was composed of six questions probing on a tool's perceived usefulness and another six questions for evaluating usability. All questions were presented on a Likert scale with seven answer options[83]. The modifications include reducing the answer options to five (1 being the worst and 5 the best), which was shown to yield better data quality[84]. The questions were also reduced to nine to exclude questions not directly applicable for assessing our tool. Finally, two open-ended questions were added to enable the subjects to express their opinions on the tool. Figure 17 shows the box plots for each of the items in our modified TAM.

[83] *F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," PhD thesis, Massachusetts Institute of Technology, 1985.*

[84] *M. A. Revilla, W. E. Saris, and J. A. Krosnick, "Choosing the number of categories in agree–disagree scales,"* Sociological Methods & Research, *vol. 43, no. 1, pp. 73–97, 2014.*

The learners rated the system favorably, with an overall mean rating of 3.606 (standard deviation equal to 1.168) and a median of 4. The third quartile values for each Likert scale are at least 4. This was despite a few bugs due to changes in handling Django sessions reported by the learners during the research. These bugs most likely prevented the learners from working with the tool at their best. Several learners appreciated working at their own pace and the quizzes to assist them in their study.

2.4 CONCLUSIONS ON THE METACOGNITIVE TUTOR

We first revisited RA, which was developed to be a generic model for metacognitive instruction. RA is designed to be used with other

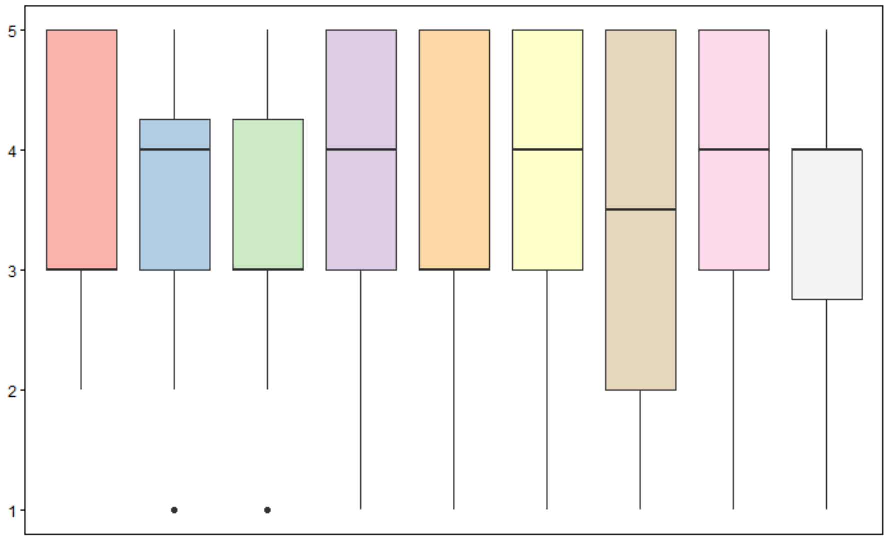


Figure 17: Usability Scores. Box plot of the scores for each usability survey item.

cognitive instruction but does not make assumptions on the cognitive domain's underlying thinking models. Thus, it can be used for various domains, making it a good fit for metacognitive instruction on online learning platforms. We then analyze how RA can be further optimized for online learning platforms. We constructed POALS to increase usability on LMS.

Several studies were conducted to test POALS Metacognitive Tutor's effectivity. Results reveal that while POALS Metacognitive Tutor can be beneficial to developing ROC, it may not be sufficient for improving KOC. A possible approach that can be taken in the future is introducing a short lesson on metacognition to help the learners appreciate the value of planning for and reflecting on their knowledge.

While the pilot study results were used as inputs for the full experiment, there could be many reasons why the same pilot study success was not evident during the full experiment. For one, the pilot study participants are more mature, with some in their third and fourth years of college. On the other hand, the full experiment participants are mostly freshmen. Maturity can then be a confounding variable as existing research have shown that cognitive development stages affect metacognition. Additionally, the pilot study participants enrolled in the class out of their own volition, hence they may have been more motivated about education-related activities from the start. On the other hand, the full experiment was conducted on a mandatory course, so motivation levels may not be as high as those in the pilot study. Attempts to reduce this effect were done by setting measures to reduce desirability bias and having control conditions while managing placebo effect.

2.5 LOW-TECH SOLUTIONS FOR METACOGNITIVE DEVELOPMENT

We started out on this work acknowledging that online learning is a viable solution to making learners future ready in the face of changing job demands. Consequently, metacognition is important to online learning. However, even in the case where online learning is not feasible, metacognition will still be important since it is part of 21st century skills needed to be competitive in the knowledge economy. But face-to-face format does not make metacognitive instruction easier: the same prerequisites of making the learners see its value and putting in context still exists[85].

The following approach can be used for planning for metacognitive instruction[86]:

1. **Define and describe key metacognitive elements to be targeted.** As we have learned, there are many aspects to metacognition. What aspect will you specifically focus on? It can be as generic as giving the learners the free reign to choose the metacognitive strategy that suits them best, or it can be as specific as learning how to build mnemonics for a module that requires a lot of memorization.

2. **Articulate metacognitive best practices.** Sometimes, even the best techniques can be less optimal. For example, creating to-do lists was long held to be beneficial. But to-do lists that did not consider prioritization in advance can lower effectiveness and efficiency[87].

3. **Adapt instructional materials to fit context.** Not only should you fit your metacognitive instruction to your context, but you must also consider the additional cognitive load introduced by metacognitive instruction. As previously mentioned, some tasks benefit more from metacognition than others.

4. **Identify possible ways the learner will respond to the instruction.** Consider your metacognitive goal as part of your learning outcomes. Designing rubrics for assessing metacognitive outcomes can help you think of metacognitive indicators that you can watch out for.

5. **Formulate supportive feedback.** The acronym *BACE* can be used as a guide:
   - *Believe* in the learner's potential.
   - *Affirm* achievements as well as honest efforts to improve.
   - *Challenge* the learners to deepen their knowledge.
   - *Encourage* the learners to continue to persevere.

[85] *I. Roll, V. Aleven, B. M. McLaren,* et al., *"Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking,"* Metacognition and Learning, *vol. 2, no. 2, pp. 125–140, 2007.*

[86] *P. Cunningham, H. Matusovich, and S. Blackowski, "Teaching metacognition: Helping students own and improve their learning," in* Workshop presented at the American Society for Engineering Education Annual Conference and Exposition, *2018.*

[87] *Oregon State University. (2021). Make better to-do lists, [Online]. Available:* `https://success.oregonstate.edu/learning/better-lists`.

6. **Acknowledge how metacognitive learning can be rewarding.** This goes back to the prerequisite that if the learners do not appreciate metacognition's importance, it will be hard to see positive effects.

Indeed, metacognitive instruction and learning can be rewarding. After the pilot study was conducted, the Educational Technology students were given a short lecture about metacognition. To assess their learning outcomes, the students were instructed to embed metacognition in their final project which is a group work involving the creation of an online class. Most groups went out of their way: some structured their courses using a didactic approach, most were mindful to add formative exercises, and some even created interactive components such as shown in Figure 18.

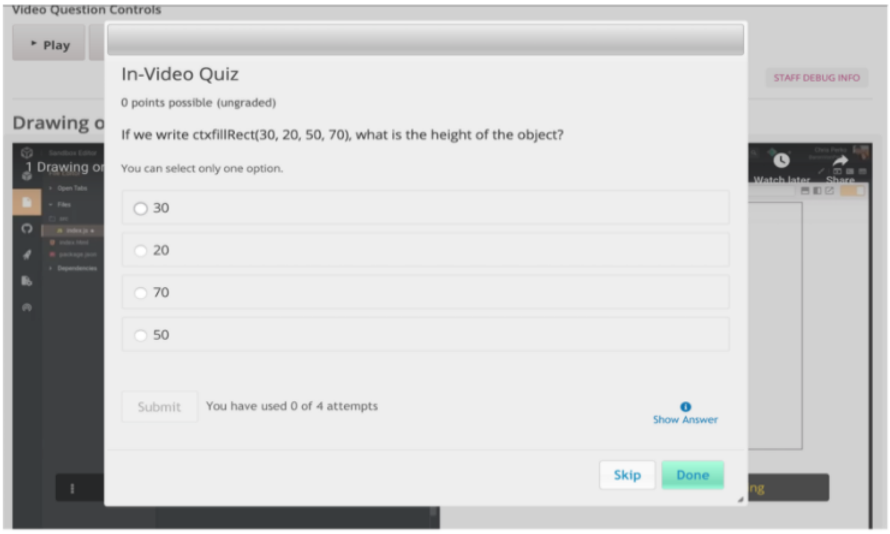


Figure 18: Student-Produced Metacognitive EdTech. One of the groups in the Educational Technology class created a plugin for edX that pops up a question in the middle of an instructional video to promote metacognition.

2.6 REMAINING CONCERNS

The lack of conclusivity in some aspects investigated can possibly be resolved by improving sampling methods. The most straightforward solution is increasing the data collection period. This can be extremely time-consuming especially in the university environment where there will be a need to wait for at least one academic term for the next data collection period to start. Improving sampling can also include shortening the experiment period similar to what is done with other previous research. However, as previously mentioned, this has the down-side of not being able to measure authentic learning outcomes since learning will often take more than an hour to happen. Another option is to provide attractive incentives such as monetary rewards. This has been attempted in some of the other courses where POALS Metacognitive Tutor has been deployed with no success. One way to rectify this is to advertise the opportunity to populations who would be more receptive to such incentives. In social sciences, crowdsourced human resources such as Mechanical Turk has been shown to offer affordable and highly diverse research participant pool[88].

[88] *A. M. Mellis and W. K. Bickel, "Mechanical Turk data collection in addiction research: Utility, concerns and best practices,"* Addiction, *vol. 115, no. 10, pp. 1960–1968, 2020.*

This research will definitely benefit from a diverse research applicant pool. Since POALS Metacognitive Tutor requires self-assessment, we often receive questions on whether demographic factors such as culture, gender, age, and others can affect the Learner Profile. Unfortunately, we could not probe this question since the class where POALS Metacognitive Tutor was thoroughly tested was overwhelmingly male. Being delivered in Japanese, the course did not attract many international students. It was also a freshman course; thus, there was not much variance in age.

As with any new tool, POALS Metacognitive Tutor can benefit from a more in-depth user study. In this research, the learners were asked to give their insights about the tool, but they were not given the opportunity to exchange ideas with the researchers through interviews or focus group discussions. There were also no sessions where the researchers can watch the learners use the tool live and observe for possible bottlenecks (e.g., will the prompts cause interruptions with the learners' problem-solving?).

Another important thing to consider is the learners' cognitive resources. First and foremost is to assess whether the exercises where the prompts will be included are suitable for metacognitive training. Metacognitive prompts may be counterproductive when added to simple tasks such as remembering facts. Even in cases where metacognitive training may be warranted, it may be more prudent to select very targeted metacognitive skills (e.g., help-seeking) instead of multiple skills such as in the case of the POALS Metacognitive Tutor. As a recall, metacognitive training is found to be primarily practical when done within a cognitive context. However, doing so can put an unsustainable cognitive load on the learners[89]. Thus, it is advisable to monitor the learners' metacognitive development through online measures that are regularly updated such as the Learner Profile. Another option is to consider other ways to help manage cognitive load. An AI solution for managing cognitive load that has been gaining traction is **adaptive learning**.

[89] *I. Roll, V. Aleven, B. M. McLaren,* et al., *"Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking,"* Metacognition and Learning, *vol. 2, no. 2, pp. 125–140, 2007.*

3 THE ADAPTIVE ENGINE

M. K. J. Carlon and J. S. Cross, "Knowledge tracing for adaptive learning in a metacognitive tutor," *Open Education Studies*, Submitted

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where-" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"–so long as I get SOMEWHERE," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

–Lewis Caroll, *Alice in Wonderland*

Learners are as varied as their points of view of learning. One way to understand learning points of view is through the *Modes of Learning* framework as shown in Figure 19[90]. In this framework, learning is viewed from two axes. One is classifying whether learning is hierarchical (i.e., there is an order to which learning contents are ideally presented to learners according to their ability) or distributed (i.e., a piece of knowledge can have multiple uses and the order to which it should be presented to the learners depend on its utility). The other axis is classifying whether learning is an individual or social (collective) activity.

Where an individual predominantly stands in the quadrant can influence their learning philosophies. For example, a person who falls in the hierarchical individual quadrant will see the value of perseverance or grit[91] and the crucial role the teacher plays for a learner's success[92]. On the other hand, a person who falls in the distributed collective quadrant would find connectivist MOOCs (cMOOCs) where learning happens through information exchange between loosely connected individuals[93] to be more worthwhile. Because of these varied stances, it is essential to consider how to cater to different learners. This is more crucial when there is an attempt to apply solutions at scale where catering to significantly different learners is required.

Adaptive learning refers to educational technologies that monitor learner progress and interaction and uses that information to modify the instruction. A gamified example of this is the instruction delivery method inspired by the *Choose Your Own Adventure* book series. This style has been used in teaching highly complex topics such as social exchange theory[94], interactive topics such as cybersecurity[95], and highly personal topics such as ethics[96].

[90] *R. Elmore,* Modes of learning framework, *edX, Ed., 2020. [Online]. Available:* https://www.edx.org/course/leaders-of-learning.

[91] *A. L. Duckworth. (Apr. 2013). Grit: The power of passion and perseverance. T. Talks, Ed., [Online]. Available:* https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance.

[92] *R. F. Pierson. (May 2013). Every kid needs a champion. T. Talks, Ed., [Online]. Available:* https://www.ted.com/talks/rita_pierson_every_kid_needs_a_champion.

[93] *O. Skrypnyk, S. Joksimovic, V. Kovanovic,* et al., *"Roles of course facilitators, learners, and technology in the flow of information of a cMOOC,"* The International Review of Research in Open and Distributed Learning, *vol. 16, no. 3, 2015.*

[94] *E. Ribarsky, "Choose your own adventure: Examining social exchange theory and relational choices,"* Communication Teacher, *vol. 27, no. 1, pp. 29–32, 2013.*

[95] *T. Chothia, S. Holdcroft, A.-I. Radu,* et al., *"Jail, hero or drug lord? Turning a cyber security course into an 11 week choose your own adventure story," in 2017* {USENIX} Workshop on Advances in Security Education ({ASE} 17)*, 2017.*

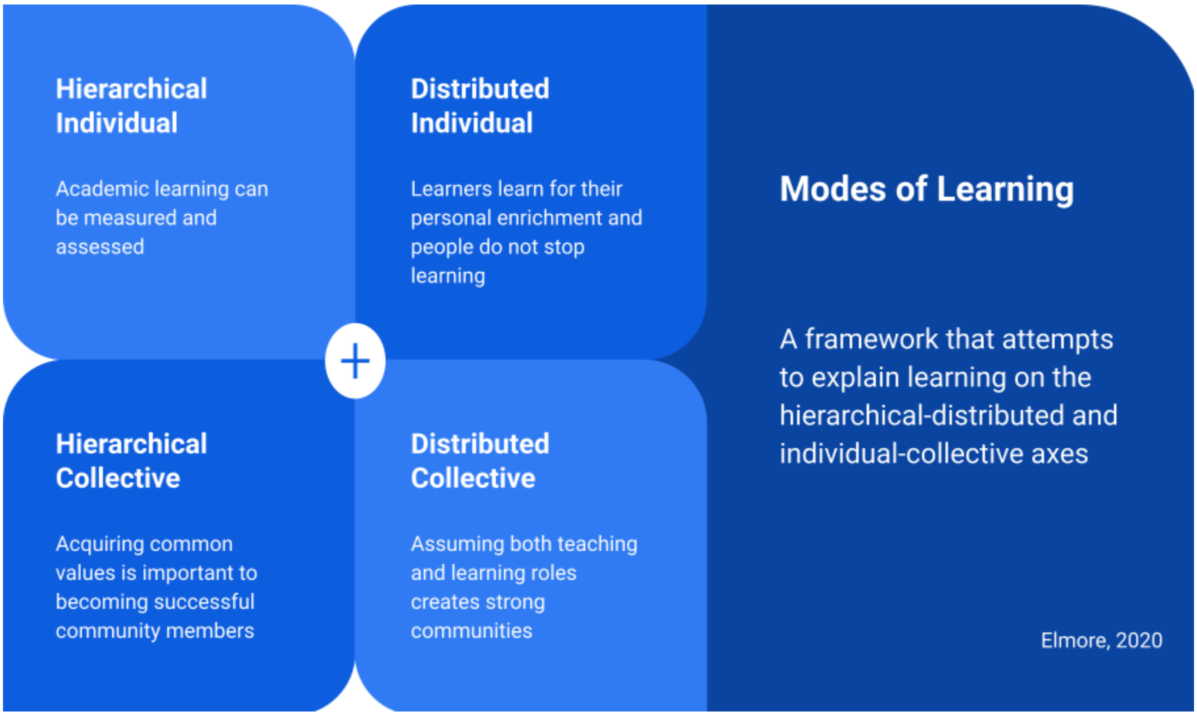


Figure 19: Modes of Learning. A framework to understand learning based on how learning is acquired (hierarchical or distributed) and who is the focal point of learning (individual or collective).

Another popular adaptive learning approach is recommender systems, or information filtering systems driven by individual user preferences. Outside the academic setting, most people are familiar with recommender systems through e-commerce platforms, where shoppers are shown suggestions on what they should check out next based on the behaviors of previous similar shoppers. Recommender systems in educational settings were shown to be usable at micro and macro levels. Micro-levels involve making small decisions, such as helping learners decide which module to study next based on what previous successful learners did[97]. Macro-levels involve making far-reaching decisions, such as the entire learning experience being personalized at the onset through an adaptive learning management system (LMS) based on a skills assessment conducted before a course is started[98].

One of the classic adaptive learning approaches is spaced repetition systems, where the psychological concept of spacing effect is utilized[99]. Spaced repetition refers to information being encoded in long-term memory through spaced study sessions instead of cramming. This is because aside from the need to account for the time and experience to learn (learning curve), we must also take into consideration that forgetting happens through time[100]. A step-up to spaced repetition is knowledge tracing, where the learner's mastery, which

[96] *The Office of Research Integrity. (nd). The lab, [Online]. Available:* `https://ori.hhs.gov/content/thelab`.

[97] *Z. A. Pardos, S. Tang, D. Davis,* et al., *"Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework," in* Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, *2017, pp. 23–32.*

[98] *S. Jagadeesan and J. Subbiah, "Real-time personalization and recommendation in Adaptive Learning Management System,"* Journal of Ambient Intelligence and Humanized Computing, *pp. 1–11, 2020.*

may not be readily visible, is estimated and used to decide the learning path[101].

With knowledge tracing, it is not just the last time and the number of times a question was answered correctly that matters in determining learner mastery for the targeted knowledge component. It also anticipates latent factors such as mistakes, guessing, or gaming-the-system behaviors. However, knowledge tracing is limited chiefly to understanding learner mastery only. It does not attempt to explore engagement and motivation levels as other adaptive learning systems do; hence it may be more limited in scope. This limitation makes knowledge tracing research, which is primarily on creating adaptive engines, advantageous by being more focused (e.g., less confounding variables) and more readily applicable on multiple platforms (i.e., once interoperability standards are established, adaptive engines can be usable with other learner and content models[102]).

[99] *P. Smolen, Y. Zhang, and J. H. Byrne, "The right time to learn: Mechanisms and optimization of spaced learning,"* Nature Reviews Neuroscience, *vol. 17, no. 2, p. 77, 2016.*

[100] *H. Ebbinghaus, "Memory: A contribution to experimental psychology,"* Annals of Neurosciences, *vol. 20, no. 4, p. 155, 2013.*

[101] *A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge,"* User Modeling and User-Adapted Interaction, *vol. 4, no. 4, pp. 253–278, 1994.*

[102] *Adaptive Instructional Systems (C/LT/AIS) P2247 Working Group. (2021). Adaptive instructional systems (c/lt/ais) p2247.1, [Online]. Available:* https://sagroups.ieee.org/2247-1/.

3.1 LATENT VARIABLE MODELS

There is more to learning than what meets the eye. In most educational environments, learning assessment activities using quizzes and homework are used to evaluate learners' knowledge and comprehension. However, a learner's performance in these activities can be influenced by factors other than learning. These can include the assessment material's quality, a learner's environmental conditions, or emotional state during the assessment. Therefore, researchers have been using latent variable models to reveal attributes hidden in observable phenomena. Latent variable models are statistical models that attempt to relate observable variables with non-observable or latent variables. For instance, Item Response Theory (IRT) is a popular latent variable modeling technique that uses learner responses to give insight into assessment item difficulty, learner ability, and learning estimate, among others[103].

[103] *B. Deonovic, M. Yudelson, M. Bolsinova,* et al., *"Learning meets assessment,"* Behaviormetrika, *vol. 45, no. 2, pp. 457–474, 2018.*

Because of IRT 's ability to differentiate learners, it has been exploited by intelligent tutoring system (ITS) to introduce adaptive learning. However, since IRT is mainly used to assess the quality of the evaluation, it has become inherently cross-sectional (taking a snapshot of learning states) as opposed to being longitudinal (being able to track the progression of learning), which is more suitable for ITSs. Hence, knowledge tracing algorithms that decide whether the learner needs more exercises to master a module or can already move on to succeeding modules are developed to fit ITSs better. Knowledge tracers are also latent variable models that treat learner mastery as its latent variable and learner performance, along with other variables, as its observables.

Knowledge tracing enables personalized learning as the pacing for each learner is adapted according to their abilities. However, despite ensuring the learners were exposed to sufficient exercises, there is a possibility that the learning is still not robust even with knowledge tracers[104]. For example, the learner may only master the skill but have difficulty applying the current learning to future learnings, or so-called shallow learning. Developing metacognitive skills can be crucial to overcoming shallow learning[105].

3.2 METACOGNITION AND ADAPTIVE LEARNING

As a recall, metacognition, or the knowledge and regulation of one's thinking process, includes skills such as goal setting and knowledge monitoring, among others[106]. This can be seen through various manifestations. Examples include learners realizing that they do not understand the topic enough to explain it in their own words. Another is when learners decide to create to-do lists to help them organize their learning activities. Multiple studies have shown that metacognition contributes to learners' academic performance and improves their learning[107]. This is even more important with the emergence of online learning, which might be here to stay long after the needs for social distancing measures imposed in the face of the COVID-19 pandemic are no longer needed[108]. Moreover, metacognitive skills allow learners to calibrate their learning and are better learning predictors in online learning environments than other factors such as time spent on assignments[109].

However, creating a tutoring system that teaches metacognitive skills to students is challenging. Training for metacognition is only practical when done in context, such as when learning a cognitive domain-specific skill (e.g., mathematics, language, and others) alongside. This puts a strain on the learners' cognitive resources[110]. The learners must spend effort on gaining metacognitive skills on top of learning in the cognitive domain. Fortunately, research on applying adaptive learning to metacognitive instruction already exists[111]. Research studies show that shallow learning (learning not being deep enough for the learner to apply in another context than where it was taught) could be addressed by metacognitive tutoring in a cognitive tutor[112]. Nevertheless, research on adaptive learning for metacognitive instruction alongside cognitive instruction is yet to be conducted.

When metacognitive instruction is done alongside cognitive domain instruction, learners might concentrate more on mastering the cognitive domain content. Cognitive domain learning will be more visible to the learners through markers such as higher grades, making it more critical for them. Metacognitive development will be harder to see, especially when the learners cannot apply their learning outside the tutoring environment. As such, developing metacognitive skills can be easier to take for granted when cognitive resources seem to be just enough for the cognitive part. What remains to be investigated is

[104] *R. S. Baker, S. M. Gowda, A. T. Corbett*, et al., *"Towards automatically detecting whether student learning is shallow," in* International Conference on Intelligent Tutoring Systems, *Springer, 2012, pp. 444–453.*

[105] *V. Aleven and K. R. Koedinger, "An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor,"* Cognitive Science, *vol. 26, no. 2, pp. 147–179, 2002.*

[106] *J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry,"* American Psychologist, *vol. 34, no. 10, p. 906, 1979.*

[107] *K. Ohtani and T. Hisasaka, "Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance,"* Metacognition and Learning, *vol. 13, no. 2, pp. 179–212, 2018.*

[108] *S. Gallagher and J. Palmer, "The pandemic pushed universities online. The change was long overdue,"* Harvard Business Review, *2020.*

[109] *L. Zhao and C. Ye, "Time and performance in online learning: Applying the theoretical perspective of metacognition,"* Decision Sciences Journal of Innovative Education, *vol. 18, no. 3, pp. 435–455, 2020.*

how to combine metacognitive tutoring and cognitive adaptive learning to manage cognitive resources.

Bayesian Knowledge Tracing (BKT) is a versatile, adaptive learning algorithm to which several researchers have previously introduced modifications. Some examples include estimating the learner's prior knowledge based on the correctness of their first response[113], estimating a problem's difficulty in a traditional setting[114] and a massive open online course (MOOC) setting[115], individualizing prior knowledge and learning rate estimates[116], and even using brain scans as observation inputs[117]. Because of its versatile nature, we are interested in how well can the BKT be modified to use metacognitive indicators in knowledge tracing.

Artificial neural networks (ANN) is another set of algorithms that is recently gaining traction among adaptive learning researchers. Artificial neural networks are typically composed of input and output layers connected by hidden layers. A particular interest is in using deep learning or neural networks with more than a single hidden layer, which can significantly improve prediction accuracy if more data is available. Since input and output layers can be defined accordingly, artificial neural networks (ANNs) is also a plausible candidate for developing knowledge tracing with metacognitive tutors.

[110] *I. Roll, V. Aleven, B. M. McLaren,* et al., *"Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking,"* Metacognition and Learning*, vol. 2, no. 2, pp. 125–140, 2007.*

[111] *K. Agustianto, A. E. Permanasari, S. S. Kusumawardani,* et al., *"Design adaptive learning system using metacognitive strategy path for learning in classroom and intelligent tutoring systems," in* AIP Conference Proceedings, *AIP Publishing LLC, vol. 1755, 2016, p. 070 012.*

[112] *R. S. Baker, S. M. Gowda, A. T. Corbett,* et al., *"Towards automatically detecting whether student learning is shallow," in* International Conference on Intelligent Tutoring Systems, *Springer, 2012, pp. 444–453.*

[113] *Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a Bayesian networks implementation of knowledge tracing," in* International Conference on User Modeling, Adaptation, and Personalization, *Springer, 2010, pp. 255–266.*

3.3 RELATED WORK: EXISTING IMPLEMENTATIONS

We envision the Adaptive Engine of Personalized Online Adaptive Learning System (POALS) to be a knowledge tracer that works with a metacognitive tutor on top of a cognitive tutor. To construct the Adaptive Engine, we need to revisit the mechanism behind the Metacognitive Tutor and existing knowledge tracing algorithms.

3.3.1 *Reflection Assistant (RA)*

The RA discussed in Chapter 2 THE METACOGNITIVE TUTOR did account for the learners' cognitive load by introducing the metacognitive tutoring during the preparation and verification phases. Metacognitive tutoring is not conducted during the actual problem-solving phase, where the demand for cognitive resources is expected to be highest. However, its metacognitive tutoring is still on top of cognitive tutoring. Thus, while the cognitive load may not be as much as other metacognitive tutors integrated with cognitive tutors, it is still an additional burden to the learners.

3.3.2 *Bayesian Knowledge Tracing (BKT)*

BKT is a knowledge tracing algorithm that is a hidden Markov model (HMM) where learner knowledge is represented as a binary variable (whether a knowledge component is mastered or not) for each knowledge component[118]. A graphical representation of BKT is shown in Figure 20.

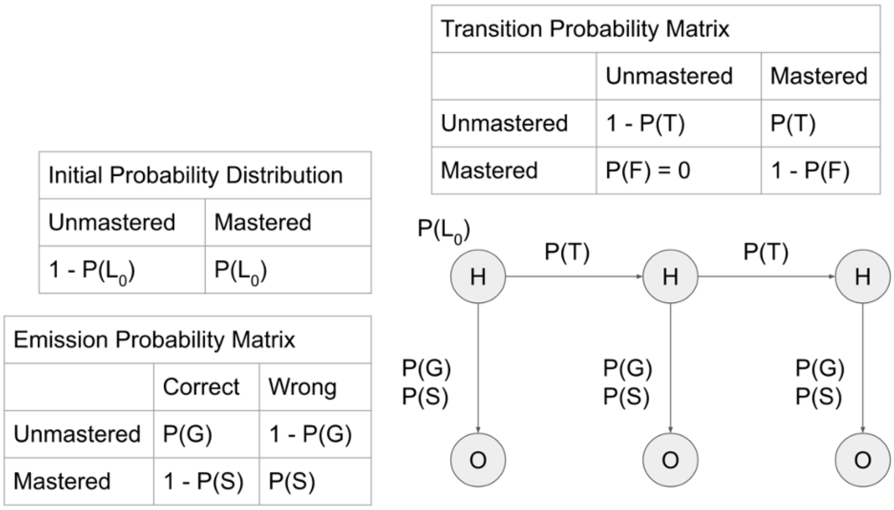


Figure 20: Bayesian Knowledge Tracing. Graphical representation of BKT as an HMM.

An HMM is composed of the following components:

- A set of hidden **states**. For BKT, the states are whether the knowledge component is **mastered** or **unmastered**.
- A **transition probability matrix** that indicates the probability of transitioning from one state to another (e.g., from unmastered to mastered).
- An **initial probability distribution** that indicates the probability of starting at a hidden state (e.g., when the learner has prior knowledge).
- A sequence of **observations** drawn from a vocabulary. For BKT, the vocabulary includes whether the answer is $\mathtt{correct} = 1$ or $\mathtt{wrong} = 0$.
- An **emission probability matrix** that indicates the probability of an observation being generated from a hidden state (e.g., when the learner answers correctly by guessing).

As an HMM, the following assumptions are held:

- The probability of a particular hidden state depends on the previous hidden state only.
- The observations are conditionally independent of all other variables given their current hidden state.

[114] *Z. A. Pardos and N. T. Heffernan, "KT-IDEM: Introducing item difficulty to the knowledge tracing model," in* International Conference on User Modeling, Adaptation, and Personalization, *Springer, 2011, pp. 243–254.*

[115] *Z. A. Pardos, Y. Bergner, D. T. Seaton,* et al., *"Adapting Bayesian knowledge tracing to a massive open online course in edX," in* Sixth International Conference on Educational Data Mining, *International Educational Data Mining Society, 2013, pp. 137–144.*

[116] *M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," in* International Conference on Artificial Intelligence in Education, *Springer, 2013, pp. 171–180.*

[117] *D. Halpern, S. Tubridy, H. Y. Wang,* et al., *"Knowledge tracing using the brain,"* International Educational Data Mining Society, *2018.*

[118] *A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge,"* User Modeling and User-Adapted Interaction, *vol. 4, no. 4, pp. 253–278, 1994.*

The BKT fits the following knowledge parameters, which are used in the HMM 's transition probability matrix and initial probability distribution:

- $p(L_0)$ *Initial Learning*: Probability that the knowledge component is already mastered even before the first opportunity to solve a problem is presented,
- $p(T)$ *Acquisition*: Probability that the knowledge component is mastered from solving the problem, and
- $p(F)$ *Forget*: Probability that the knowledge component was previously mastered but is not currently mastered. This is traditionally set to 0 and is not included among the calculated parameters.

Additionally, the following performance parameters are also fitted, which are used in the HMM 's emission probability matrix:

- $p(G)$ *Guess*: Probability that the knowledge component is not yet mastered, but the learner was able to apply it correctly on the problem, and
- $p(S)$ *Slip*: Probability that the knowledge component is already mastered, but a mistake was made when applying it to the problem.

The correctness of the learner's response at opportunity n can be predicted with Equation 3.

$$p(Correct_n) = p(L_n) * p(\neg S) + p(\neg L_n) * p(G) \tag{3}$$

The probability that a knowledge component is mastered given that the problem is correctly answered is usually inferred using Equation 4.

$$p(L_n|correct) = \frac{p(L_n) * p(\neg S)}{p(L_n) * p(\neg S) + p(\neg L_n) * p(G)} \tag{4}$$

When the answer is wrong, Equation 5 is used instead.

$$p(L_n|wrong) = \frac{p(L_n) * p(S)}{p(L_n) * p(S) + p(\neg L_n) * p(\neg G)} \tag{5}$$

Equation 6 gives the probability that the knowledge component is then mastered on the next problem.

$$p(L_n) = p(L_{n-1}|obs_{n-1}) + p(\neg L_{n-1}|obs_{n-1}) * p(T) \tag{6}$$

Here, obs_{n-1} is the observation (correct or wrong) at opportunity $n-1$.

One risk that comes with introducing *Guess* and *Slip* parameters that offer counter-intuitive explanations for observed behavior is the possibility of *model degeneracy*: the resulting model may not behave as it was intended to be[119]. For knowledge tracing, model degeneracy occurs when the link between learner knowledge and learner performance is lost. For example, despite the learner performing well on problems, the model still predicts the learner has not mastered the knowledge component because *Guess* is given more weight than *Acquisition*. Model degeneracy is likely to occur when either $p(S)$ or $p(G)$ is greater than 0.5.

Model degeneracy can be rare in practice. However, several factors, such as confusingly worded questions, can result to model degeneracy[120]. A measure to avoid model degeneracy is bounding $p(S)$ and $p(G)$ to a small range of values, with some choosing to fix the values for said parameters. However, this exposes the problem of deciding the best values for $p(S)$ and $p(G)$ as their values will affect the other parameters. Also, fixing values for $p(S)$ and $p(G)$ deprives the chance to investigate the factors causing model degeneracy.

[120] *S. Doroudi and E. Brunskill, "The misidentified identifiability problem of Bayesian knowledge tracing,"* International Educational Data Mining Society, *2017.*

3.3.3 *Knowledge Tracing with Deep Learning*

Artificial neural networks such as the one shown in Figure 21 had been drawing interest among researchers since Deep Knowledge Tracing (DKT) using recurrent neural networks (RNNs) was introduced[121]. ANNs are conceptually derived from biological neurons, where neurons have inputs that can produce outputs through activation and are passed on to other neurons[122].

[121] *C. Piech, J. Bassen, J. Huang,* et al., *"Deep knowledge tracing," in* Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, *2015, pp. 505–513.*

[122] *W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity,"* The Bulletin of Mathematical Biophysics, *vol. 5, no. 4, pp. 115–133, 1943.*

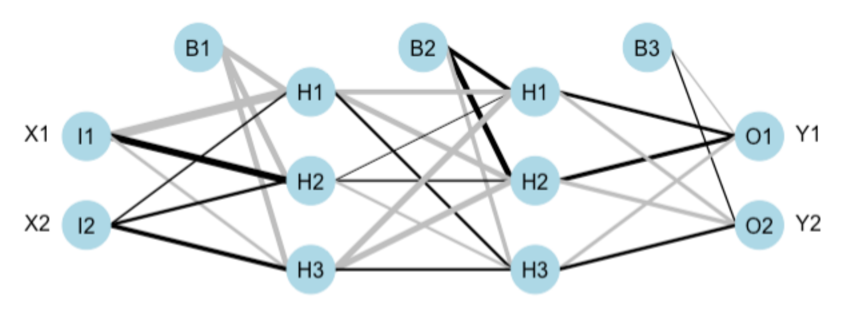


Figure 21: Deep Neural Network DNN. A sample showing a DNN as a network of nodes in multiple layers.

An ANN can be characterized by the following:

- It has an **input layer** with at least one node. In Figure 21, *I1* and *I2* correspond to the nodes of the input layer.
- It has an **output layer** with at least one node. In Figure 21, *O1* and *O2* correspond to the nodes of the output layer.
- It has at least **one hidden layer**. In Figure 21, there are two hidden layers, with each of the hidden layers having three nodes.

A neural network learns through each layer; thus, more layers (i.e., deeper) may be better models if sufficient data is used for training. A sample heuristic being used is about tens of individual samples for each parameter to be estimated, essentially the weights.

- Nodes are connected with synapses associated with **weights**. In Figure 21, the weights were visualized with the connecting lines' thickness, where black lines are positive weights, and gray lines are negative weights. During the training of a neural network model, the combination of nodes and layers is used as a starting point. The corresponding weights (and, optionally, the bias values) are calculated through adjustments. Passing the inputs through the weighted nodes and activation function is closest to the expected output.
- When predicting an output given an input, the connections' weighted sums are passed through an **activation function**. The activation function decides whether a node should fire or not; hence its return values are mostly either just 0 or 1. An example of this is the sigmoid function illustrated in Figure 22.
- A **bias** value may also be needed to shift the activation function. In Figure 21, *B1*, *B2*, and *B3* are the bias values.

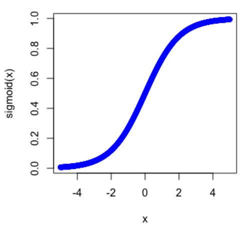


Figure 22: Sigmoid Function. This is a typical choice for activation function because of its ability to separate 0s and 1s smoothly.

Backpropagation, an application of chain rule for minimizing the error function, is used to derive the weight values given the activation function. Backpropagation executes weights finetuning through gradient descent (or ascent, depending on how the optimization problem is formulated). Gradient descent is taking iterative steps in the direction of the derivative or gradient, following the steepest slope.

The **error function** calculates the difference between the model's predicted outputs $predicted_i$ and the actual outputs $observation_i$. **Forward propagation** is used to compute for the predicted outputs, where inputs are fed to the network passing through the hidden layers until the output is derived. An example of an error function is the **mean square error**, which can be expressed as Equation 7, where N = number of samples.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(observation_i - predicted_i)^2 \qquad (7)$$

3.4 PROPOSED MECHANISM FOR POALS ADAPTIVE ENGINE

A system where metacognitive tutoring is done on top of cognitive tutoring while employing adaptive learning is proposed. Figure 23 illustrates such a system, where LMS hosts the cognitive tutor and RA as the metacognitive tutor. An adaptive engine influences how the learner interacts with the LMS based on the metacognitive inputs by using knowledge tracing.

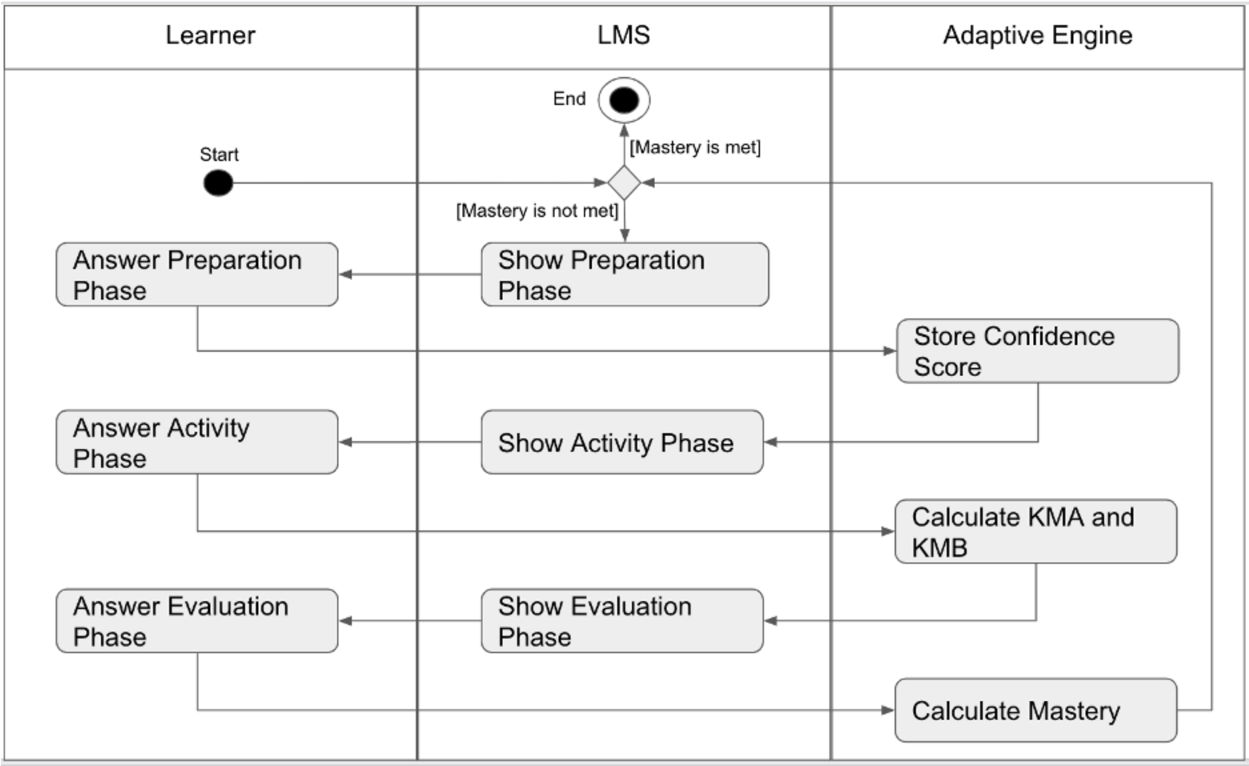


Figure 23: Adaptive Engine Activity Diagram. Envisioned activity diagram for RA model usage with an adaptive engine.

3.4.1 *BKT with RA: RA-BKT*

There are myriad reasons a learner's response might be attributed to a guess or slip. For example, simple errors, fatigue, or the learner giving up due to frustration with the way a problem is worded despite mastering the associated knowledge may lead to slips. On the other hand, guesses can be attributed to sheer chance, assessment tool errors, or confusing a related theory with the intended theory and somehow still arriving at the correct answers.

When combining RA and BKT, it may be thought that when the learner has high Knowledge Monitoring Accuracy (KMA) and predicted that they could answer the problem but was not able to do so, then it is more likely that the incorrect answer was due to a slip. It is probably not due to the learner not being able to learn the knowledge component yet. Similarly, the case when the learner who has an optimistic Knowledge Monitoring Bias (KMB) and a low KMA predicted that they could answer the problem and were indeed able to answer correctly must be scrutinized further. The learner already gaining mastery cannot be quickly assumed because it is likely that they simply guessed the answer correctly.

It is tempting to include the KMA and KMB in the formulation of the BKT. The most straightforward way is adding the KMA and KMB values (or their resulting classifications) in the observation vocabulary. However, since KMA and KMB are cumulative values across opportunities, adding them to the observation vocabulary would violate the HMM assumption that observations are conditionally independent.

The confidence self-reports and performance are combined to form the observation vocabulary that preserves the conditional independence of the observations. The original observation vocabulary is $\{1, 0\}$, which corresponds to the possible Answer values. The new observation vocabulary is $\{c\_0, p\_0, i\_0, c\_1, p\_1, i\_1\}$ where c, p, and i corresponds to learners predicting to answer correctly, partially correctly and incorrectly respectively. For the performance, 0 and 1 correspond to learners answering the problem incorrectly and correctly, respectively. This BKT reconstruction is referred to as RA-BKT.

The new emission probability matrix for RA-BKT is shown in Table 11. While BKT has four parameters, RA-BKT will have 14 parameters. Since the hidden states are not changed, the knowledge parameters which make up the probability transmission matrix will be the same. On the other hand, the performance parameters will be replaced by the new emission probability matrix items.

Table 11: RA-BKT Emission Probability Matrix. Mapping of probabilities of hidden states given observation markers.

State	Observation					
	c_0	p_0	i_0	c_1	p_1	i_1
Unmastered	$P(U_{c\_0})$	$P(U_{p\_0})$	$P(U_{i\_0})$	$P(U_{c\_1})$	$P(U_{p\_1})$	$P(U_{i\_1})$
Mastered	$P(M_{c\_0})$	$P(M_{p\_0})$	$P(M_{i\_0})$	$P(M_{c\_1})$	$P(M_{p\_1})$	$P(M_{i\_1})$

The RA allows for being able to answer the problem partially correctly. But for parallelism with the standard BKT, we will not be considering partially correct answers. Therefore, computations for KMA and KMB will still be the same, even if there are no partially correct answers. From here on, KMA and KMB will be referred to as learner awareness and outlook, respectively.

3.4.2 *ANN with RA: RA-ANN*

The formulation derived above still did not allow taking advantage of the KMA and KMB values in predicting whether mastery is achieved or not. ANNs do not have the conditionally independent restriction on inputs as in BKT. Therefore, KMA and KMB may be used alongside the confidence reports for our input. Neural networks that use the number of chances given to the learner (i.e., opportunities) and another that uses all other inputs were created to parallel comparisons with BKT. These are labeled ANN and RA-ANN respectively. The neural networks are constructed as classification problems: that is, the neural networks will predict whether the learner will answer the problem correctly or not. This is summarized in Table 12.

Table 12: Neural Network Input and Output Layers. Additional inputs are provided for RA-ANN.

Layer	ANN	RA-ANN
Input	Opportunity	Opportunity
		Predicts to answer correctly*
		Predicts to answer incorrectly *
		KMA
		KMB
Output	Answer	Answer

* *Predicts to answer partially correctly* is no longer used as input as it can be inferred from these inputs.
These are one-hot encodings of the confidence self-report.

3.5 BUILDING THE ADAPTIVE ENGINE

The following questions are to be answered:

- How do the models compare with each other in terms of training efficiency and accuracy?
- Which model might be the best in reducing cognitive load?
- How closely will the models follow learning intuitions (e.g., model degeneracy; the relationship between awareness, outlook, and mastery)?

3.5.1 *Dataset*

A dataset was created for this experiment by defining learner personas. Table 13 lists possible learner behaviors based on their performance (8 behaviors) and confidence report (13 behaviors). The performance and confidence report behaviors are combined, resulting in 104 learners ($8 * 13 = 104$).

In this dataset, each learner has ten opportunities to answer the problem or demonstrate learning the component. This formulation's imagined setup is a quiz with ten items for a knowledge component that the learner answers once. Alternatively, this can also mean a single problem that the learner can attempt to answer up to ten

Table 13: Assumed Learner Behaviors. The experimental dataset was created based on personas exhibiting these assumed learner behaviors.

Performance	Confidence Report
· Always answers correctly	· Always predicts to answer correctly
· Always answers incorrectly	· Always predicts to answer partially correctly
· Occasionally answers correctly	· Always predicts to answer incorrectly
· Occasionally answers incorrectly	· Always predicts to answer correctly but occasionally predicts to answer partially correctly
· Progressively performs better	· Always predicts to answer correctly but occasionally predicts to answer incorrectly
· Regresses in performance	· Always predicts to answer partially correctly but occasionally predicts to answer correctly
· Progressively performs better then regresses in performance	· Always predicts to answer partially correctly but occasionally predicts to answer incorrectly
· Regresses in performance then progressively performs better	· Always predicts to answer incorrectly but occasionally predicts to answer correctly
	· Always predicts to answer incorrectly but occasionally predicts to answer partially correctly
	· Progressively improves in prediction
	· Regresses in prediction
	· Progressively improves then regresses in prediction
	· Regresses then improves in prediction

times. In most cases, the situation will be somewhere in between the said scenarios. BKT 's assumption that learners gain more knowledge the more opportunities they are given (supposing the learner exerts honest effort to learn and is not merely gaming the system) is to be followed. Having ten opportunities is reasonable without being overbearing if the first interpretation (a ten-item quiz for every knowledge component) is applied.

The following data were also created to prevent model degeneracy:

- The performance behavior where the learner progressively improves is repeated. Data were created such that the learner initially gets the answer wrong and consistently gets the answer correctly afterward. In our dataset, each learner has ten opportunities. The improving performance condition is repeated for each opportunity count, where the learner starts to get the answer correctly from opportunity N, with N between 2 to 10. Combining these nine new conditions with the 13 confidence report behaviors adds 117 more data ($13 * 9 = 117$).

- The same is done for confidence report behavior, where the learner gets the prediction right from opportunity N, with N between 2 to 10. The confidence report allows for partially correct prediction. For example, for $N = 2$, there is the case where the learner is predicting either incorrectly or partially correctly during the first opportunity. Hence, this condition is repeated ten times instead of just nine. Other than for $N = 2$, the learner is set to have a partially correct prediction for opportunities above $\texttt{floor}(N/2)$ but below N. After combining with the initial eight performance behaviors, 80 additional learner data were created.

- Additionally, the combination of improving performance prediction (nine times) and improving confidence prediction (ten times) is repeated. This added 90 learner data, finally resulting in 391 ($104 + 117 + 80 + 90 = 391$) learner data.

In an actual learning scenario, the learners will be exposed to more than one knowledge component. However, for BKT, each knowledge component is modeled separately; each knowledge component model does not affect other knowledge components. While it may be argued that related knowledge components could be affecting each other, this case could be handled by estimating for the initial mastery. Having more than one knowledge component is thus inconsequential. Hence, only data for a single knowledge component is created.

The 391 learners were randomly assigned to five groups, with four groups combined as model training dataset and the remaining group reserved as validation dataset. The resulting division between training and validation sets is detailed in Table 14, where the distributions across key factors are relatively even. Thus, unbalanced data, which is typically an issue in machine learning, is not a significant concern for this research. With 391 learner data having ten opportunities each, there is a total of 3910 observation records.

The standard deviation and unlikeability coefficient[123] of the Answer and Confidence values are also shown to indicate the efforts made to make the training and validation sets as identical as possible. A locally estimated scatterplot smoothing (LOESS) fitting was also done for Opportunity (independent variable) and Answer (dependent variable), and the residual error was obtained as an additional measure for variability.

[123] G. D. Kader and M. Perry, *"Variability for categorical variables,"* Journal of Statistics Education, *vol. 15, no. 2, 2007.*

Table 14: Modeling Data Distribution. Relative data balance was attempted to be achieved.

Description	Training	Validation
Data count	3120 (79.79%)	790 (20.20%)
Answer (standard deviation)	0.500	0.499
· 0	1553 (49.77%)	366 (46.32%)
· 1	1567 (50.22%)	424 (53.67%)
Confidence (Unalikeability coefficient)	0.645	0.666
· C (answer correctly)	1391 (44.58%)	279 (35.31%)
· P (answer partially correctly)	770 (24.67%)	275 (34.81%)
· I (answer incorrectly)	959 (30.73%)	246 (31.13%)
LOESS Residual standard error	0.472	0.454

The validation set appears marginally less varied than the training set, as seen from the Answer's standard deviation and the LOESS residual error. This could have led to lower training accuracy than testing accuracy that can be seen in Section 3.5.2.

During test data creation, no assumptions were made about the relationship between the metacognitive measures and their mastery level. This is in line with the general assumption that metacognition is a domain-independent skill[124]. Based upon the above information, the following columns were created for the dataset:

[124] V. Aleven and K. R. Koedinger, *"Limitations of student control: Do students know when they need help?" In* International Conference on Intelligent Tutoring Systems, *Springer, 2000, pp. 292–303.*

- **learner**: An identifier for the learner created using personas.
- **skill**: The knowledge component; currently, only 1 knowledge component is created with the assigned value "A". This information is not used for the current experiment.
- **opportunity**: An opportunity to answer a question; an opportunity results to one observation. Currently, 10 opportunities are created for each learner.
- **confidence**: The confidence self-report provided by the learner in the POALS Metacognitive Tutor preparation phase; possible

values are C, P, and I which correspond to predicting to be correct, partially correct, and incorrect, respectively.

- **answer**: The performance; possible values are $0 = wrong$ and $1 = \mathtt{correct}$.
- **awareness**: The KMA score. This was computed as detailed in 2.2.1 The Reflection Assistant.
- **outlook**: The KMB score. This was computed as detailed in 2.2.1 The Reflection Assistant.
- **train**: Whether the data will be used for training (0) or validation (1).

The trend between opportunities and correctness, awareness, and outlook was visualized using generalized linear modeling to ascertain how realistic the synthetic data is. The upward trend made in the assumption is visible in Figure 24. First, both awareness and outlook values were adjusted to limit the range of values to $[0, 1]$, with 1 being the best value, just like correctness. Next, the awareness was normalized to change the range of values from $[-1, 1]$ to $[0, 1]$. Then, the corrected outlook was set to $1 - |value|$ since the original outlook values are also in the $[-1, 1]$ range, with 0 being the desirable value. This way, all values we are interested in are presented in the same scale $[0, 1]$ and desired increasing progression.

The resulting trends were compared with the data from ASSISTments (2015) and the Geometry Angles dataset. The ASSISTments is a widely used platform for learning research[125]. The Geometry Angles dataset resulted from research on metacognition[126]. The attribute "Opportunity" was introduced to both datasets after sorting by Log ID attribute for ASSISTments and ID for Geometry Angles to ensure parallel comparison. The Opportunity is then incremented, starting from 1 for each Student and Problem combination. Finally, the data are further filtered such that only the Student and Problem combinations with exactly ten Opportunity counts are included.

Both ASSISTments and Geometry Angles exhibit the same upward trend, thus matching the trends of the synthetic dataset. Furthermore, we see an upward trend for awareness and outlook, which follows the intuition that with more practice, the more the metacognitive ability becomes apparent. Just like what was done during the POALS Metacognitive Tutor spotcheck, the p-values resulting from the linear modeling used for trend analysis were also calculated. When using the typical cut-off of 0.05 for rejecting the null hypothesis, we can say that statistical significance is observed.

[125] *N. T. Heffernan and C. L. Heffernan, "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching,"* International Journal of Artificial Intelligence in Education, *vol. 24, no. 4, pp. 470–497, 2014.*

[126] *V. Aleven and K. R. Koedinger, "Limitations of student control: Do students know when they need help?" In* International Conference on Intelligent Tutoring Systems, *Springer, 2000, pp. 292–303.*

3.5.2 *Modeling*

For training the BKT variants, an existing R code for BKT[127] was modified to accommodate the RA-BKT. The correspondence between the BKT and RA-BKT parameters is shown in Table 15. The Baum-Welch algorithm, a special case of the expectation maximization (EM) algorithm that uses the forward-backward algorithm for the expectation step[128], was used for model fitting. In EM, the parameters are first initialized randomly, and the expected values of observations are computed based on the parameter values[129]. The expected values are compared against the actual observations, and the parameters are re-calibrated based on the comparison. A threshold of 1×10^{-9} and a maximum step of 100 is set. The iteration continues until the difference between the previous and the current results is less than the threshold or when 100 iterations are met. A seed is set for the pseudo-random number generator to ensure reproducibility.

[128] *L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes,"* Inequalities, *vol. 3, no. 1, pp. 1–8, 1972.*

[129] *A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm,"* Journal of the Royal Statistical Society: Series B (Methodological), *vol. 39, no. 1, pp. 1–22, 1977.*

Table 15: BKT and RA-BKT Correspondence. This information is useful for understanding model degeneracy later on.

BKT	$\leftrightarrow$	RA-BKT
$p(L_0)$	$\leftrightarrow$	$p(L_0)$
$p(T)$	$\leftrightarrow$	$p(T)$
$p(G)$	$\leftrightarrow$	$p(U_{c\_1}) + p(U_{p\_1}) + p(U_{i\_1})$
$p(S)$	$\leftrightarrow$	$p(M_{c\_0}) + p(M_{p\_0}) + p(M_{i\_0})$
$1 - p(G)$	$\leftrightarrow$	$p(U_{c\_0}) + p(U_{p\_0}) + p(U_{i\_0})$
$1 - p(S)$	$\leftrightarrow$	$p(M_{c\_1}) + p(M_{p\_1}) + p(M_{i\_1})$

A ten-fold cross-validation repeated five times was used for model training using the training dataset previously defined. Each iteration involved randomly distributing the dataset into ten bins, followed by ten training and testing rounds. A different bin is used for testing each round, while the remaining bins are used for training. Using cross-validation allows for more reliable results than having a single train and test set since it is less likely for the results to be just due to a convenient test/train data split. Additionally, at the beginning of each repeat, the parameters are once again re-initiated randomly. As an EM algorithm, the Baum-Welch algorithm can maximize a function with multiple peaks or optima, such as the one shown in Figure 25. However, the global optimum may not be reachable depending on the initial parameters. Therefore, it is necessary to have different initial parameters to ensure that the global optimum is found and not just a local optimum.

***Gradient methods**, or methods where optima are searched by looking at the gradient or slope of the current point's surrounding and moving in the direction of the best slope, is typically used for optimization problems.*

For each repeat, the model is taken to be the average of the parameters in each iteration. The model corresponding to the repeat with the highest average cross-validation accuracy is selected. Equation 8 gives the accuracy based on the contingency mapping in Table 16.

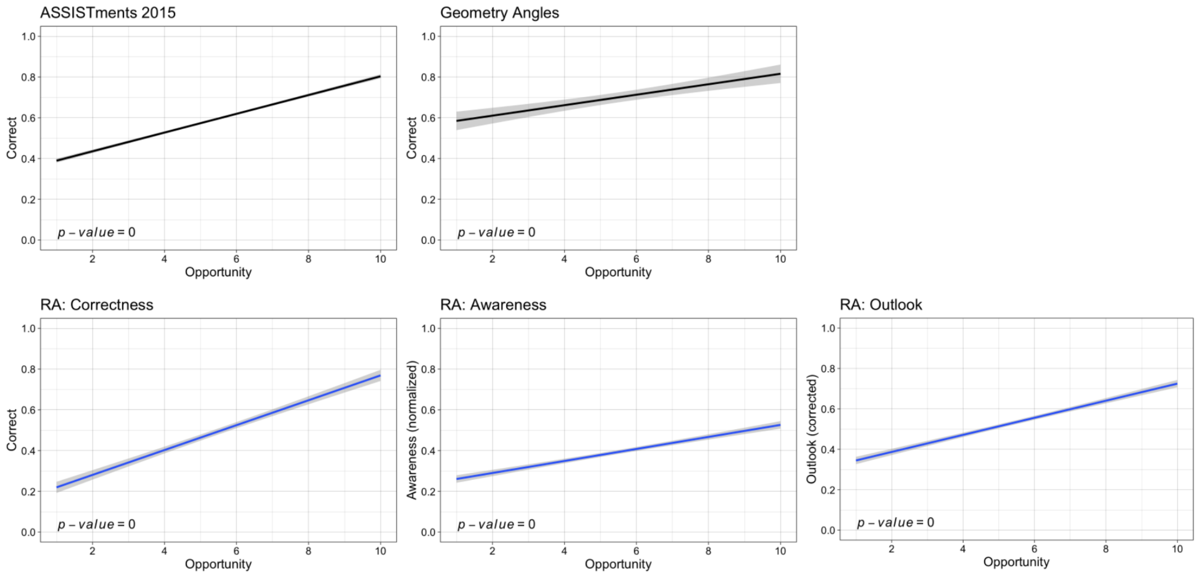


Figure 24: Other Datasets and RA Dataset Comparison. Trend comparison between ASSISTments, Geometry Angles, and RA's correctness, awareness, and outlook was used to validate the created dataset.

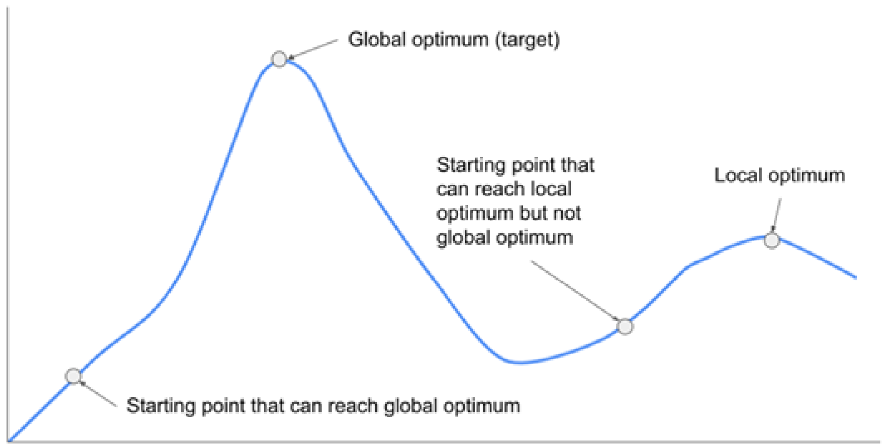


Figure 25: Multiple Optima. Depending on the initial parameter set, using gradient methods may not lead to the global optima.

Here, the first model is the actual observation and the second model is the trained model.

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} \tag{8}$$

When validating, typically, the resulting model is used to predict the associated observation. However, the observation vocabulary for the RA-BKT is different from the other models. The probabilities for $\{c\_0, p\_0, i\_0\}$, and $\{c\_1, p\_1, i\_1\}$ are summed up as analogs to $\{0, 1\}$ when validating RA-BKT to be able to make a parallel comparison between RA-BKT and the other models. For all other models, whether the observation corresponds to the answer is correct is predicted.

Table 16: Contingency Table Mapping. This is used for calculating model accuracy and statistical difference across models.

		First Model's Prediction	
		Correct	Wrong
Second Model's	Correct	A	B
Prediction	Wrong	C	D

The R packages **deepnet**[130] and **caret**[131] were used for training the DL models. The **caret** package does the repeated cross-validation described earlier automatically. Additionally, **caret** searches for the best combination of the number of nodes and layers using the same accuracy formula mentioned earlier.

The search space included up to three layers (thus, DL) to tap the potential of higher accuracy rates. Each layer is set to possibly have 0, 3, 5, 7, or 10 nodes (0 not included for the first layer to ensure there is at least one layer). In addition, candidate hidden dropouts (0 and 0.1) were also set. Dropout involves randomly ignoring neurons during training to prevent overfitting or the resulting model being too specific to the training data[132]. This makes the entire search space have a size of 200 (four options for the first layer × five options for the second layer × five options for the third layer × two options for hidden dropouts).

The validation dataset predictions were used to compute the testing accuracy and the statistical difference between models using McNemar's test[133] with the same contingency table as in Table 16. The resulting statistic from McNemar's test, calculated in Equation 9, is said to follow the χ^2 distribution. For sample sizes more than 25, the statistical degree of freedom (different from the modeling degree of freedom) is typically assumed to be 1. The corresponding p-value is

[130] X. Rong, deepnet: Deep learning toolkit in r, *R package version 0.2, Mar. 2014. [Online]. Available:* `https://CRAN.R-project.org/package=deepnet`.

[131] M. Kuhn, *"Building predictive models in R using the caret package,"* Journal of Statistical Software, *vol. 28, no. 5, pp. 1–26, 2008.*

[132] *N. Srivastava, G. Hinton, A. Krizhevsky,* et al., *"Dropout: A simple way to prevent neural networks from overfitting,"* Journal of Machine Learning Research, *vol. 15, no. 1, pp. 1929–1958, 2014.*

[133] *Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages,"* Psychometrika, *vol. 12, no. 2, pp. 153–157, 1947.*

obtained from the readily available distribution table for χ^2 using the computed statistic from these assumptions.

$$\chi^2 = \frac{(B - C)^2}{B + C} \tag{9}$$

The null hypothesis is that none of the models predict better than the other. The p-value should be less than α, which was chosen to be 0.05 to follow conventions. When the p-value does not meet the said condition, we cannot reject the null hypothesis (i.e., that the model with better accuracy scores is better).

For knowledge tracing, mastery prediction is more important than correctness prediction. The statistical difference of mastery predictions between models is also calculated with the Mann-Whitney U test. This test was selected because it does not require the items to be normally distributed and allows for comparing continuous variables, unlike McNemar's test[134]. The same assumptions for the null hypothesis as in McNemar's test are held.

[134] *H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other,"* The Annals of Mathematical Statistics, *pp. 50–60, 1947.*

3.6 RESULTS AND DISCUSSION

The training task was not resource-intensive, so a simple machine (1.4 GHz Quad-Core Intel Core i5 processor, 8 GB 2133 MHz LPDDR3 memory, Intel Iris Plus Graphics 645 1536 MB graphics) was sufficient. The R programming language (version 4.0.2) was used. A naïve model that always predicts that the learner will always answer correctly was used for baseline comparison. Based on Table 14, there are marginally more observations with correct answers for both training and validation sets. Thus, the baseline that always predicts answering correctly has better odds than random chance.

Table 17 shows the resulting training metrics. For accuracy checking, the predicted answers will be taken into account. The training and prediction times are also noted.

Description	Baseline	BKT	RA-BKT	ANN	RA-ANN
Training time (minutes)	-	10.356	11.537	12.429	13.929
Prediction time (seconds)	-	0.058	0.098	0.023	0.007
Training Accuracy	0.502	0.786	0.846	0.648	0.864

Table 17: Training Data Comparison. Models are compared based on their training time, execution time, and training accuracy.

Training for RA-BKT took longer than BKT; the same is also observed for the DL models. While this result is not desirable, it is still reasonable considering the hardware specifications used for training and the increase in parameters estimated. The RA models are expected to take longer since they have more inputs than the standard models. The DL models are also expected to take more time than the HMM models since the DL 's search space size is relatively larger.

In practice, the training will be done before use in tutoring software. Hence, long training times can be acceptable if it will not demand expensive hardware. However, it is crucial to keep the prediction time low to avoid lag when the mastery prediction is being used for deciding learning paths while the learners are using software using knowledge tracing models. The prediction times being small for all models is thus a positive result.

Figure 26 shows the resulting networks for ANN and RA-ANN. The resulting network for ANN is relatively deep with two layers, while the RA-ANN consisted of only a single hidden layer. However, the ANN has lower accuracy; with insufficient data, producing good deep networks may be difficult. It can be seen from the RA-ANN that the network does not have to be deep to model the data decently for some problems. Most other DL applications in knowledge tracing had more complicated problem construction than presented here. For instance, the learner's skill is also added as an input to potentially use the network to inform course developers which skills are related to each other[135].

[135] *C. Piech, J. Bassen, J. Huang,* et al., *"Deep knowledge tracing," in* Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, *2015, pp. 505–513.*

Table 18 clarifies that RA-ANN is the best performer based on accuracy, followed by RA-BKT. Note that, as previously pointed, DL models will require more data for training. Hence, if data is not enough, using RA-BKT is worth considering. Each of the models is also significantly different from the other in terms of correctness prediction except when comparing BKT with RA-BKT and ANN with RA-ANN. This means that the difference between the accuracy across models is not merely due to random chance except for BKT against RA-BKT and ANN against RA-ANN. With this information, we know that choosing between HMM-based algorithms and DL-based algorithms matter, but whether the standard or the RA models should be used is a toss-up.

Table 18: Model Differences. Testing accuracies (diagonal) and significant differences of correctness prediction using McNemar's test (upper triangle) and mastery prediction using Mann-Whitney U test (lower triangle) between models.

	Baseline	BKT	RA-BKT	ANN	RA-ANN
Baseline	0.463	< 0.001	< 0.001	< 0.001	< 0.001
BKT	< 0.001	0.768	0.745^{*}	< 0.001	< 0.001
RA-BKT	< 0.001	< 0.001	0.865	< 0.001	< 0.001
ANN	< 0.001	< 0.001	< 0.001	0.682	$< 0.842^{*}$
RA-ANN	< 0.001	< 0.001	< 0.001	< 0.001	0.899

* Not statistically significant

The original motive for creating RA-BKT and investigating the DL models is to reduce the learners' cognitive load when using a tutor-

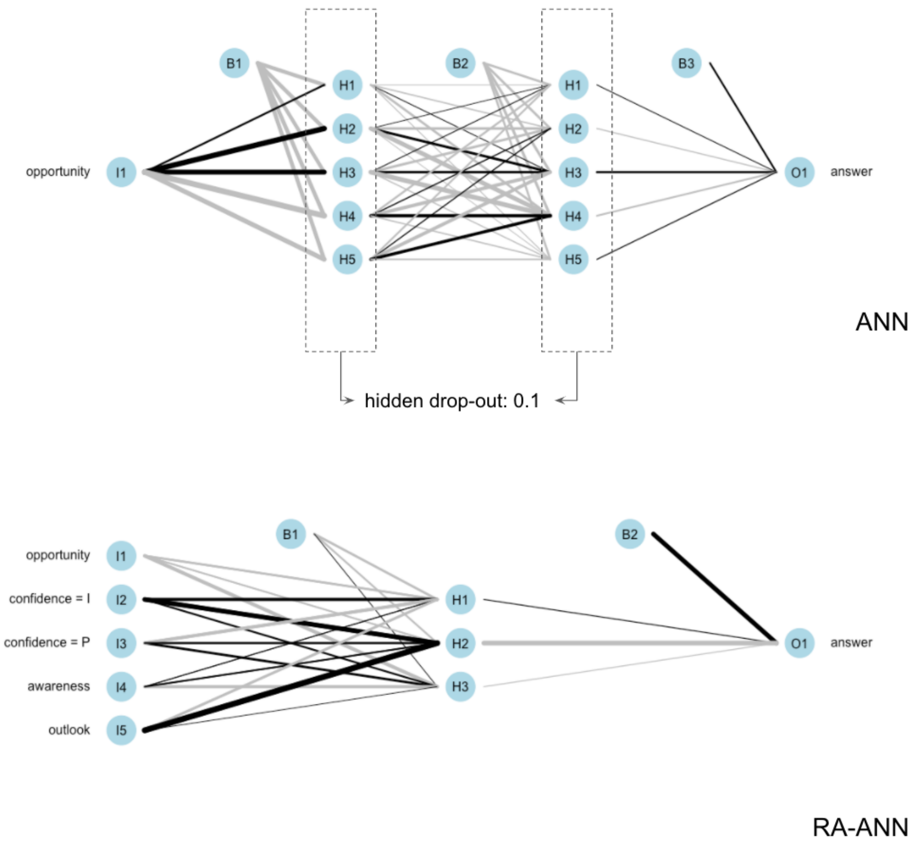


Figure 26: ANN and RA-ANN. Resulting networks with (RA-ANN) and without (ANN) metacognition values.

ing system that has both cognitive and metacognitive tutoring elements. With knowledge tracing, cognitive load is managed by giving the learners just enough opportunities until they master the knowledge component (e.g., provide opportunities until predicted mastery has reached a predefined value, say, 0.90). Once a predefined mastery level is reached, the learner can move on to the next knowledge component, thus spending less time and cognitive resources on the current knowledge component. For the BKT and RA-BKT, the mastery corresponds to the probability that the learner has mastered the knowledge component, which can be seen from the HMM 's transition probability matrix. For the ANN and RA-ANN, this was taken to be the prediction that the answer will be correct.

When looking at mastery predictions, which matters during knowledge tracing, all models differ significantly from each other. With this result, the choice between standard and RA models is no longer a toss-up. While correctness predictions between standard and RA models may differ only due to random chance, choosing one over the other will considerably affect adaptive learning.

Table 19 shows the knowledge parameters from the trained HMM models. Despite the efforts to create a data set intended to prevent degeneracy, BKT still resulted in a potentially degenerate model where the **Guess** and **Slip** are both more than 0.5. The resulting **Initial Learning** is also high, which could be contrary to typical assumptions on the use of tutoring software (learners might use tutoring software to learn concepts not familiar to them, to begin with). Note that no assumptions were made during dataset creation related to prior knowledge, guess, and slip. Fortunately, these same problems are not evident in the resulting RA-BKT model. This can be evidence of RA-BKT's better compliance with the learning intuitions we had set out to investigate.

Table 19: HMM Knowledge Parameter Values. Resulting knowledge parameter values for HMM-based models that are used to check for the models' faithfulness to the modeling intuition.

Description	BKT	RA-BKT
Initial Learning	0.493	0.044
Acquisition	0.010	0.118
Guess	0.794	0.255
Slip	0.770	0.046

For this analysis, the focus is on the RA-BKT parameters directly mappable to the BKT parameters. Those that map to the derivable parameters **Not Guess** $(1 - p(G))$ and **Not Slip** $(1 - p(S))$ were not accounted for. Nevertheless, the RA-BKT parameters that comprise these derivable parameters ($p(U_{c_0}) + p(U_{p_0}) + p(U_{i_0})$ and $p(M_{c_1}) + p(M_{p_1}) + p(M_{i_1}$ respectively) can provide implicit feedback about the way the knowledge component is taught when calculated using actual learner data. For example, a high value for the probability of a learner saying they can answer a question correctly despite not having mastery ($p(U_{c_0})$) may indicate misconception. On the other hand, A high probability of a learner saying they cannot answer the question correctly despite having mastery ($p(M_{i_1})$) could indicate that the question may have been confusing.

Figure 27 compares the resulting linear regression lines for the models' predicted mastery and opportunity. All linear models showed statistical significance with p-value less than 0.05 except for BKT. Ideally, the RA-BKT and the RA-ANN curves sit higher than the BKT and ANN curves to reduce the opportunities required (i.e., higher predicted mastery, thus less work). However, this is not the case especially for the HMM models due to BKT's high **Initial Learning**. With higher **Initial Learning**, the BKT assumes that the learner starts with more prior knowledge than what the RA-BKT predicts.

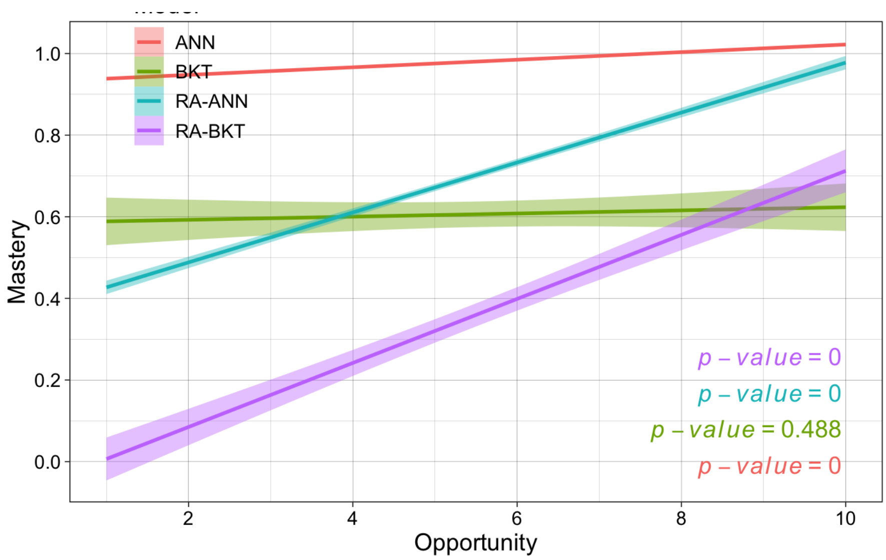


Figure 27: Predicted Mastery Against Opportunity. Predicted mastery against learning opportunity linear regression lines with 95% confidence interval and p-values.

Looking at the regression lines, the RA-BKT has the steepest slope. It can be deduced that had the starting point been the same, the RA-BKT would reach the desired predicted mastery before the other models. The BKT line slope, in particular, is gentler, which can be explained by the potential degeneracy discussed earlier based on the low **Acquisition** and high **Slip** and **Guess**.

Unlike in BKT and RA-BKT where the models can be inspected for possible degeneracy, ANN and RA-ANN would require post hoc modeling to explain the original models. This is a disadvantage of the models based on neural networks. What is known is that the ANN model had lower accuracy, to begin with, making the RA-ANN model the better choice between ANN and RA-ANN. Additionally, Figure 27 shows that the RA-ANN model better reflects the learning intuition the dataset was built on than the ANN case, similar to what is observed with the BKT and RA-BKT models. Thus, the RA models better follow the intuition that the more chances that are given to the learners, the more they gain mastery.

To illustrate how adaptive learning will work with the created models more concretely, suppose the teacher decides that 50% mastery for a given module is sufficient for the learners to proceed to the next module. Without adaptive learning, all learners would have to attempt all ten opportunities before moving on to the next knowledge component. If the curves in Figure 27 are crudely followed for adaptive learning, a learner who answers the first opportunity correctly can already proceed based on the standard BKT and ANN. This makes the standard models suspiciously lenient. The learner would need to answer the first four or nine opportunities to proceed to the next knowledge component if the RA-ANN or RA-BKT respectively are followed, which is reasonably more than the opportunities that the standard models will require but still less than the complete set.

One of the motivations for developing the RA-BKT model is that the awareness and outlook scores could indicate learners' mastery (i.e., guesses and slips could be less frequent if they have desirable awareness and outlook scores). Intuitively, one might say that higher mastery could mean better awareness since the learners have better domain knowledge to understand their knowledge levels. Figure 28 shows the linear regression lines of the predicted mastery against awareness scores. Again, all linear models showed statistical significance with p-value less than 0.05 except for BKT. While all models exhibit an upward trend as expected, the ANN 's ascent is abrupt, and the BKT 's barely noticeable. Once again, the RA models outperform the standard models from this perspective.

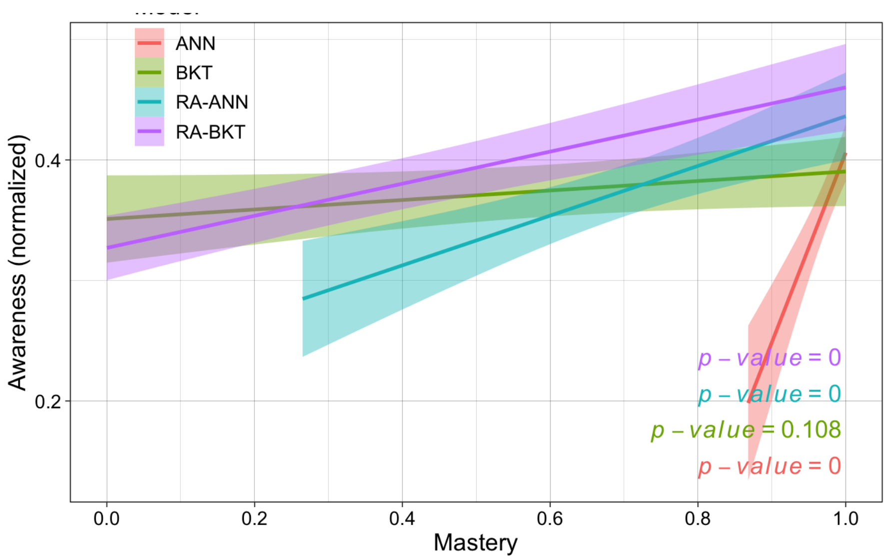


Figure 28: Awareness Against Predicted Mastery. Awareness scores against predicted mastery linear regression lines with 95% confidence interval and p-values.

A similar argument can be applied for the outlook measure. While being optimistic is generally seen as a positive attitude, optimism can have its costs, and pessimism has its benefits in learning[136]. Hence,

136 *K. Sweeny and J. A. Shepperd, "The costs of optimism and the benefits of pessimism.,"* Emotion, *vol. 10, no. 5, pp. 750–753, 2010.* DOI: 10.1037/a0019016.

a neutral outlook is desirable. Figure 29 shows the linear regression lines of the predicted mastery of the models against the corrected outlook scores, like what was used for analyzing the dataset trends. Just like in the previous figures, all linear models showed statistical significance with p-value less than 0.05 except for BKT. It can be seen from the figure that BKT does not follow the intuition of having an upward trend. ANN 's ascent is abrupt, just like the awareness case. On the other hand, the RA models continue to follow intuition.

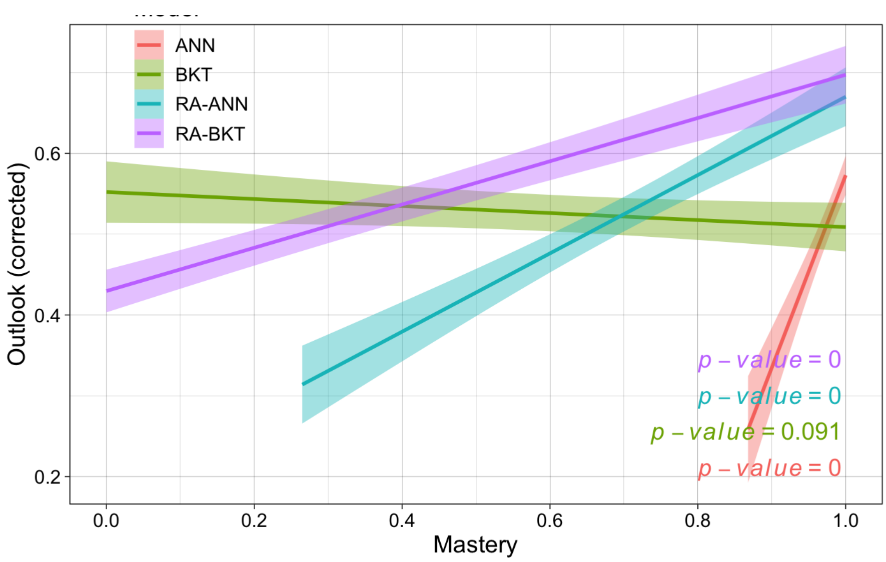


Figure 29: Outlook Against Predicted Mastery. Corrected outlook scores against predicted mastery linear regression lines with 95% confidence interval and p-values.

3.7 CONCLUSIONS ON THE ADAPTIVE ENGINE

The feasibility of using knowledge tracing to manage cognitive resources on a metacognitive tutor using the RA model was explored. The RA-BKT was constructed by expanding the observation vocabulary to include the correctness of the learners' answers at each opportunity and their confidence in self-reports. The goal was the understand the usefulness of adding metacognitive measures in knowledge tracing when compared to standard inputs. In addition, DL models using metacognitive inputs and not using metacognitive inputs were also constructed to situate this work within the recent research trends. All resulting models are then compared with each other. A dataset based on assumed learner behaviors was created for this purpose.

Creating a synthetic learner dataset is an original approach to conducting modeling based on a learning theory expected to be applied at scale but is yet to be tested. The approach was validated by comparing the trends from the created dataset with existing large-scale datasets. Showing similarities can be valuable when testing educational theories that have no precedent before subjecting learners to experimentation.

Even though training times for the models varied, all of them had decent prediction times. RA-ANN had the best test accuracy, followed by RA-BKT. As such, the models using metacognitive inputs performed better than those without in terms of accuracy. A possible reason for ANN 's lack-luster accuracy is insufficient data to produce the deep network that it has resulted in. While the RA-ANN had the best performance, seeing its resulting network only had a single hidden layer can leave doubt. It could have performed better if there were sufficient data to construct a deeper network. All models are statistically different from each other in terms of mastery prediction, while correctness predictions were not statistically significant when comparing between standard and RA models.

The resulting parameters for RA-BKT are non-degenerate. The observation follows from the predicted hidden state (i.e., whether the learner answers correctly or not follow from the prediction whether the learner has learned the knowledge component or not). The same was not observed for the BKT: the estimated guess and slip probabilities, as well as the estimated prior knowledge, are all too high (greater than 0.5). For the RA-BKT to be a better alternative to BKT in cognitive load management, it should return higher mastery predictions. This was not the case since the BKT 's prior knowledge prediction is too high. However, when regression lines of opportunity against mastery are checked, RA-BKT had a steeper slope. Had the BKT and RA-BKT ended up with similar prior knowledge predictions, the desired mastery level would be more quickly achieved with RA-BKT. Similar observations were also seen when comparing ANN and RA-ANN.

Another point of interest is the relationship between RA measurements and predicted mastery. The RA-BKT and RA-ANN show upward trends when comparing mastery with awareness and outlook more closely than the standard models. However, the BKT does not consistently show upward trends, and the ANN has too abrupt slopes. These observations show that RA-BKT and RA-ANN follow cognitive and metacognitive learning intuitions better than the standard models.

In summary, RA-BKT and RA-ANN could be viable options for managing cognitive load while metacognitively tutoring given their high accuracy, efficient prediction times, and more intuitive predictions. RA-BKT can be looked at more favorably when constraints such as training time, dataset size, or hardware are present. Additionally, since the meaning of the underlying structure behind the HMM-based models are known from the offset not like in the case of the neural networks, RA-BKT can potentially be easier to diagnose for problems based on the resulting parameter values (e.g., Guess and Slip probabilities that can cause model degeneracy). This is critical as metacognition is crucial in succeeding in learning environments that require significant autonomy. With the emergence of online learning, concepts that can be challenging even with teacher-based support find their way in the online medium. This exacerbates the need to teach metacognition while ensuring deep cognitive learning.

3.8 WHAT'S NEXT?

A central weak point of this experiment is the dataset used. No actual data that includes both learner performance and metacognitive measures exist as far as researchers know. Furthermore, while cognitive performance might be estimated using other learning theories such as the IRT[137], a similar theory is yet to be formulated for confidence ratings.

On the other hand, the synthetic dataset may be a strength of this study. Using fictional characters or personas for evaluating preliminary work is done in fields like user-centered design and marketing, especially in cases where the work is not yet ready for use by their intended audience. This strategy is not yet validated for educational research purposes. A way to validate synthetic datasets is presented in this study. Comparing the results of this theoretical approach with data from actual usage can open new possibilities.

The availability of data from actual usage would raise more interesting investigation points. For one, metacognition is generally seen as not domain-specific. There is still a belief that the domain where the metacognitive opportunity is presented does matter[138]. The RA models may provide link between metacognitive and cognitive knowledge like what was attempted by plotting mastery predictions against awareness and outlook scores. Having data from the metacognitive tutoring tool usage on different cognitive domains will be beneficial for this investigation.

A possible drawback of the RA models is modeling fairness. The RA models use confidence ratings that can be influenced by personal characteristics such as gender[139] and culture[140]. As such, the resulting mastery predictions may unjustly penalize some learners, not because of their lack of cognitive or metacognitive mastery, but because of their innate attitudes. Therefore, fairness enhancing interventions might need to be considered[141].

When thinking of what is most "fair" to the learners, there is a need to seek the expertise of the teachers. Traditionally, the teachers have regular (possibly daily) interactions with the learners, developing the teachers' insight into facilitating learning. However, for the teachers to be equally effective in online learning environments where the interactions can be very different from offline classrooms, tools must be provided to "look into" their learners' situations.

[137] *S. P. Reise, A. T. Ainsworth, and M. G. Haviland, "Item response theory: Fundamentals, applications, and promise in psychological research,"* Current Directions in Psychological Science, *vol. 14, no. 2, pp. 95–101, 2005.* DOI: 10.1111/j.0963-7214.2005.00342.x.

[138] *I. Roll, V. Aleven, B. M. McLaren,* et al., *"Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking,"* Metacognition and Learning, *vol. 2, no. 2, pp. 125–140, 2007.*

[139] *C. L. Colbeck, A. F. Cabrera, and P. T. Terenzini, "Learning professional confidence: Linking teaching practices, students' self-perceptions, and gender,"* The Review of Higher Education, *vol. 24, no. 2, pp. 173–191, 2001.*

[140] *E. S.-c. Ho* et al., *"Characteristics of East Asian learners: What we learned from PISA,"* Educational Research Journal, *vol. 24, no. 2, p. 327, 2009.*

[141] *S. A. Friedler, C. Scheidegger, S. Venkatasubramanian,* et al., *"A comparative study of fairness-enhancing interventions in machine learning," in* Proceedings of the Conference on Fairness, Accountability, and Transparency, *2019, pp. 329–338.*

4

THE ANALYTICS DASHBOARD

"I quite agree with you," said the Duchess;
"and the moral of that is 'Be what you would seem to be'
—or if you'd like it put more simply—
'Never imagine yourself not to be otherwise than what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise.'"
"I think I should understand that better," Alice said very politely, "if I had it written down: but I can't quite follow it as you say it."
"That's nothing to what I could say if I chose," the Duchess replied, in a pleased tone.
"Pray don't trouble yourself to say it any longer than that," said Alice.

–Lewis Caroll, *Alice in Wonderland*

Learners must understand not just how they performed in class, but also how they could improve[142]. Such information can be provided to the learners through **feedback**. For the feedback to be effective, it should elaborate on the performance criteria and not just give out grades. It should also provide actionable improvement points and have the right balance of critique and support to build confidence. We attempt to provide feedback to the learners with the Personalized Online Adaptive Learning System (POALS) Metacognitive Tutor through the Evaluation Phase's Learner Profile. Additionally, there is an option for the teachers to provide hints to the learners when the incorrect answer is selected in the Problem-Solving Phase.

It is not just the learners that need feedback, but the teachers, too. Learner-provided feedback is critical for improving teaching quality for future learners[143]. Equally importantly, feedback to the teacher is crucial for them to provide interventions to prevent learners from losing motivation[144]. However, learner evaluation surveys, which are the most common source of teacher feedback, can suffer from lack of validity (different learners in the same class giving drastically different evaluation on the same teacher), bias (learners more likely giving negative feedback when they did not receive good grades), and lack of utility (scores are usually normalized across different teachers whose teaching workloads can vary widely, making resulting recommendations not applicable)[145].

Periodic feedback is also needed to emphasize the role of the learners as stakeholders in their and their peers' learning and to prevent learners from losing motivation. An alternative source of feedback that recently has been gaining attention is learning analytics dashboards (LADs). While LADs can benefit learners in developing self-

[142] *P. Ferguson, "Student perceptions of quality feedback in teacher education,"* Assessment & Evaluation in Higher Education, *vol. 36, no. 1, pp. 51–62, 2011.*

[143] *D. Feistauer and T. Richter, "How reliable are students' evaluations of teaching quality? A variance components approach,"* Assessment & Evaluation in Higher Education, *vol. 42, no. 8, pp. 1263–1279, 2017.*

[144] *D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice,"* Studies in Higher Education, *vol. 31, no. 2, pp. 199–218, 2006.*

[145] *H. W. Marsh and L. A. Roche, "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility,"* American Psychologist, *vol. 52, no. 11, p. 1187, 1997.*

directed learning skills[146], they might prevent teachers from exercising creativity in instructional planning and decision making[147].

4.1 GETTING FEEDBACK

In a traditional classroom, learners have multiple avenues of interaction with the teacher available to them. For example, when the learner needs something, they can raise their hand during class or approach the teacher after the class finishes or during office hours. Likewise, the teacher can gauge the learners' engagement by observing what is happening inside the classroom and in the follow-up interactions outside of class. Even if the classroom interactions are minimal, there are still opportunities for the teachers and learners to exchange feedback during several activities, including homework, projects, and exam evaluation.

In an online classroom, such communication can be more difficult. For those in synchronous classes, such as through teleconferencing, the learners are just muted in most cases. The teacher may not even be sure if the learners are present. In other forms of the online classroom, such as in asynchronous classes like massive open online courses (MOOCs), learners and teachers can interact in the discussion forums. However, research has shown that, in some cases, only about three percent of learners post on these forums[148]. Another possibility for interaction is when learners provide feedback during course surveys. However, these surveys are frequently done at the start and the end of the course. Hence, there is not much opportunity for the teacher to adjust their teaching based on learner feedback. Additionally, most online classrooms are self-paced, so it is hard for them to see how learners are faring overall. An added complication is that some learners lack self-directed learning skills, which are essential to succeeding in online learning environments[149].

An important thing to consider moving forward is that online learning has its strengths and weaknesses, and it is likely here to stay. For instance, the COVID-19 pandemic situation remains uncertain despite vaccination roll-out worldwide. We also have learned that some learners benefit tremendously from online platforms[150]. For example, some learners have jobs with schedules conflicting with their classes, or some learners have disabilities that are better accommodated in an online learning platform. So, it is only fair to keep online classrooms or hybrid forms as an option for everyone who might need them now that we know that it is plausible for all levels of learning.

Self-directed learning can be handled separately by introducing metacognitive tutors in the learning environment[151]. An example of this is the POALS Metacognitive Tutor that we developed. This tool was designed to work as a stand-alone web-based application and as an add-on to learning management systems (LMSs) like Canvas or

[146] *G. Sedrakyan, J. Malmberg, K. Verbert,* et al., *"Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation,"* Computers in Human Behavior, *vol. 107, p. 105 512, 2020.*

[147] *M. Brown, "Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data,"* Teaching in Higher Education, *vol. 25, no. 4, pp. 384–400, 2020.*

[148] *D. F. Onah, J. E. Sinclair, and R. Boyatt, "Exploring the use of MOOC discussion forums," in* Proceedings of London International Conference on Education, *2014, pp. 1–4.*

[149] *M. Zhu, C. J. Bonk, and M. Y. Doo, "Self-directed learning in MOOCs: Exploring the relationships among motivation, self-monitoring, and self-management,"* Educational Technology Research and Development, *pp. 1–21, 2020.*

[150] *J. Reich, "Ed tech's failure during the pandemic, and what comes after,"* Phi Delta Kappan, *vol. 102, no. 6, pp. 20–24, 2021.*

[151] *C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in* International Conference on Intelligent Tutoring Systems, *Springer, 2004, pp. 668–677.*

Moodle. POALS Metacognitive Tutor has been instrumental in developing the learners' ability to regulate their knowledge, which is vital for self-directed learning[152].

4.2 LEARNING ANALYTICS DASHBOARDS

Learning analytics dashboards LAD are interfaces showing multiple data visualizations aimed to provide information about learning experiences. LADs had been created for various purposes, such as providing generic information on learner activities in a MOOC[153], enabling communicating at-scale[154], providing early warnings about at-risk learners[155], supporting academic advising[156], and possibly giving insights on peripheral learning communities[157].

LADs, in one way or another, provide feedback to different target audiences. For example, a typical LAD accessible to the teacher that shows learner activities (e.g., learners who submitted exercises, number of logins, and others) can provide feedback on the engagement levels of the learners. Learners who consult LADs that show their progress in their online courses can make adjustments to ensure they meet a deadline. Additionally, administrators such as those involved in academic advising can consult LADs in understanding whether their curricula can support learner growth and prepare them for jobs.

LADs have several appealing points. First, they visualize information, making it easier to digest as opposed to large quantities of text. Secondly, they aggregate multiple information in a single view, making related information more accessible. Finally, the information is updated regularly, if not in real-time. These regular updates can facilitate a continuous course design cycle, which in turn can reduce confusion and frustration, clarify expectations, and maximize learning[158].

While natural language processing had been used in MOOC quality assurance[159], as far as we know, there is no research involving the use of natural language processing (NLP) with metacognitive reflections for LADs yet. This is a critical exploration point since metacognitive tutoring based on open-response prompts such as the one provided by POALS Metacognitive Tutor can provide unique opportunities to enhance LADs. For one, POALS Metacognitive Tutor is more private than discussion forums, making them more conducive to learners who feel anxious with a larger audience. Secondly, learner interaction with POALS Metacognitive Tutor is as frequent as the learning exercises provided; hence, feedback can be obtained on a more regular basis compared to course surveys. Finally, since the prompts are reflective in nature, the implicit feedback that can be harvested may be very different from the feedback explicitly provided by learners, yet still be very relevant to their learning experiences. This chapter demonstrates how the POALS Metacognitive Tutor can be used as in-

[152] *M. K. J. Carlon and J. S. Cross, "Open response prompts in an online metacognitive tutor," in* The 2021 Annual Spring Conference of Japan Society for Educational Technology, *2021.*

[153] *R. Cobos, S. Gil, A. Lareo,* et al., *"Open-DLAs: An open dashboard for learning analytics," in* Proceedings of the third (2016) ACM conference on learning@ scale, *2016, pp. 265–268.*

[154] *C. V. Le, Z. A. Pardos, S. D. Meyer,* et al., *"Communication at scale in a MOOC using predictive engagement analytics," in* International Conference on Artificial Intelligence in Education, *Springer, 2018, pp. 239–252.*

[155] *K. Sun, A. H. Mhaidli, S. Watel,* et al., *"It's my data! Tensions among stakeholders of a learning analytics dashboard," in* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, *2019, pp. 1–14.*

[156] *F. Gutiérrez, K. Seipp, X. Ochoa,* et al., *"LADA: A learning analytics dashboard for academic advising,"* Computers in Human Behavior, *vol. 107, p. 105 826, 2020.*

[157] *D. A. Joyner, "Peripheral and semi-peripheral community: A new design challenge for learning at scale," in* Proceedings of the Seventh ACM Conference on Learning@ Scale, *2020, pp. 313–316.*

put for a LAD by constructing a wireframe of the POALS Analytics Dashboard. A user study was also conducted to evaluate the usefulness and usability of the POALS Analytics Dashboard.

[158] *Texas A & M University. (nd). Course design, [Online]. Available:* *https://cte.tamu.edu/Instructional-Resources/Course-Design*.

[159] *J. S. Cross, N. Keerativoranan, M. K. J. Carlon,* et al., *"Improving MOOC quality using learning analytics and tools," in* 2019 IEEE Learning With MOOCS (LWMOOCS), *IEEE, 2019, pp. 174–179.*

4.3 POALS ANALYTICS DASHBOARD WIREFRAME

Presented as: M. K. J. Carlon and J. S. Cross, "POALS Analytics Engine: A student affect dashboard," in *The Eighth UK Japan Engineering Education League Workshop*, 2021

To help the teacher somehow receive feedback from their learners, we created POALS Analytics Dashboard. POALS Analytics Dashboard is a LAD that shows feedback that the learners might not say directly extracted using NLP from the prompt responses. A wireframe of the Analytics Dashboard is shown in Figure 30.

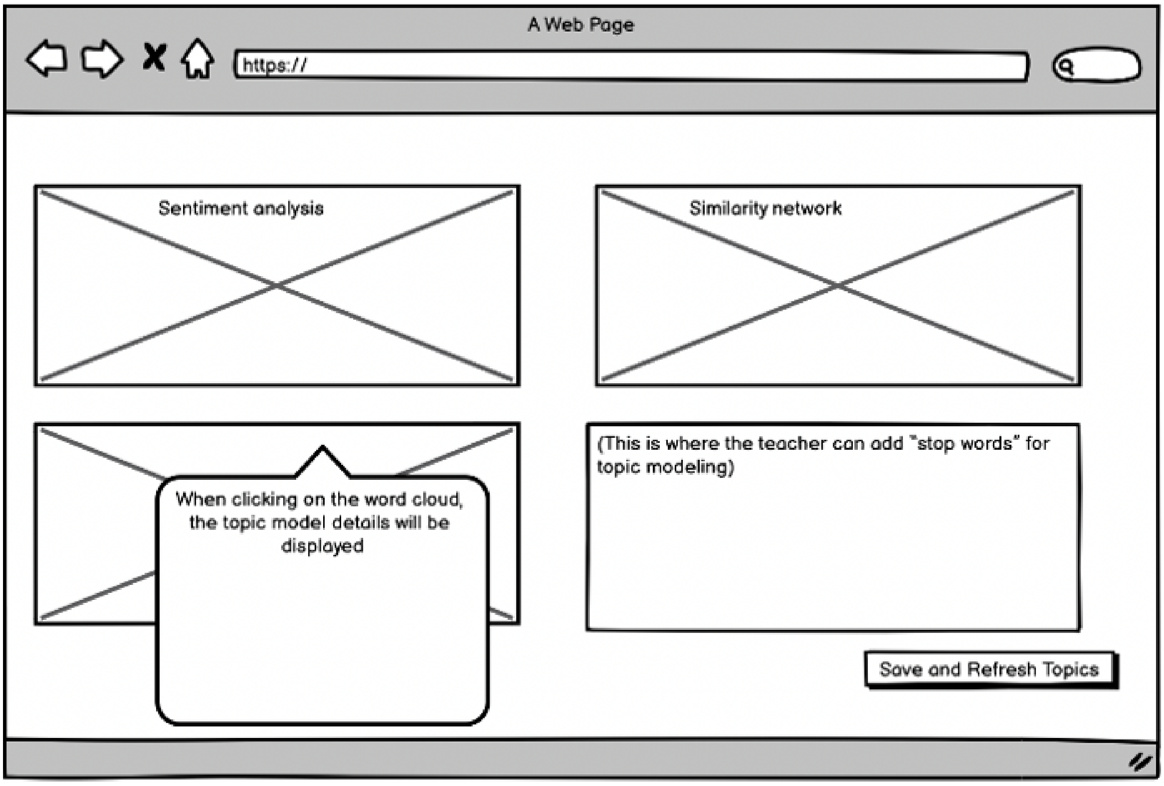


Figure 30: Analytics Dashboard Wireframe. A sketch of the POALS Analytics Dashboard that contains an aggregate visualization for sentiment analysis, a word cloud of prompt response topics, and a similarity network of the course contents.

Wireframes are cost-effective visualizations of a system to be made that puts focus on the content (what information to be shown) as opposed to specific interface components (e.g., buttons, links, and others) as in the case of prototypes, and expected look and feel as in mock-ups[160]. A wireframe is created instead of a higher fidelity

[160] *S. Sutipitakwong and P. Jamsri, "Pros and cons of tangible and digital wireframes," in* 2020 IEEE Frontiers in Education Conference (FIE), *IEEE, 2020, pp. 1–5.*

prototype or mock-up to be able to focus on the NLP algorithms and not be caught up with aesthetics.

Analytics dashboards have received popularity due to the need to make sense of the increasingly available big data. In education, LADs have been utilized to provide insight, prompt user reflection, and inform interventions[161]. When theoretically grounded and designed with user-centered approaches, LADs such as POALS' Analytics Dashboard may provide practical learning support.

The visualizations for the Analytics Dashboard are created using Highcharts[162] JavaScript software library, while the NLP were done using relevant Python libraries. The data gathered from the Metacognitive Tutor full experiment in section 2.3.3 was used for this demonstration.

4.4 SENTIMENT ANALYSIS

The online teacher might be interested to know how happy their learners are in their class. While ensuring that the learners are in a cheerful disposition seems to be "common sense" important, this has theoretical underpinnings. Subtle, elusive, unverbalized emotions are the basis of thought, meaning, and language, and thus affect perception and eventually, cognition[163].

To address this, we conducted sentiment analysis. Figure 31 is a composite visualization of learner sentiments. The column graph shows the average absolute sentiment polarity score from the prompt responses. Its color changes to red when the sentiment is negative, orange when neutral, and green when positive. The ratio of learners that have negative, neutral, or positive sentiments is shown through the half-donut chart. This visualization allows the teacher to see if there are strong sentiments and how prevalent is such sentiment within the class.

Sentiment analysis is an NLP task where the affective state of a given text is quantified. A common approach is to treat it as a classification problem where the target text is classified to a particular sentiment, typically either positive or negative. Sentiment analysis can be done using knowledge-based techniques (i.e., defining rules, typically based on some lexicon), statistical approaches (e.g., machine learning techniques such as the use of support vector machines[164]), or a combination of said approaches. Sentiment analysis has been previously used in other instructional quality assurance activities, such as assessing improvements introduced after revising a MOOC[165]. In this research, we are using machine learning (transformers in particular).

The learner responses in the preparation and evaluation prompts were used as inputs for the sentiment analysis. For this particular case, all inputted text was presumed to be Japanese. This assumption is essential since the choice of language dictates the possible libraries

[161] W. Matcha, D. Gašević, A. Pardo, et al., *"A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective,"* IEEE Transactions on Learning Technologies, *vol. 13, no. 2, pp. 226–245, 2019.*

[162] *Highcharts. (2021). Interactive JavaScript charts for your webpage, [Online]. Available:* `https://www.highcharts.com/`.

[163] *Y. Kanazawa, "Emotion as "deeper" than cognition: Theoretical underpinnings and multidisciplinary lignes de faits to the Emotion-Involved Processing Hypothesis (EIPH),"* 国際学研究= Journal of International Studies, *vol. 9, no. 1, pp. 185–206, 2020.*

[164] *G. Song, "Sentiment analysis of Japanese text and vocabulary learning based on natural language processing and SVM,"* Journal of Ambient Intelligence and Humanized Computing, *pp. 1–12, 2021.*

[165] *M. K. J. Carlon, M. R. Gaddem, C. A. Hernández Reyes,* et al., *"Investigating mechanical engineering learners' satisfaction with a revised monozukuri MOOC," in* European MOOCs Stakeholder Summit 2021, *2021.*

and pre-trained models that can be used for sentiment analysis. This also affects the other NLP-based analysis done by the Analytics Dashboard. Listing 1 shows the code snippet used for sentiment analysis.

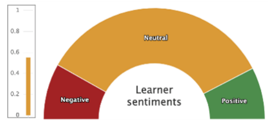


Figure 31: Planned Sentiment Analysis Visualization. The half-donut shows the ratio of learners in each sentiment and the column graph shows the average sentiment.

Listing 1: Sentiment Analysis Code Snippet

```
from transformers import BertJapaneseTokenizer
from transformers import pipeline,
    AutoModelForSequenceClassification

tokenizer = BertJapaneseTokenizer.from_pretrained('cl-tohoku/bert
    -base-japanese-whole-word-masking')
model = AutoModelForSequenceClassification.from_pretrained('daigo
    /bert-base-japanese-sentiment')
sentiment_analyzer = pipeline("sentiment-analysis",model=model,
    tokenizer=tokenizer)
```

The Python library **BertJapaneseTokenizer** was used for tokenization. Tokenization is the process of splitting an input text (possibly a sentence or even a paragraph) into smaller units called tokens. Each token could correspond to a word, or depending on the choice of algorithm, parts of a word. For instance, *"Alice is running"* may be tokenized as `["Alice", "is", "running"]` or `["Alice", "is", "run", "ing"]`.

The transformer-based deep learning technique Bidirectional Encoder Representations from Transformers (BERT) was used as the pre-trained system for tokenizing[166]. More specifically, the model trained by Tohoku University specific for the Japanese language, *bert-base-japanese-whole-word-masking*, was used[167]. Transformers are designed to handle sequential data (e.g., a sequence of words that form a paragraph) without needing the sequence to be processed in order[168]. BERT, in particular, uses a bidirectional approach to improve context learning by using masked language pre-training objective which will be described later.

The sentiment analysis task used is a classification task where an input text is either positive or negative. The *bert-base-japanese-sentiment* was used as the pre-trained model, which labels the input as either ポジティブ(positive) or ネガティブ(negative)[169]. The calculated probability for the resulting label was also extracted.

Other English sentiment dictionaries provide mechanisms to identify whether a sentiment score is positive, neutral, or negative. For example, the Natural Language Toolkit (NLTK), a Python library, has the Valence Aware Dictionary and sEntiment Reasoner (VADER) functionality that calculates the probabilities of a given text being positive, neutral, or negative. In our case, it is important to see the percentages of learners who have polar sentiments. Hence we need to establish a neutral range that is not predefined in our sentiment dictionary choice. The label was changed to neutral for both negative and positive cases when the probability was less than 0.25. The cut-off is

166 *J. Devlin, M.-W. Chang, K. Lee,* et al., *"BERT: Pre-training of deep bidirectional transformers for language understanding,"* arXiv preprint arXiv:1810.04805, *2018.*

167 *Tohoku University. (Jan. 2021). BERT base Japanese (IPA dictionary, whole word masking enabled), [Online]. Available:* https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking.

168 *A. Vaswani, N. Shazeer, N. Parmar,* et al., *"Attention is all you need,"* arXiv preprint arXiv:1706.03762, *2017.*

169 *Daigo. (Dec. 2020). bert-base-japanese-sentiment, [Online]. Available:* https://huggingface.co/daigo/bert-base-japanese-sentiment.

just heuristically set and must be further investigated. These labels (positive, negative, neutral) then corresponded to the green, red, and orange colors in the half-donut chart, respectively. Finally, the probabilities were multiplied by -1 when the original label is negative, and the values are summarized to get the value for the column graph.

Here is the result from the electrical engineering course. It is interesting to see that there are not many neutral responses. In a typical classroom, learners are primarily neutral. Learners who are very happy or very sad are most likely very rare. When the sentiment analysis model was investigated, it was observed that even very a simple sentence like *"I live in Hachioji"* gives a very positive score.

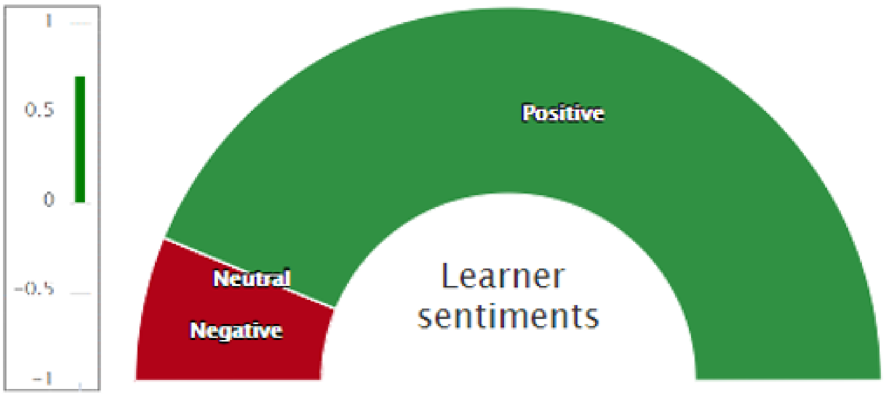


Figure 32: Resulting Sentiment Analysis Visualization. The sentiment of the electrical engineering students appeared to be suspiciously too positive.

This is then a significant challenge for sentiment analysis, particularly in non-English languages. The usual strategy in the past is to translate to English first then use the English models since English models are typically more mature. However, because we are transforming information, there is a chance that there might be some information loss. An alternative is to build our sentiment model, but this will require extensive data collection.

Additional pretraining can also be done where a corpus specific to the target domain is used to train a model under a masked language modeling task. A masked language modeling task, or cloze task[170], involves filling in a masked portion of a given passage *"The quick brown **** jumped over the lazy dog"* in the given context. This is the same technique used for training BERT.

Manual labeling was done to all responses (positive, neutral, or negative) to have a sense of how well pretraining can improve our results. Ideally, labeling should be done by multiple persons using a rubric and tested for reliability[171]. The process will then be iterated until the labelers arrive at reliably similar results. Since this work is exploratory and our intention is not to make a sentiment analysis model for public consumption, labeling was only done once. The labeled responses are divided into training and test datasets as shown in Table 20.

[170] *W. L. Taylor, ""cloze procedure": A new tool for measuring readability,"* Journalism Quarterly, *vol. 30, no. 4, pp. 415–433, 1953.*

[171] *J. Lappeman, R. Clark, J. Evans,* et al., *"Studying social media sentiment using human validated analysis,"* MethodsX, *vol. 7, p. 100 867, 2020.*

Sentiment	Total	Training	Test
All	903	722 (80%)	181 (20%)
Positive	249	199 (80%)	50 (20%)
Neutral	512	410 (80%)	102 (20%)
Negative	142	113 (80%)	29 (20%)

Table 20: Sentiment Analysis Pretraining Data. The split between training and test data set for sentiment analysis pretraining after manual coding.

The dataset was set to have one label feature only, set to 0 for negative sentiment, 1 for neutral sentiment, and 2 for positive sentiment. Another option is to have different label features for each sentiment and do one-hot encoding (e.g., set the negative label to 1 and neutral and positive labels to 0 when the response has negative sentiment). This is similar to what was done to the confidence ratings during model training the POALS Adaptive Engine. However, since labeling is done manually, only one label feature was created to prevent human error.

Again, BertJapaneseTokenizer was used with the cl-tohoku/bert-base-Japanese-whole-word-masking model. Because the label feature is not binary, we are conducting multiclass classification. PyTorch library's Dataset was used for supplying subsets of our dataset to the neural network in mini-batches (4 for training and 2 for testing) to prevent memory issues. It also helps make the data entries have a uniform shape, varying vastly with text-based data. In our case, the maximum data length is set to 512, which is the maximum sequence that BERT can manage. To prevent overfitting , Dataset reshuffles the data every epoch (this was set to 3), or a complete pass through the entire data.

Overfitting occurs when the resulting model being too specific to the training data set that it cannot generalize to cases not seen during training

A neural network was also created with BERT for finetuning based on the pre-trained model used for the tokenizer. Neural network model training involves iteratively predicting the outputs, comparing the prediction with the actual output, and updating the weights based on a loss function. The Adam optimizer, which was shown to work well with data with sparse gradient such as in natural language processing by calculating an exponential moving average of the gradient and the squared gradient[172], was used for these iterations (learning rate was set to $1\text{x}10^{-5}$). The cross-entropy loss (shown in Equation 10) was used for the loss function, which works well for multiclass classification[173].

[172] *D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization,"* arXiv preprint arXiv:1412.6980*, 2014.*

[173] *E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss,* et al.*, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning,"* arXiv preprint arXiv:2011.05231*, 2020.*

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{10}$$

In the above formula,

- M is the number of classes (three in our case: positive, neutral, and negative),

- log is the natural log,
- c is the classification,
- o is the observation,
- y is a binary indicator $(0, 1)$ as to whether c is the correct classification for o, and
- p is the probability that o is of class c.

The training accuracy is 91.97%, and the test accuracy is 81.77%. When the manual labels were compared with the output of the previously selected generic sentiment analysis model, the accuracies are 37.40% and 34.25% for training and test datasets, respectively, both of which are lower than random chance. Having a customized sentiment analyzer can thus be beneficial and worthy of further research. Future research directions can include developing a sentiment analyzer that can generalize well in the academic setting. This could also be generic research on continuous updates on the sentiment analysis model as new data comes in, usually referred to as lifelong machine learning (not to be confused with lifelong learning in an educational context)[174].

[174] X. Hong, P. Wong, D. Liu, et al., *"Lifelong machine learning: Outlook and direction," in* Proceedings of the 2nd International Conference on Big Data Research, *2018, pp. 76–79.*

4.5 LEARNER MISCONCEPTIONS

Another thing that the online teacher might think is, had the class been in-person, what will the learners be talking about? Another implicit feedback we will retrieve from the prompts is the topics discussed by the learners. A word cloud similar to Figure 33 can help the teachers gauge if the learners can pick up the essential key points in the lessons.

Figure 33: Planned Topic Model Word Cloud. First, the topics are gathered and the word counts are calculated. Words found in the modeled topics are then ranked according to the word count, and a word is selected for each topic.

While knowing the topics the learners are thinking of in itself is interesting, learning about these topics can also help the teacher uncover misconceptions. To elaborate, if a topic from a particular module comes up, it may be intuitive to think that the class average for the said module may be high since enough learners have thought more carefully about it. If, however, an emerging topic ends up coming from a low-scoring module, then it may be worth investigating if there is a misconception that the learners repeat over.

Misconception is a prevalent problem in education, may it be in chemistry[175], mathematics[176], or any other field. Misconceptions can stem from various causes, such as tricky wording or unfamiliar words not properly introduced in class that the learner then searches for in other possible sources. This becomes especially more important in the online learning environment as the internet is a frequent source of misconceptions[177].

[175] M. Üce and I. Ceyhan, *"Misconception in chemistry education and practices to eliminate them: Literature analysis.,"* Journal of Education and Training Studies, *vol. 7, no. 3, pp. 202–208, 2019.*

[176] R. Rochmad, M. Kharis, A. Agoestanto, et al., *"Misconception as a critical and creative thinking inhibitor for mathematics education students,"* Unnes Journal of Mathematics Education, *vol. 7, no. 1, pp. 57–62, 2018.*

[177] B. Acar Sesen and E. Ince, *"Internet as a source of misconception,"* Turkish Online Journal of Educational Technology-TOJET, *vol. 9, no. 4, pp. 94–100, 2010.*

To aid with the visualization, we used the Python **WordCloud** library. The actual result from the word cloud was just used for sanity check of the resulting topics. The word cloud displayed on the Analytics Dashboard is determined by the most prominent word in each topic, and word size is the probability assigned to the topic.

To get the topic information, we conducted topic modeling on the metacognitive prompt responses. The algorithm latent Dirichlet allocation (LDA) was used on term frequency-inverse document frequency (tf-idf) token vectors available on the **gensim** library after tokenizing using **MeCab** library. After tokenization, fields that are are not a noun, verb, or adjective are removed for simplicity.

In LDA, observations (e.g., words and documents) are used to explain latent or unobserved groups (i.e., topics) by looking where the observations are present[178]. The "presence matrix" was constructed with tf-idf where the weight of words (term frequency) is matched up with how much information does a word provides, that is, in how many documents does the said word appear (inverse document frequency). The code used for topic modeling is simplified in Listing 2.

[178] *D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation,"* Journal of Machine Learning Research, *vol. 3, pp. 993–1022, 2003.*

Listing 2: Topic Modeling Code Snippet

```
import MeCab
from gensim.corpora.dictionary import Dictionary
from gensim.models import LdaModel
from collections import defaultdict

mt = MeCab.Tagger('')
train_texts = []
for index, row in prompts.iterrows():
  text = []
  node = mt.parseToNode(row["response"])
  while node:
    fields = node.feature.split(",")
    if fields[0] == '<noun>' or fields[0] == '<verb>' or fields
        [0] == '<adjective>':
      text.append(node.surface)
    node = node.next
  train_texts.append(text)

dictionary = Dictionary(train_texts)
corpus = [dictionary.doc2bow(text) for text in train_texts]
lda = LdaModel(corpus=corpus, num_topics=NUM_TOPICS, id2word=
    dictionary)
# NUM_TOPICS was set to 5, which is the number of modules + 1
```

Figure 34 is the result from the electrical engineering class after translation to English. During topic modeling, sometimes some irrelevant words will pop up. For example, in our case, several learners mentioned they wanted to review the videos, which might not be attractive to the teacher. These unwanted words can make interpretation difficult. Therefore, we added a feature where the teacher could specify the words they do not want to see, and then the Analytics Dashboard will update the result. When the responses become too numerous, topic modeling may take a long time to execute. Hence, we initially intended for the modeling to be done on a nightly basis. However, it may be problematic for the teacher to wait for their settings to take effect for an entire day.

Figure 34: Resulting Topic Model Word Cloud. These were the English translations of selected words from emerging topics after excluding irrelevant words.

Unless someone is part of the course' teaching team, the resulting word cloud may not make sense. This is as intended as we hope to tap into the teacher's expertise in taking action based on what the POALS Analytics Dashboard provides. Nevertheless, the word cloud can be improved by providing a drill-down level where the teacher can pick a specific topic to inspect further and see more details.

4.6 SURFACE LEARNING

Finally, maybe the teacher will be interested in knowing which topics the learners can remember well. For example, some information from Module 1 might be necessary to understand Module 3. If the learners got good scores in Module 1 but lower scores in Module 3, the teacher might want to intervene to ensure that the learners still remember Module 1.

Surface learning is the phenomenon where a learner picks up just enough knowledge to pass immediate tests but possibly not enough to sustain them in future learning activities[179]. The choice between surface learning or deep learning (intending to understand topics and relate to other ideas) relies not just on the learners' motivation but also on the accompanying learning activities. Therefore, the teacher must see if surface learning is widespread in their class, which can al-

[179] *D. H. Dolmans, S. M. Loyens, H. Marcq,* et al., *"Deep and surface learning in problem-based learning: A review of the literature,"* Advances in Health Sciences Education, *vol. 21, no. 5, pp. 1087–1112, 2016.*

lude to how the learning activities are constructed rather than several individual motivations.

The teacher may be able to detect surface learning through a network graph just like shown in Figure 35. Nodes represent modules, and modules that are related are connected. Hovering on a node shows the average learner score in the exercises for the said module. By inspecting the average learner scores on related modules, the teacher may be able to detect surface learning or the case where learners can only master a lesson but could not apply the learning to succeeding lessons.

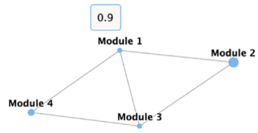


Figure 35: Planned Similarity Network Graph. Each module is represented by a node, with nodes meeting similarity cut-off being connected.

To measure similarities between modules, the corresponding texts must first be converted into numerical representations or **vectors**. In a previous attempt, metacognitive prompt responses were vectorized with tf-idf. However, since tf-idf relies on word counts, word ordering and semantics are not taken into account. This is workable for dictionary-based models, such as in some sentiment analyzers or cases where we are only interested in making salient points more visible such as in topic modeling. This can be problematic specifically for our similarity network graph since when we measure similarities between documents that use the same sets of words repeatedly, the ordering and semantics will matter.

The **doc2vec** algorithm available in the **gensim** library was used instead for vectorization. For **doc2vec**, semantics and order are preserved by using entire documents (or paragraphs) in creating neural networks that will be used to predict a word's presence in a given document[180]. The resulting node weights then serve as the vector representation of the words. The resulting model was spotchecked by picking a few portions of a datum used for training (about 0.5%), introducing very minor changes, and checking for similarity with all other data. With a few attempts (seven, to be exact), all randomly picked text were scored to be most similar to their source data.

[180] *Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in* International Conference on Machine Learning, *PMLR, 2014, pp. 1188–1196.*

The vectors derived for each module might have different magnitudes or scales (e.g., some modules having more words). We are more interested in the "intrinsic" similarity rather than "sizes." The cosine similarity, which measures the angle between two vectors, was thus selected instead of the more popular Euclidean distance, which magnitudes can influence. Listing 3 shows a code snippet used for calculating similarities between modules.

Listing 3: Network Similarity Graph Code Snippet

```
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics.pairwise import euclidean_distances

tagged_data = [TaggedDocument(words=tokenizer.tokenize(doc), tags
    =[i]) for i, doc in enumerate(contents)]
```

```
model_d2v = Doc2Vec(vector_size=100,alpha=0.025, min_count=1)
model_d2v.build_vocab(tagged_data)

for epoch in range(100):
    model_d2v.train(tagged_data,
                total_examples=model_d2v.corpus_count,
                epochs=model_d2v.epochs)

document_embeddings=np.zeros((5,100))
for i in range(len(document_embeddings)):
    document_embeddings[i]=model_d2v.docvecs[i]

pairwise_similarities=cosine_similarity(document_embeddings)
```

The **cosine_similarity** function available in **scikit learn** was used for measuring similarity. The cosine similarity between two vectors X and Y can be expressed by Equation 11, where $\cdot$ is vector dot product and $*$ is scalar multiplication:

$$K(X, Y) = \frac{X \cdot Y}{(\sqrt{X \cdot X} * (\sqrt{Y \cdot Y})} \tag{11}$$

Figure 36 shows the result for the electrical engineering course. All values are between 0.42 and 0.51. In this figure, we decided to make the nodes connected if their similarity score is above 0.4 due to proximity of similarity scores, so all nodes ended up being connected.

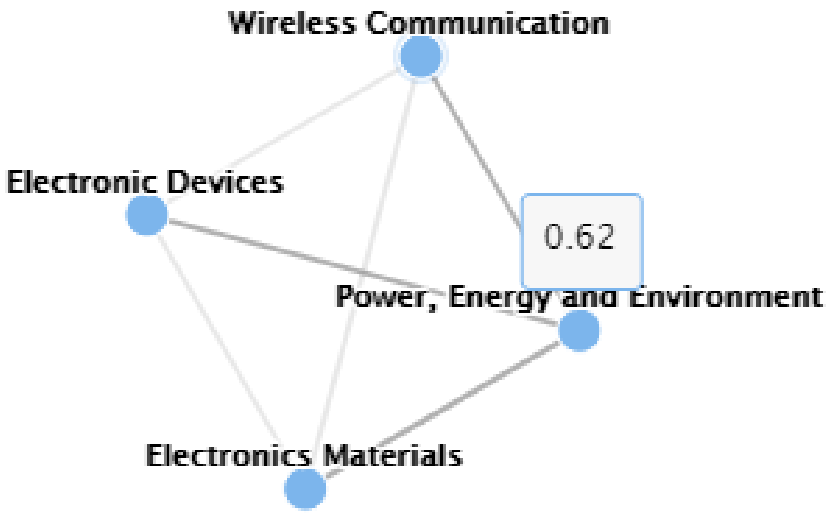


Figure 36: Resulting Similarity Network Graph. The value 0.4 was used as similarity cut-off

That is then the challenge with similarity network. What is the proper cut-off for similarity? We need to conduct similar experiments to have a better idea. Another challenge, not just for the similarity network but for all Analytics Dashboard visualization, is how do we make it multilingual? Currently, we are creating English dubs for the electrical engineering class, so we can expect to have more English responses in the future. We need to consider how to manage data that is a mixture of English and Japanese.

In this exercise, only the degree of similarity of the modules was considered. This makes the assumption that the current module organization is the best way to present the material. In some cases, the educator may want to check if the quality of the course content itself is sufficient. Other NLP-based quality assurance measures on course content such as clustering and readability analysis were done on MOOCs[181]. In this case, the educator may be interested if the content teaches the material they intend to teach. This was previously done through topic modeling, similar to what was discussed previously[182].

181 *M. K. J. Carlon, N. Keerativoranan, and J. S. Cross, "Content type distribution and readability of MOOCs," in* Proceedings of the Seventh ACM Conference on Learning@ Scale, *2020, pp. 401–404.*

182 *M. K. J. Carlon, A. D.D. C. Asa, N. Keerativoranan,* et al.*, "Topic modeling in MOOCs: What was to be discussed, what the instructor discussed, and what the learners discussed," in* 2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), *Submitted.*

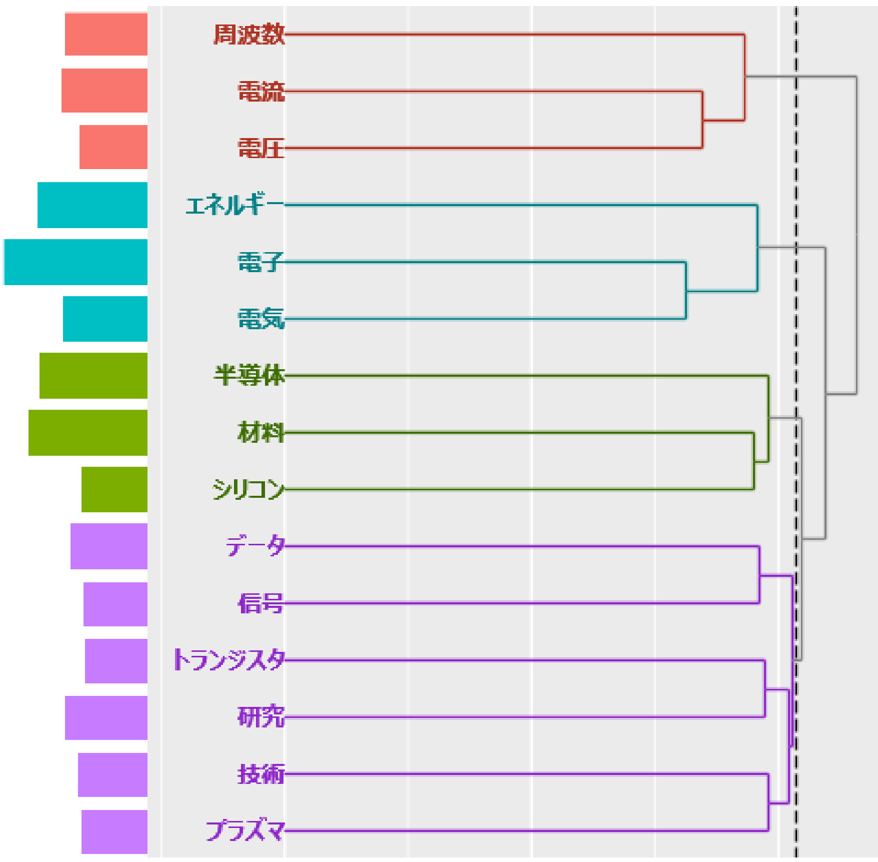


Figure 37: Hierarchical Clustering Graph. This was constructed using KH Coder with Ward's method and cosine distance.

Another point of interest could be on understanding the relationship of the concepts taught. Our current network graph is directionless; what if we put direction to it? Is it possible to say that Module 3 builds upon Module 1, but not the other way around? A possible visualization for this is a hierarchical clustering graph, where the direction can be from parent node to child nodes. Figure 37 was produced from this course's content using **KH Coder**, a tool created for quantitative text analysis[183]. Just like in topic modeling, we also specified the words that are to be excluded or stop words. In this case, we only included nouns for analysis and the word "説明" (explanation) was set as a stop word. Aside from the stop words, we also manually gathered the group of words discovered through the ChaSen[184] Japanese morphological parser that should always be taken together (e.g., 太陽電池: solar cell, 磁気浮上: magnetic levitation, and 高周波信号: high frequency signal, among others). Finally, we set several parameters in creating this graph: the minimum and maximum term and document frequencies for a word to be considered for clustering, the clustering method, and the distance measurement.

For a typical teacher to be able to produce this visualization, not only must they be guided in interpreting the results, but they must also be trained possibly extensively to understand all the variables that they can manipulate. Except for those who were privileged with time and resources to be adept with educational technologies, most teachers might be discouraged to use tools that require too much overhead. This phenomenon is in fact observed over and over again in several educational technologies, leading to the so-called **EdTech Matthew effect**, where educational technologies, even those available for free, lead to increasing inequalities since only the affluent benefit from them[185].

Other more established methods of graphing similarities are available such as graph constructions based on k-nearest neighbor algorithm[186] or topological representations[187]. These alternative options may even provide more information, such as intrinsic clusters of information. However, as we had seen during hierarchical clustering, these approaches may warrant more investigation to enable educators of varied backgrounds to manipulate the necessary parameters and interpret the results easily.

4.7 DESIGN CONSIDERATION: HUMAN-CENTERED AI

The Analytics Dashboard is based on the human-in-the-loop (HITL) design or interfaces that require human interaction. While HITL was previously attributed to increased problems in computer systems due to human errors[188], HITL is being increasingly recognized as a means to build fairer AI systems[189]. Most modern AI systems rely on data to produce the models needed for execution. Unfortunately, data collection can be prone to social biases. Freely available large-scale data such as those gathered from the internet, e.g., corpora used by Generative Pre-trained Transformer 3 (GPT-3), can reflect toxic behavior of unmoderated internet users[190]. This is not ideal for an educational setting. Hence, we highlight the importance of the teachers' expertise in using these AI models in the classroom.

[183] K. Higuchi, *"KH Coder 3 reference manual,"* Kioto (Japan): Ritsumeikan University, *2016.*

[184] *Y. Matsumoto, A. Kitauchi, T. Yamashita,* et al., *"Japanese morphological analysis system ChaSen version 2.0 manual,"* NAIST Technical Report, *1999.*

[185] *J. Reich, "Two stances, three genres, and four intractable dilemmas for the future of learning at scale," in* Proceedings of the Seventh ACM Conference on Learning@ Scale, *2020, pp. 3–13.*

[186] *W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in* Proceedings of the 20th International Conference on World Wide Web, *2011, pp. 577–586.*

[187] *L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for dimension reduction,"* arXiv preprint arXiv:1802.03426, *2018.*

[188] *L. F. Cranor, "A framework for reasoning about the human in the loop," in* Proceedings of the 1st Conference on Usability, Psychology, and Security, *Advanced Computing Systems Professional and Technical Association, 2008, pp. 1–15.*

4.8 USER STUDY

> Portions drafted as: M. K. J. Carlon and J. S. Cross, "Learning analytics dashboard prototype for implicit feedback from metacognitive prompt responses," in *29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education*, 2021

The POALS Analytics Dashboard was introduced to nine educators working in secondary schools, professional training, after-school support, and higher education (undergraduate and graduate) from Japan, the Philippines, the United States, and Finland. They have experience in face-to-face and hybrid formats, and one has experience in a fully online format. They answered a questionnaire made up of four parts: a written interview to inquire about their experiences engaging with learners and their opinions about recent educational trends; an introduction to POALS Metacognitive Tutor and Analytics Dashboard; a Likert scale based on the Technology Acceptance Model (TAM) with a rating from 1 – Strongly Disagree to 5 – Strongly Agree[191]; and a free-response form for further feedback. Some of the educator responses are quoted.

> It's really hard to monitor and measure learning. Sometimes even assessment scores or grades are not enough gauge of what someone has learned.

> The most frustrating is that each student has his or her own way of learning things and since the class is heterogeneous, it's somehow hard to gauge each student using one way of assessment.

The educators described their interactions with their learners primarily within the allotted face-to-face time during the written interview. The interactions are also mainly with the class as a whole instead of with individuals or small groups. Their usual strategy for eliciting engagement is through exhibiting a positive outlook (e.g., asking for their learners' welfare, being cheerful, and others) and fostering discussions. These interactions allow them to assess their learners' understanding even before homework or quizzes are given. Their discussion-based assessments can be through their instructional design (e.g., one educator described their typical lesson to include a warm-up, a build-up, and an expansion) or through non-verbal communication (e.g., facial expressions). These engagements allow them to detect when the learners are not picking up as expected (e.g., learners who could move up a level after passing a critical exam without understanding the material). Early identification of concerns enables them to adjust their instructional materials or provide remedial support. In rare cases, their final resort includes consulting other educators about their learners' progress.

> Less participation to [sic] students as the environment students are in (mostly in their house) is beyond my control.

[189] *F. M. Zanzotto, "Human-in-the-loop artificial intelligence,"* Journal of Artificial Intelligence Research, *vol. 64, pp. 243–252, 2019.*

[190] *E. Strickland. (Feb. 2021). OpenAI's GPT-3 speaks! (Kindly disregard toxic language), [Online]. Available:* `https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business`.

[191] *F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," PhD thesis, Massachusetts Institute of Technology, 1985.*

> Teaching online was very challenging. Other than having to transform face to face materials to online teaching, there is also the internet connectivity issues and leaning [sic] the online platform mandated by the university.

With the onset of the COVID-19 pandemic, most of them had allocated timeslots for synchronous discussions and provided asynchronous communication channels when face-to-face interaction was prohibited. Most of them found that their learners are struggling with their lessons due to various reasons: hardware and connection issues, online instructional materials constructed in a rushed manner by a team that did not include the educator, and home environmental issues. The problem that frustrates the educators the most is the seeming lack of participation (e.g., learners being off-camera and not responding) or even private communication that could help them assist their learners with lesson difficulties. This is despite them providing multiple venues to be accessible to their learners (e.g., LMS, videoconferences, chats, recordings, emails, telephone, and others).

> There are cases that students simply resort to copying from the internet to answer questions/problems. Also, students are not yet ready to do independent learning.

Despite teaching in the online format for more than a year due to the COVID-19 pandemic, the respondents expressed difficulty assessing learning and engagement without face-to-face time with the learners. The situation has become more challenging by an increased risk of dishonesty as learners search for answers online and schools deciding to remove examinations to ease the learners' pressure.

The educators were introduced to POALS after the written interview. More specifically, the Metacognitive Tutor screens were shown and their purposes explained. Afterwards, a mock-up of the POALS Analytics Dashboard (see: Figure 38) with a higher fidelity than the wireframe is shown. Each portion of the mock-up was also explained. All screens are shown as static view (i.e., the educators cannot interact with them), though they were advised they can ask for clarifications anytime. Details of the POALS introduction for the user study are in Section B.4.2 Introducing POALS. The educators are then asked to assess how easy to use or how usable they perceive the POALS Analytics Dashboard to be using the modified TAM.

Figure 39 shows the boxplots of the modified TAM responses. All the respondents agreed that the POALS Analytics Dashboard could help them respond to their learners' unvoiced needs and assess their progress. The only item that had a mean below 4 – Agree is the perception of maximizing the use of the tool. The respondents should be given more time to use the tool in their classes to understand this better.

When asked to elaborate their opinions about the POALS Analytics Dashboard, they appreciate being able to check their learners'

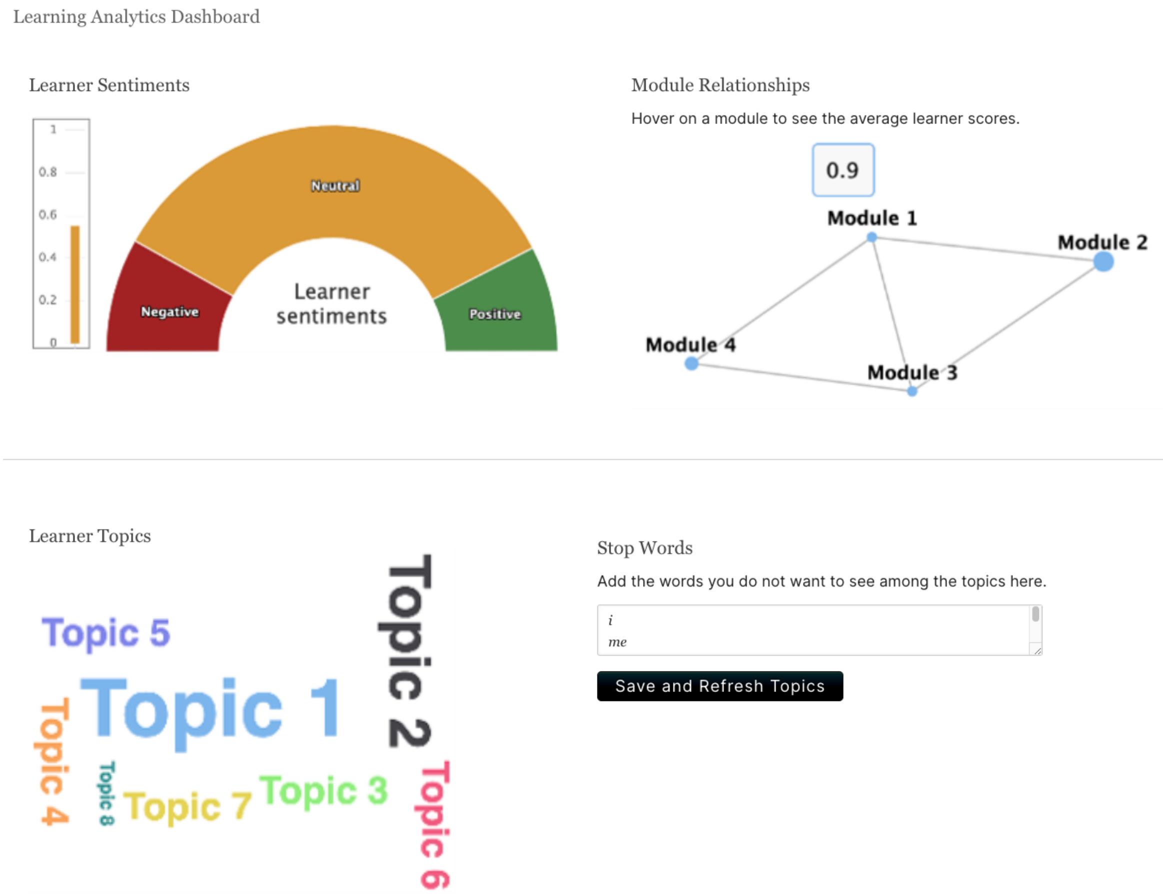
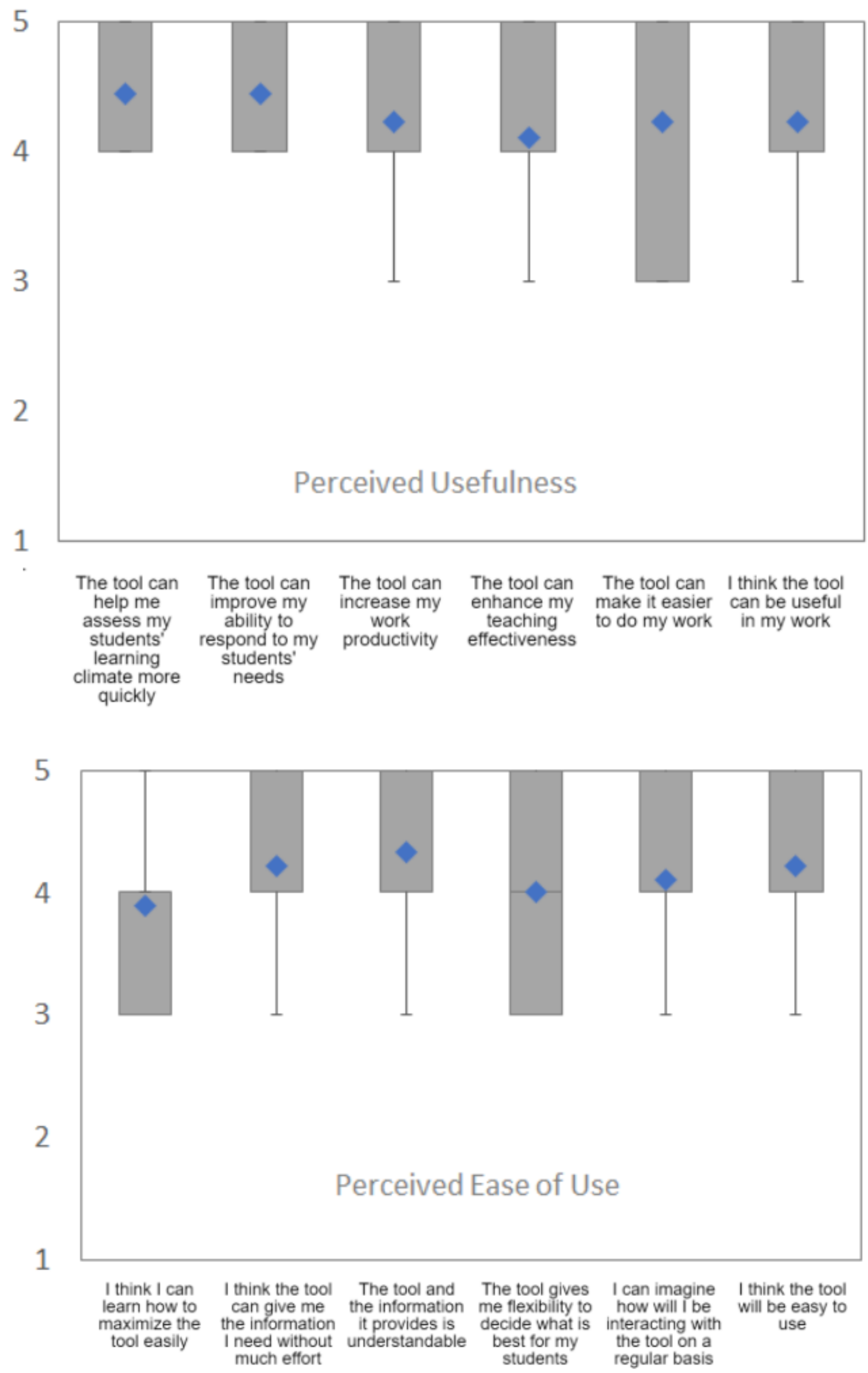


Figure 38: Analytics Dashboard High Fidelity Mock-up. The educators are shown this mock-up after the POALS Metacognitive Tutor screens are introduced. Each part of the mock-up is introduced afterwards.

Figure 39: Analytics Dashboard TAM Results. Boxplots of modified TAM results with means illustrated as blue diamonds.



progress on-demand. They think the word cloud feature can help assess how the learners scaffold their learning. An added strength of the tool is that it gives insights into the learners' feelings which may not be vocalized. However, it is susceptible to accessibility problems that are common with any other online tool (e.g., learners having connectivity problems). One educator commented that they would appreciate being given tips on how to address the potential problems detected by the Analytics Dashboard, but this may lead back to the problem identified by previous research where too much information on LADs stifles teachers' creativity[192]. As such, further research is needed to identify an optimal amount of information for LADs.

[192] *M. Brown, "Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data,"* Teaching in Higher Education, *vol. 25, no. 4, pp. 384–400, 2020.*

> Informative of students' feelings regarding lessons so the teacher will know what to do. These feelings are usually not expressed without a tool because students tend to be shy to admit that they don't understand something.

> The word cloud feature and the ability to assess how well students are scaffolding their learning.

4.9 CONCLUSIONS ON THE ANALYTICS DASHBOARD

The POALS Analytics Dashboard is a collection of visualizations that are grounded on teacher intuition and learning theories. While the Analytics Dashboard provides details that can be used for detecting poor sentiments, misconceptions, and surface learning, the judgment of whether interventions should be made or not is left to the teacher by design. This not only taps the teacher's expertise that could lead to a better learning experience but also prevents inadvertent adverse effects of AI unfairness that is yet to be uncovered.

We learned that creating the POALS Analytics Dashboard for at-scale consumption is mostly feasible (i.e., can be done using readily available software with minimal tweaking). Furthermore, our user study reveals that the POALS Analytics Dashboard can be helpful to teachers. This is important for online learning environments where the teachers may find it hard to assess the learners' engagement and progress from mere observations.

4.10 FUTURE WORK

Some future work identified includes finetuning sentiment analysis for educational settings, exploring alternative similarity network visualizations, and creating a minimum viable product (MVP) that can be tested in an actual setting. User study indicates a positive outlook towards the concept; thus, developing further the POALS Analytics Dashboard (e.g., multiple language support, more interactivity, and others) is worthwhile.

5 CONCLUSION

> *"Who are YOU?" said the Caterpillar.*
> *This was not an encouraging opening for a conversation.*
> *Alice replied, rather shyly, "I–I hardly know, sir, just at present– at least I know who I WAS when I got up this morning, but I think I must have been changed several times since then."*
>
> –Lewis Caroll, *Alice in Wonderland*

Several challenges were overcome in order to get to this point. The Personalized Online Adaptive Learning System (POALS) needed to be developed almost from scratch to ensure that necessary precautions (e.g., maintainability, security, and others) are taken care of. After the coding was completed, necessary changes were identified during the pilot study. Additionally, unforeseen changes happened in the related systems, such as the operating systems used for development, the learning management system (LMS) implementation, and the hardware that required considerable time and effort to accommodate. Earlier recruitment attempts were also a struggle, and when a feasible subject body was finally identified, the COVID-19 pandemic hit, disrupting the higher education schedule in Japan. While there are still many possibilities left to be explored, several vital results were uncovered.

5.1 SUMMARY OF RESULTS

Through literature review, we established a need for learners to develop metacognitive skills to succeed in online learning environments. Success in online learning environments is crucial since we enter an age where there is a need for constant skill-building. In addition, online learning is a potentially efficient way for institutions to provide necessary knowledge on a large scale.

5.1.1 *Are open-response prompts effective in developing metacognitive skills on an online learning platform?*

We analyzed how an existing domain-independent metacognitive tutor, the Reflection Assistant (RA), can be optimized for online learning platforms. We constructed POALS Metacognitive Tutor to increase usability on LMS. The development of POALS overall also took into account the need for information security, which is critical for any tool accessed via the internet.

POALS Metacognitive Tutor was first pilot-tested on an undergraduate class on educational technology. We learned from the pilot study that there is a potential for POALS Metacognitive Tutor to be effective in improving learner metacognition. However, there are concerns about cognitive load, as seen from the lack of learner engagement on the open-response prompts.

The prompts were simplified, and a full experiment was conducted on an undergraduate class on electrical engineering. Results reveal that while POALS Metacognitive Tutor can be beneficial to developing regulation of cognition (ROC), it may not be sufficient for improving knowledge of cognition (KOC). There is a potential increase in performance when using POALS Metacognitive Tutor, but there is no statistical significance on this front. A possible approach that can be taken in the future is introducing a short lesson on metacognition to help the learners appreciate the value of planning for and reflecting on their knowledge.

5.1.2 *Can we use innovative ways to improve knowledge tracing algorithms for adaptive learning?*

Metacognitive tutoring involves using cognitive resources on at least two fronts: the cognitive and the metacognitive domains. While there is an attempt to reduce cognitive load on the metacognitive front, managing cognitive resources on the cognitive front are still essential. A popular approach for managing cognitive resources on cognitive endeavors is applying knowledge tracing on learning exercises. The typical knowledge tracing approach involves estimating learner mastery of the topic and allowing the learner to move on to succeeding topics once they demonstrate mastery instead of enduring more exercises than necessary. On the flip-side, the knowledge tracing may prevent learners from moving on without mastering the current topic, which can, in turn, result in unwarranted cognitive load due to difficulty applying a prerequisite knowledge later on.

We selected an interpretable algorithm, Bayesian Knowledge Tracing (BKT), and an explainable algorithm, Deep Neural Network (DNN), and customized these to use inputs from POALS Metacognitive Tutor in creating POALS Adaptive Engine. Models were created using a synthetic dataset based on predefined learner personas. We learned that these existing algorithms perform better with the additional inputs from POALS Metacognitive Tutor than without those. While the DNN-based model has high accuracy, the BKT-based model had equally competitive accuracy while following learning intuitions.

With the above result, we resolve to use the interpretable algorithm for POALS Adaptive Engine. We saw that using the interpretable version did not require significant sacrifice in terms of accuracy. For one, interpretable algorithms do not require as much data as explainable

__Interpretability__ is the extent to which a cause-and-effect can be seen from the system. For example, in BKT, the expected parameters __Guess__, __Slip__, __Initial Learning__, __Acquisition__, and __Forget__ were already decided beforehand. The resulting models then make it easier for the user to see how the inputs are used in estimating the parameters.

__Explainability__ is the extent to which the results can be justified. A typical approach to this is creating a second post-hoc model to explain the primary model.

algorithms. While this puts us at a disadvantage of possibly not uncovering more hidden patterns, this can help us ensure that the algorithmic results will always be understandable and faithful to the intuitions used behind modeling[193].

[193] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

5.1.3 *Can we find alternative sources of feedback to assess the effectiveness of an online class?*

Even learners who can learn autonomously can benefit from teacher guidance. However, the teachers need some feedback from the learners to plan for the best interventions to optimize learning. This can be not easy to obtain in an online learning environment where not every learner is visible to the teacher.

POALS Analytics Dashboard shows data visualizations derived chiefly from the learner inputs in POALS Metacognitive Tutor. POALS Analytics Dashboard can show the learner sentiments to allow the teacher to address negative concerns promptly. It also shows the emerging topics from the open response prompts to detect potential misconceptions and the module similarities with the average performance in each module to uncover unwanted surface learning.

5.2 CONTRIBUTIONS

The most obvious contribution of this work is the creation of POALS itself. Once fully implemented, POALS can be a tool for any online learning configuration that provides support to both the learners and the teachers. From a more theoretical point of view, we were able to explore the following research directions.

5.2.1 *It reassesses the effectiveness of metacognitive prompting on an online learning environment.*

The RA, a tutor for an interactive learning environment and basis of this research, was conceptualized in the early 2000s. Since then, the world has experienced significant advances in technology. For instance, Google had its initial public offering in August 2004, and the first iPhone was released in June 2007. From then on, the way we handle technology and the types of jobs becoming available has drastically changed[194]. In addition, more and more classrooms are moving online[195], and as such, it is imperative to re-investigate computer-based learning interventions. However, since massive open online courses (MOOCs) are introduced, the RA was not revisited to the best of our knowledge.

[194] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" Technological Forecasting and Social Change, vol. 114, pp. 254–280, 2017.

[195] D. Shah. (Dec. 2020). The second year of the MOOC: A review of MOOC stats and trends in 2020 - class central, [Online]. Available: https://www.classcentral.com/report/the-second-year-of-the-mooc/.

5.2.2 *It uses an affective measure (learner's personal assessment) and not just performance in modeling the knowledge tracing algorithms.*

According to the Emotion-Involved Processing Hypothesis, learning is just as much emotional as it is cognitive[196]. For example, in the POALS Metacognitive Tutor, we ask for the learners' confidence in solving a problem successfully. While confidence ratings maybe not be entirely based on emotions, they can manifest learner anxieties and attitudes. The POALS Metacognitive Tutor's Learner Profile also keeps track of the learners' outlook score, which shows a learners' pessimism or optimism. The literature says that too much of either would be detrimental to the learning experience.

With the POALS Adaptive Engine, we explored using the confidence ratings and the outlook scores in creating knowledge tracing algorithms. While there are several attempts to create more personalized knowledge tracing algorithms[197,198,199], we are yet to find similar work embedding affective measures in knowledge tracing.

5.2.3 *It explores alternative sources of course quality feedback.*

As learning analytics continues to be a growing research field[200], more and more research studies have explored not just learning dashboard but various ways to discover learning experience improvement opportunities. NLP-based approaches are also very common, mostly in analyzing discussion board posts[201,202,203]. While these approaches are a step forward from the usual end-of-course surveys as sources of course quality feedback, these can still be lacking as the majority of learners opt not to post on discussion boards for various reasons[204].

With the POALS Analytics Dashboard, the source input, which is the metacognitive prompts, are visible to the learners and the select teachers only. The classmates will not see it, so it does not have the same problem as the discussion forums where the learners might be too shy to post. Also, we are getting implicit feedback or feedback that the learners might not tell the teachers directly because the learners themselves do not realize that they have such difficulties. This implicit feedback can be missing from the equally private end-of-course surveys.

5.2.4 *It creates a direct link between metacognitive skills development, learning analytics, and adaptive learning.*

With POALS, we can foster metacognitive development through the Metacognitive Tutor, use metacognitive inputs as part of knowledge tracing through the Adaptive Engine, and apply NLP techniques to derive implicit feedback from metacognitive reflections to be displayed in the Analytics Dashboard. We are confident that none has worked

[196] *Y. Kanazawa, "Emotion as "deeper" than cognition: Theoretical underpinnings and multidisciplinary lignes de faits to the Emotion-Involved Processing Hypothesis (EIPH)," 国際学研究=* Journal of International Studies, *vol. 9, no. 1, pp. 185–206, 2020.*

[197] *M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," in* International Conference on Artificial Intelligence in Education, *Springer, 2013, pp. 171–180.*

[198] *P. Nedungadi and M. Remya, "Incorporating forgetting in the personalized, clustered, Bayesian knowledge tracing (PC-BKT) model," in* 2015 International Conference on Cognitive Computing and Information Processing (CCIP), *IEEE, 2015, pp. 1–5.*

[199] *J. Zhao, S. Bhatt, C. Thille,* et al., *"Interpretable personalized knowledge tracing and next learning activity recommendation," in* Proceedings of the Seventh ACM Conference on Learning@ Scale, *2020, pp. 325–328.*

[200] *V. Kovanovic, M. Scheffel, N. Pinkwart,* et al., *"LAK 2020 program chairs' welcome," in* ACM International Conference Proceeding Series, *2020, p. iii.*

on the same problem as we did since we are building on the work we have started.

5.3 SOCIETAL IMPACT

Intuitively, we can say that artificial intelligence (AI) will have a significant impact in the future of education primarily because AI is a fast-growing field. Additionally, as saw through the COVID-19 pandemic, educational technology must be continuously improved to anticipate future contingencies. But actually, AI in education has been at peak hype even in the years before the COVID-19 pandemic.

Figure 40 shows the Gartner hype cycle charts created based on the analysis provided by the e-learning website Web Courseworks at the start of every year[205] from 2018. The Gartner hype cycle is a marketing tool used to assess the public perception of a technology trend[206]. From 2018 to 2020, AI has been at the **peak of inflated expectations**, indicating great interest and publicity. Another technology that is related to our research is MOOCs which is at the **plateau of productivity** in 2018, possibly due to the maturity of MOOCs as a delivery method for a few graduate degrees. In 2019, MOOCs regressed to the trough of disillusionment, potentially due to not being able to catch on in a corporate setting. Another technology of interest to us is analytics which appeared in the **innovation trigger** in 2019, progressing to the peak of inflated expectations in 2020.

At the onset of the COVID-19 pandemic, most educational institutions resorted to emergency remote online learning, which deviated significantly from online learning practices that were established to be effective[207]. Similarly, workplaces shifted to remote work (mostly work-from-home) schemes. These changes were reflected in the discussion around educational technology hypes. In Figure 41, we can see that people are starting to discuss more how workplaces will change from hereon and how learning can still be social despite the lack of face-to-face interactions.

Even though MOOCs, AI, and analytics were no longer hyped, the need for these technologies during and after the COVID-19 pandemic remains evident. The emerging interest in social learning can lead to improvements on MOOCs and can be supported by both AI and analytics. After all, despite the seemingly negative reception to online learning during the pandemic, its benefits are still visible. Online learning made learning accessible to several learners (e.g., those with specific disabilities, those who need to work alongside studying, and others) who would otherwise not continue studying. Now that we learned that online learning could be possible in many cases, it may be here to stay.

However, as pointed out by our respondent educators in our POALS Analytics Dashboard user study, online learning has several issues

[201] *M. Wen, D. Yang, and C. Rose, "Sentiment analysis in MOOC discussion forums: What does it tell us?" In* Educational data mining 2014, *Citeseer, 2014.*

[202] *E. Huang, H. Valdiviejas, and N. Bosch, "I'm sure! Automatic detection of metacognition in online course discussion forums," in* 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), *IEEE, 2019, pp. 1–7.*

[203] *R. Setiawan, W. Budiharto, I. H. Kartowisastro,* et al., *"Finding model through latent semantic approach to reveal the topic of discussion in discussion forum,"* Education and Information Technologies, *vol. 25, no. 1, pp. 31–50, 2020.*

[204] *D. F. Onah, J. E. Sinclair, and R. Boyatt, "Exploring the use of MOOC discussion forums," in* Proceedings of London International Conference on Education, *2014, pp. 1–4.*

[205] *Web Courseworks. (2021). eLearning hype curve predictions, [Online]. Available:* https://webcourseworks.com/elearning-predictions-hype-curve/.

[206] *A. Linden and J. Fenn, "Understanding Gartner's hype cycles,"* Strategic Analysis Report Nº R-20-1971. Gartner, Inc, *vol. 88, 2003.*

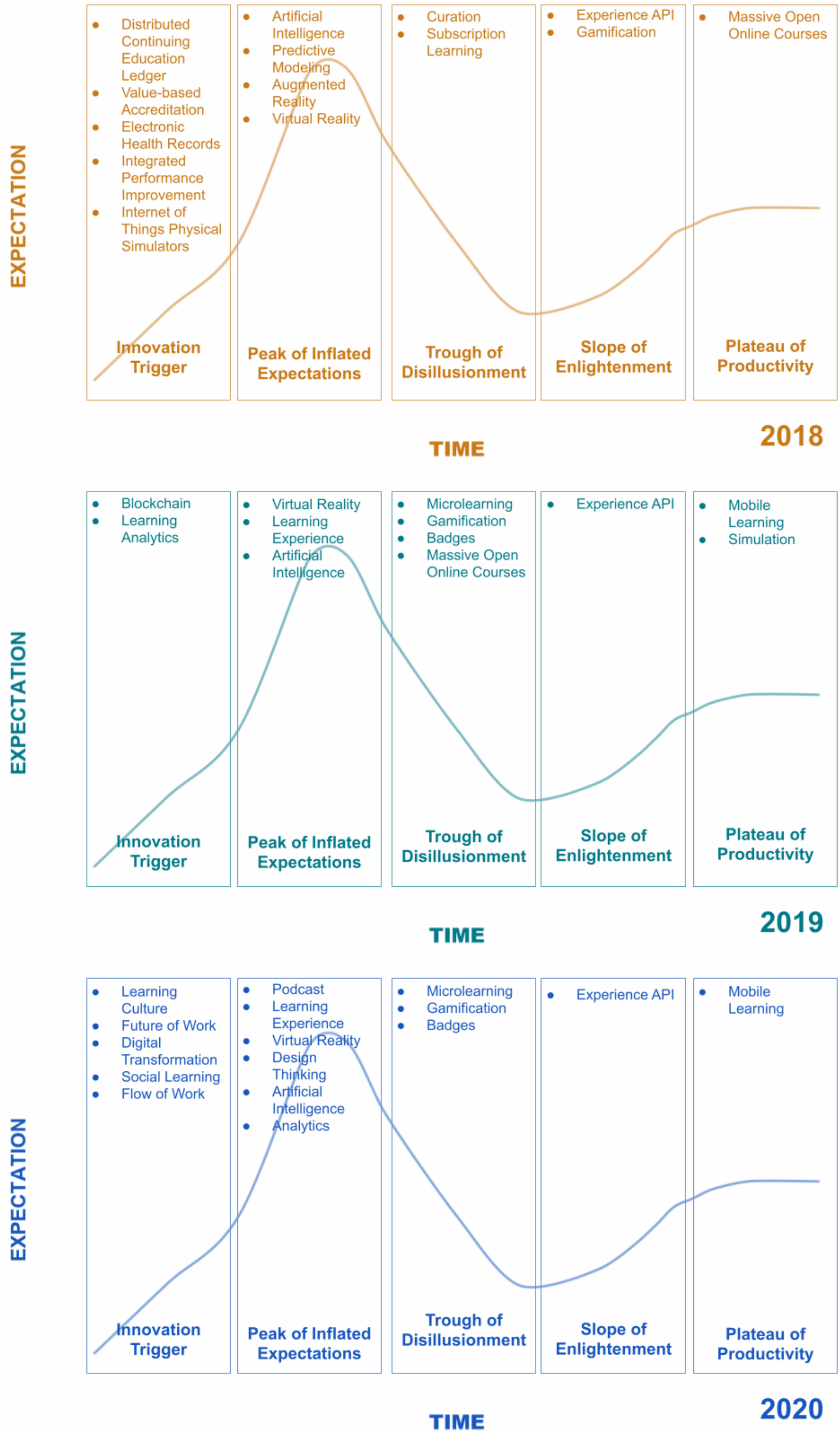


Figure 40: Pre-pandemic Hypes. Educational technology yearly hype predictions prior to COVID-19 (Web Courseworks, 2021).

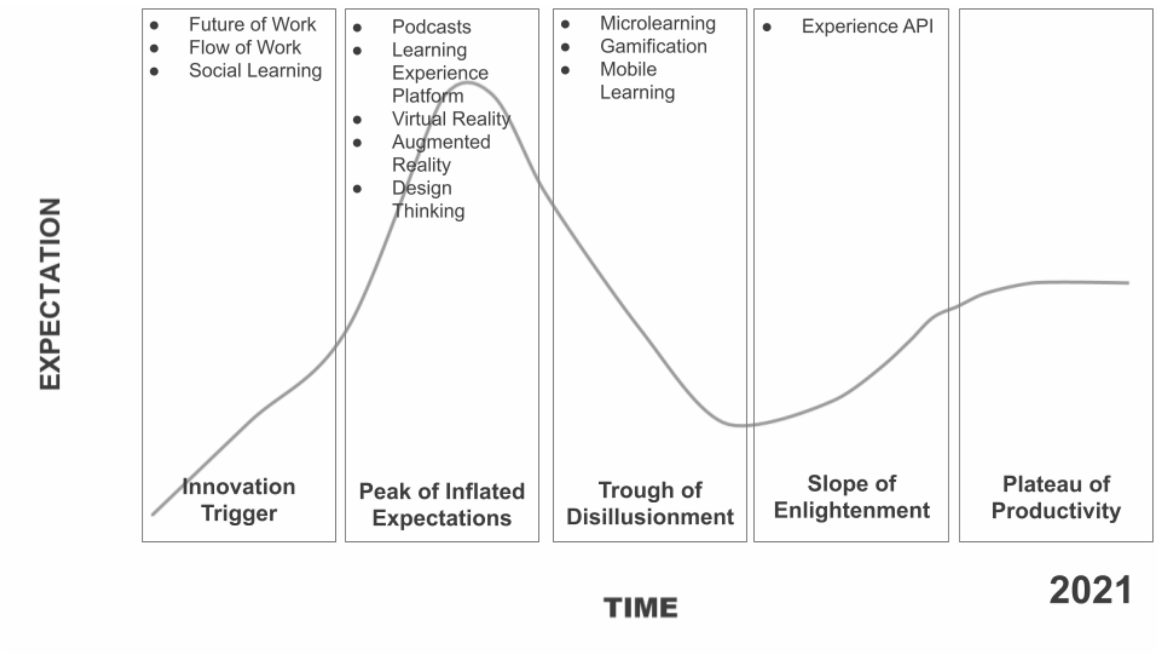


Figure 41: 2021 Hype. Educational technology hype prediction, a year after COVID-19 outbreak (Web Courseworks, 2021).

[207] C. Hodges, S. Moore, B. Lockee, et al., *"The difference between emergency remote teaching and online learning,"* Educause Review, *vol. 27, pp. 1–12, 2020.*

extending from lack of resources to lack of skills to succeed in the said format. While some concerns such as interconnectivity problems are not within our control, we learning scientists can still improve online learning environments through robust instructional design and pedagogically sound educational technologies. Figure 42 in particular summarizes how POALS can contribute to online learning success.

5.4 REFLECTIONS

A work on metacognition would be remiss without spending a few moments to reflect on what had been the weak points. These include the eventual limitations as well as inefficiencies not known from the start.

5.4.1 *Limitations*

POALS Metacognitive Tutor still needs to be tested with learners of varied demographics and cognitive domains with and without explicit instruction on metacognition. The conundrum between metacognition's generality and specificity across domains remains to be a challenge in metacognition research[208]. Another still unresolved challenge in metacognition research is establishing the relationship between metacognition and individual differences. In addition, conditions for instructing and acquiring metacognition are yet to be established. Lastly, a follow-up study conducted long after the end of the experi-

[208] R. Azevedo, *"Reflections on the field of metacognition: Issues, challenges, and opportunities,"* Metacognition and Learning, *Jun. 2020.*

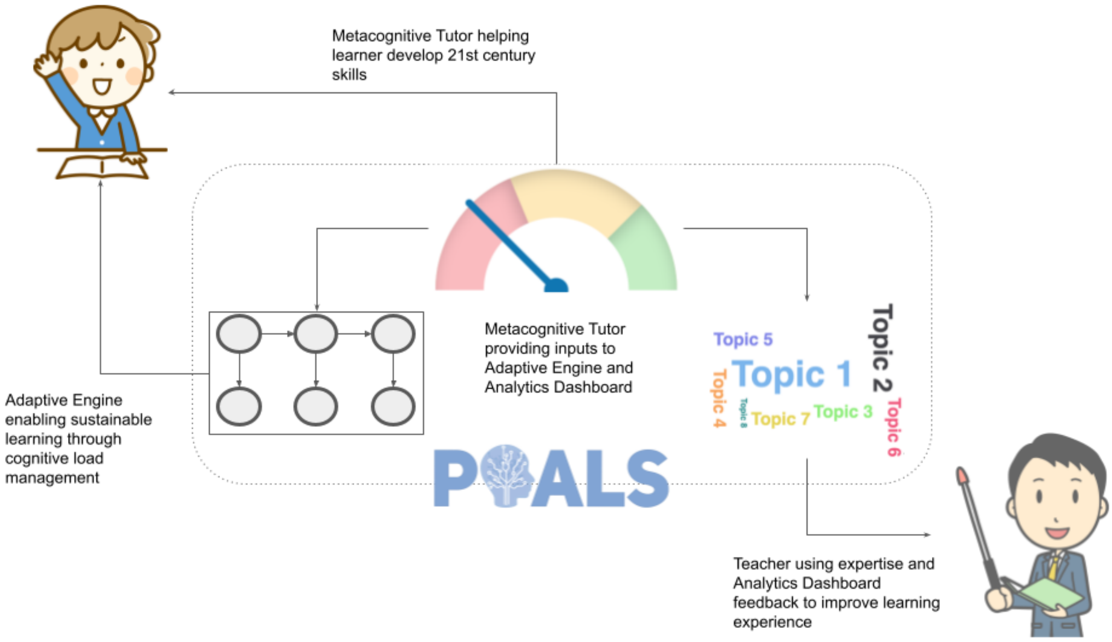


Figure 42: Learner and Teacher Interaction with POALS. POALS can help alleviate some issues associated with online learning by providing learners with learning environment sustainable for developing 21st century skills, and enabling teachers to use their expertise in supporting their learners.

ment can help us ascertain the effectiveness of POALS Metacognitive Tutor in developing metacognition in the long term.

POALS Adaptive Engine is yet to be modeled using actual learner data and tested in an actual online learning environment. Additionally, the fairness of the resulting algorithms should be verified, considering that the relationship between individuals and metacognition is still an open question. Formal methods in verifying explainability of resulting deep learning (DL) based models must be explored.

A minimum viable product (MVP) is yet to be created for POALS Analytics Dashboard. At the core of the user-centered design is eliciting user feedback. It will be remiss for a system designed to help users gather feedback not to allow its users to give feedback on the system itself.

5.4.2 *What could have been done?*

All the instructors approached for this research had been very supportive and accommodating. They were looking forward to positive results and remained patient despite several issues encountered. Unfortunately, not all results had been resoundingly positive. Better outcomes could have been achieved with closer coordination with the instructors. The software could have been designed with the teacher in mind from the start. That way, the teachers could have been empowered to create more fun and challenging exercises that will fit well with metacognitive development.

This research used design-based approach to be able to introduce new interventions as necessary. However, this has not been well taken advantage of: the intervals between interventions (from pilot study to full experiment) had been too long, and the interventions had not been aggressive enough. The need for metacognitive instruction was suspected from the pilot study result and was reinforced by literature review. However, instead of introducing metacognitive instruction, the more conservative approach of changing the metacognitive prompts was done.

Finally, there had been significant inefficiencies with the software development. The original software was designed to work as a plugin for the edX platform. Having a plugin for research though will require constant external support from edX, so it was decided to overhaul the software to support Learning Tools Interoperability (LTI) instead. This had the positive effect of being able to support other LMSs as well as develop a stand-alone version. However, there had been a huge urge at the start to keep as much of the old software as possible. Eventually, since the original plugin was very tightly-coupled with the edX platform, most of it were not reusable. The resulting software had been a patchwork that was difficult to maintain. This had a huge cost as maintenance became frequent as the main LMS, edX, continuously introduced updates to their platform.

5.5 FUTURE DIRECTIONS

Of course, the first order of business is to address the limitations identified. Further enhancements can also be introduced, such as providing tutorials on how to benefit the most with POALS, giving short introductory lessons on metacognition, and tapping more recent advances in multilingual NLP. The validation process for these enhancements can also be improved by making the target users more involved through participatory research, focus group discussions, think-alouds, and interviews. New features such as showing knowledge tracing explanations can be explored. Participant recruitment and completion rates should also be improved. These may be done by increasing incentives for full participation or designing more targeted experiments that can be completed with less time and effort. After these improvements, we can zoom out and see where POALS lies in the bigger picture.

Metacognition is an important component of self-regulation, or ones' ability to plan, monitor, and adjust activities to cater to environmental conditions. Figure 43 shows metacognition as related to other self-regulation components self-efficacy, motivation, and behavior[209]. The Cross Laboratory in the Department of Transdisciplinary

[209] *L. D. Frazier, B. L. Schwartz, and J. Metcalfe, "The MAPS model of self-regulation: Integrating metacognition, agency, and possible selves,"* Metacognition and Learning, *pp. 1–22, 2021.*

Science and Engineering, School of Environment and Society, Tokyo Institute of Technology, have several works relating to both metacognition and self-efficacy[210,211,212]. What remains to be investigated are behavior and motivation.

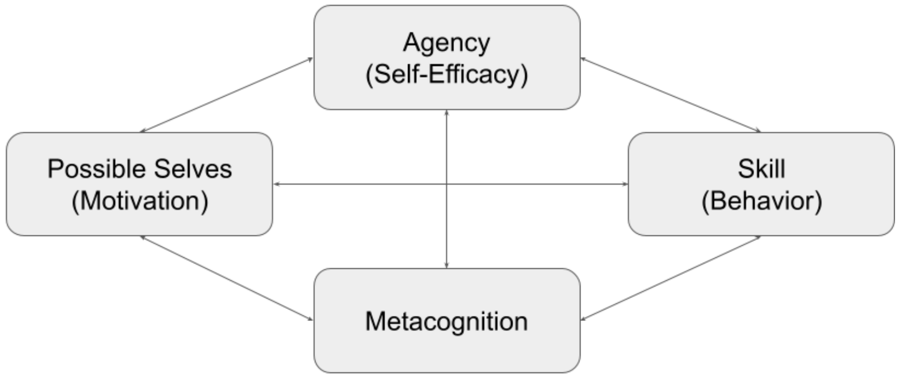


Figure 43: MAPS Model. Self-regulation as the relationship between metacognition, agency (self-efficacy), possible selves (motivation), and skill (behavior).

Nudges, or invoking positive reinforcement through indirect suggestions, cover a body of research in behavioral economics to affect human decision-making under challenging situations[213]. Thus, nudges affect both behavior and motivation. There are existing research studies on nudges in education[214]. Hence it could be interesting to know how can POALS possibly accommodate nudging to be a complete environment for self-regulation training.

From our experience with the Metacognitive Tutor experiments, finding participants committed to participating for an extended period can be highly challenging. Shorter research protocols (e.g., planning short lessons that can be completed in a few hours) may be designed at least once to gather enough data to test hypotheses to address this concern. Once results are validated and deemed usable in a classroom setting, it is recommended that teachers make a short discussion about the importance of self-regulated learning and how the learners can use the tools available to their advantage.

[210] *C. Seng, M. K. J. Carlon, and J. S. Cross, "Information literacy training effectiveness on Cambodia's province-based undergraduates,"* International Journal of Comparative Education and Development, *2020.*

[211] *C. Seng, M. K. J. Carlon, and J. S. Cross, "Research self-efficacy of Cambodian undergraduate students at province-based universities,"* International Journal of Sociology of Education, *vol. 9, no. 2, pp. 155–190, 2020.*

[212] *M. K. J. Carlon, C. Seng, and J. S. Cross, "Countering negative Matthew effect in undergraduate research with metacognition and digital learning," in* The 2020 Annual Fall Conference of Japan Society for Educational Technology, *2020.*

[213] *R. H. Thaler, "From cashews to nudges: The evolution of behavioral economics,"* American Economic Review, *vol. 108, no. 6, pp. 1265–87, 2018.*

[214] *M. T. Damgaard and H. S. Nielsen, "Nudging in education,"* Economics of Education Review, *vol. 64, pp. 313–342, 2018.*

A ABOUT THE SOFTWARE

> *Alice laughed. "There's no use trying," she said: "one can't believe impossible things."*
> *"I daresay you haven't had much practice," said the Queen. "When I was your age, I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."*
>
> –Lewis Caroll, *Through the Looking Glass*

As far as we can tell, each component of the Personalized Online Adaptive Learning System (POALS) is unique with no available similar applications existing both in the market and in research. Nevertheless, extra care was taken to reduce the development amount and increase POALS' compatibility with existing systems. When POALS Metacognitive Tutor was first conceived, it was developed as an *XBlock*, a plugin specific to the edX platform[215]. This was developed explicitly for edX to be easy to implement. Hence all the developer has to do is follow the examples given in the XBlock tutorial.

[215] *EdX Inc. (2021). Open edX XBlock tutorial, [Online]. Available:* `https://edx.readthedocs.io/projects/xblock-tutorial/en/latest/`.

Using the XBlock platform works well for organizations that manage their edX instances. However, there is additional overhead when another organization manages the edX instance. Intervention from the instance manager will always be needed when making updates to the XBlock. This is the case of TokyoTech, where the instance is managed by edX itself. Also, the look-and-feel and interactions are somehow limited by what is provided by edX. This is not ideal for a research project that expects frequent changes and requires unique learner interactions.

The original Metacognitive Tutor was re-developed to be a Learning Tools Interoperability (LTI) provider. LTI is an interoperability standard that allows communication across different educational technology software. Converting to an LTI provider means developing a new application from scratch with minimal framework support from existing tools such as edX. With the switch to being an LTI provider, it became possible to make more complicated software, giving birth to POALS. Being an LTI provider, POALS can be used not just on edX but on other educational technology supporting LTI such as Canvas, Moodle, and Blackboard among others[216]. Thus, the initial roadblock has become an enabler for other research activities.

[216] *IMS Global Learning Consortium. (Apr. 2019). Learning tools interoperability core specification 1.3, [Online]. Available:* `http://www.imsglobal.org/spec/lti/v1p3/`.

A.1 SYSTEM ARCHITECTURE

Presented as: M. K. J. Carlon and J. S. Cross, "Challenges of developing a metacognitive tutor on Open edX," in *Fourteenth Asia-Oceania Top University League on Engineering Student Conference*, 2019

The Python web programming framework Django (version 1.11) was used for POALS development. The Django framework comprises an application server that handles interaction with the users (in our case, the learners) and a database server where all data are saved. POALS is a web application that can work as a stand-alone quiz application or as an LTI provider to any LMS that can act as an LTI consumer. The Python library django-lti-provider (version 0.3.3) was used for the application server to enable the linking between POALS and other educational technology tools[217]. The use of LTI compliant connection ensures that the student information remains private and that only those given access to POALS can access it. For the case where POALS functions as stand-alone software, Django's oauth toolkit was used to provide the same functionality.

[217] *Columbia Center for Teaching and Learning. (Feb. 2018). django-lti-provider version 0.3.3, [Online]. Available:* `https://github.com/ccnmtl/django-lti-provider/commits/0.3.3`.

Information security was also accounted for during the development of POALS as can be seen from its system architecture in Figure 44. The container platform Docker was used to compartmentalize POALS. Separate containers are dedicated for the application server (running with Gunicorn) and the database server (running with PostgreSQL). Compartmentalizing the servers prevents the entire system from crashing when one of its components fails. This ensures fast recoverability in case of technical failures since only the point of failure needs to be revived. The contents of the database server are also regularly backed up to reduce information loss in case of failures (i.e., at most, only the data from the last backup point to the point of failure will be lost). Furthermore, communication between the servers and the internet passes through a reverse proxy (running with NginX). The reverse proxy performs load balancing to prevent serious technical failures from happening in the first place. Finally, all communications pass through a Secured Sockets Layer (SSL) (running with LetsEncrypt) to ensure that all communications are encrypted to prevent information theft.

A.2 APPLICATION SERVER

The application server serves the interfaces for the stand-alone version and the LTI provider. Thus, the application server hosts two applications: the LTI authenticator and the main application. In addition, the application server also hosts Django applications such as the admin page, which can be accessed by users assigned with administrator rights and oauth for providing security for the stand-alone version.

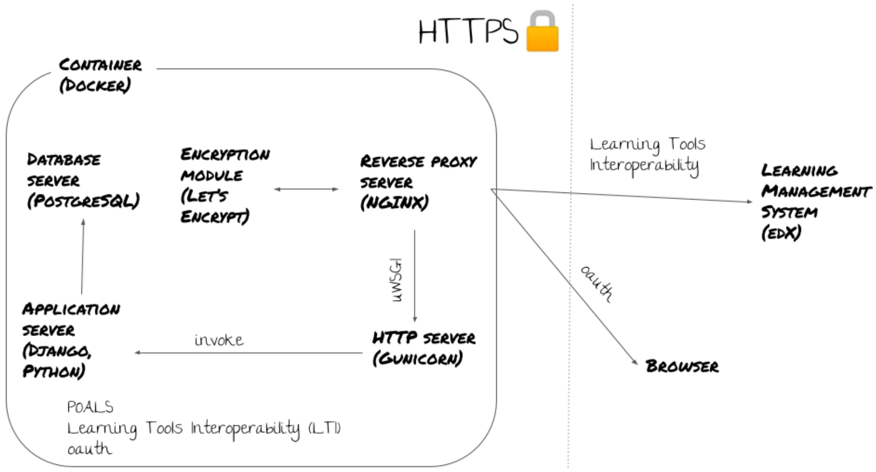


Figure 44: System Architecture. Various technologies are used in POALS to protect information security.

Upon logging in as a stand-alone user, a list of courses available to them is displayed. Next, the user can choose to click on a course link that leads them to the module list for that course. This provides a similar structure to that typically provided by LMSs. Clicking on an entry in the module list will bring the user to the problem list, which is common to both the stand-alone and LMS versions. Samples of these screens are shown in Figure 45.

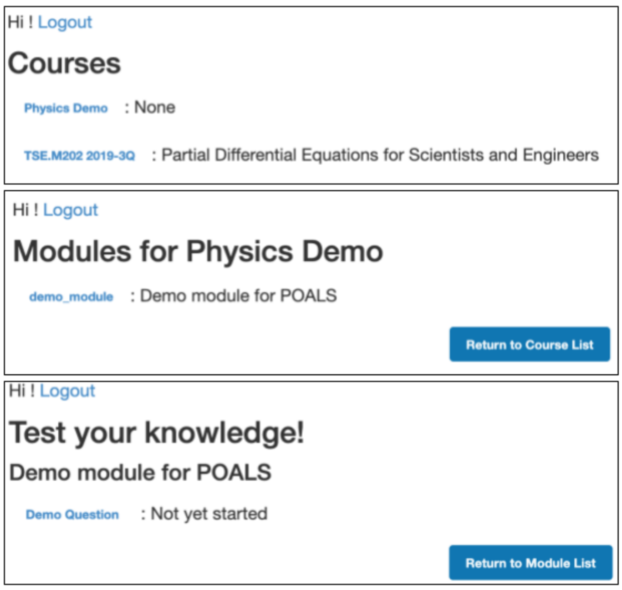


Figure 45: Stand-alone Screens. To give structure similar to that provided by LMSs, the stand-alone version has screens that allow the user to navigate through courses and modules until they reach the question list that is identical to the LMS version.

When the unit with POALS LTI provider is opened on an LMS, or when the user has navigated past the modules list on the stand-alone version, a list of exercises for the specified module is first displayed as seen in Figure 46. Currently, all learners will see the same exercises for each module. Clicking on one of the entries in the list will display the exercise details (problem-solving phase for those in the control group and the preparation phase for those in the treatment group).

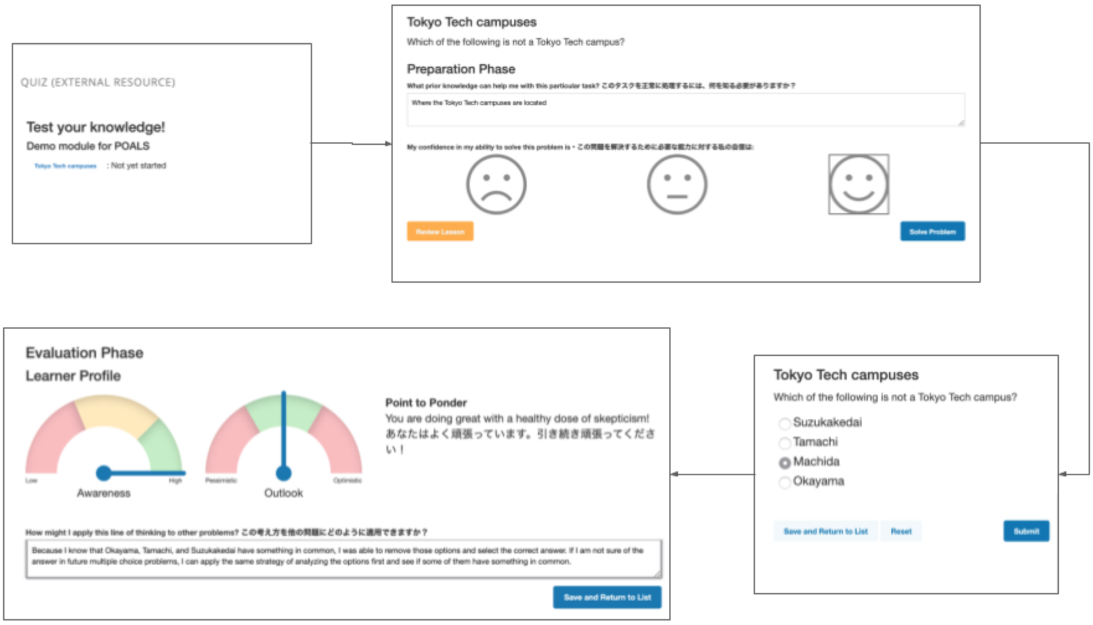


Figure 46: POALS Common Screens. First, a list of exercises for the module is displayed. Clicking on a list item will display the Metacognitive Tutor screens.

A.3 DATABASE STRUCTURE

Each course run corresponds to a **Course** instance (e.g., Science and Engineering Ethics for 2019 and 2020 will be two course instances). Each **Course** will be made up of **Module** and **Outcome** instances. It is possible for a **Module** to have multiple **Outcome** instances and an **Outcome** to be part of several **Module** instances. This relationship between **Module** and **Outcome** is managed through **Outcome Mapping**. Currently though, we have been creating Outcome and Module instances with one-to-one relation.

Each quiz item encountered by the learner corresponds to a **Problem** instance. Currently, checkboxes, multiple-choice, and short response problem types can be set. So far, no short response problem type had been set. For **Problem** instances that are of checkbox or multiple choice types, **Problem Option** instances are created for each possible answer to the problem. Multiple **Hint** instances can also be created for each **Problem** instance. Currently, the **Hint** instances are not yet utilized.

Each user (both learner and teacher) will have a **User** instance. The teacher must be provided with an administrator right to be able to create **Course**, **Module**, **Outcome**, **Outcome Mapping**, **Problem**, **Problem Option**, and **Hint** instances. A **Learner** instance will be provided to the learner **User** for each course the learner is enrolled to. That is, if the learner is enrolled in both Science and Engineering Ethics and Electrical Engineering Literacy courses, the learner will have one User instance and two Learner instances – one for each course.

Before Learner proceeds with using POALS, their consent to join the research is gathered, and relevant questionnaires are administered before and after the experiment. The questionnaire details are explained in Chapter B.

The hierarchy of these instances, which are manually created from the database server's perspective, is detailed in Figure 47.

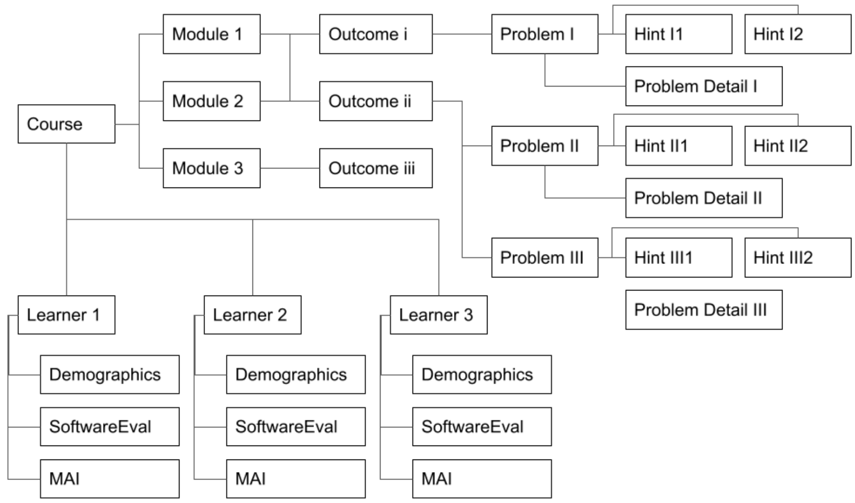


Figure 47: Hierarchy of Manually Inputted Instances. These are the information manually inputted by an administrator account.

After these manual instances are created, a few instances are automatically created by the database server and subsequently updated based on learner interaction. Figure 48 illustrates the groupings of these automatically created instances. Details of these instances are discussed in succeeding paragraphs.

Each **Problem** and **Learner** combination will have associated **Answer** instances which get updated every time the Learner attempts to answer the **Problem**. The **Answer** instance is automatically created when a **Problem** or a **Learner** record is created. An **Answer** instance is created for each allowable attempt set in the **Problem** instance.

Each **Answer** and **Problem Option** combination will have an associated **Answer Detail** instance which also gets updated every time the Learner selects the associated **Problem Option** as an answer. For **Answer** instances whose associated **Problem** instance is not a checkbox or multiple choice types, an **Answer Detail** is created, which is updated with the Learner's input when answering the problem.

When a **Learner** assigned to an experimental condition answers an exercise, the **Knowledge Monitor** and **Prompt Answer** instances associated with the **Answer** instance are created. The **Prompt Answer** instance contains the responses to the metacognitive prompts provided by the **Learner** during the Preparation and Evaluation phases. The **Knowledge Monitor** instance, on the other hand, contains the values relevant to computing for the Learner Profile metrics.

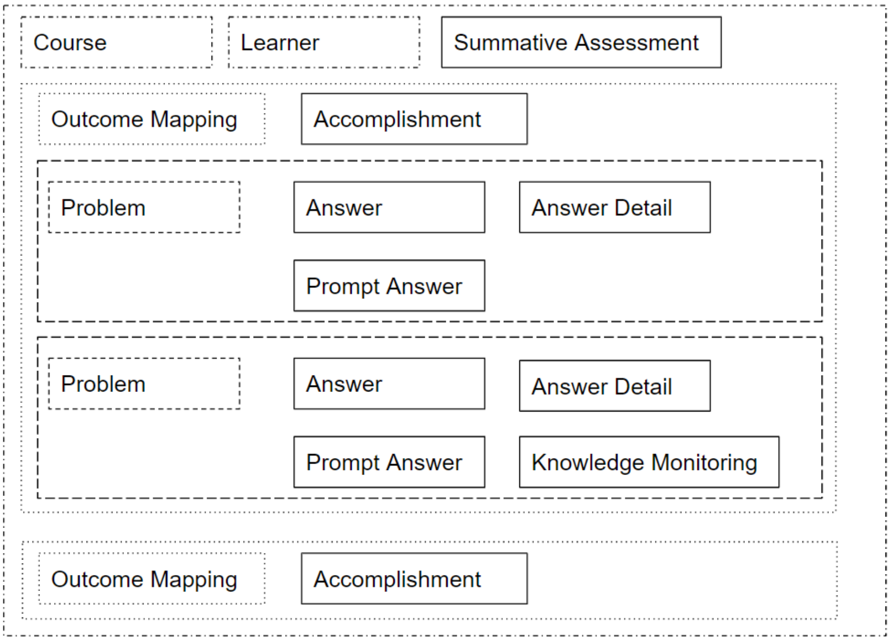


Figure 48: Grouping of Automatically Generated Instances. Instances in solid borders are automatically instantiated based on the manually inputted instances (dashed or dotted borders).

An **Accomplishment** instance will be created for each **Outcome Mapping** and **Learner** combination. The **Accomplishment** instance will be used for tracking the **Learner**'s progress for each **Outcome Mapping**. This will then be used by the Adaptive Engine to decide whether the **Learner** needs more exercises or not.

Every **Learner** and **Course** combination will have a corresponding **Summative Assessment** instance. This will be used to store the diagnostic and summative test results to help gauge the effectiveness of the Adaptive Engine.

A.4 INSTALLATION AND USAGE

The POALS source code can be downloaded from https://gitlab.com/maykristine/poals. Listing 4 shows the file directory of the source code. After downloading, the software can be installed natively, as a Python virtual environment, or as a Docker environment.

Listing 4: File Directory

```
poals
| app
| | Dockerfile
| | entrypoint.sh
| | main
| | | admin.py
| | | apps.py
| | | constant.py
| | | forms.py
| | | middleware.py
| | | migrations
```

```
| | | models.py
| | | static
| | | | css
| | | | | style.css
| | | | js
| | | | | assessment.js
| | | | | problem.js
| | | | | research.js
| | | templates
| | | | common
| | | | | about.html
| | | | | base.html
| | | | | courses.html
| | | | | error.html
| | | | | home.html
| | | | | modules.html
| | | | main
| | | | | accomplishment.html
| | | | | answer_detail.html
| | | | | answer_detail_lti.html
| | | | | answer_detail_sa.html
| | | | | answer_list.html
| | | | | answer_list_lti.html
| | | | | answer_list_sa.html
| | | | | assessment.html
| | | | | assessment_lti.html
| | | | | assessment_sa.html
| | | | | eval.html
| | | | | prep.html
| | | | | prob.html
| | | | | prob_detail.html
| | | | | prob_overview.html
| | | | | submit_error.html
| | | | | submit_info.html
| | | | registration
| | | | | login.html
| | | | research
| | | | | consent.html
| | | | | demographics.html
| | | | | description.html
| | | | | error_consent.html
| | | | | games.html
| | | | | mai.html
| | | | | research.html
| | | | | research_end.html
| | | | | research_info.html
| | | | | retract.html
| | | | | survey.html
| | | | | tam.html
| | | | | teaching_eval.html
| | | tests.py
| | | urls.py
```

```
| | | views.py
| | | __init__.py
| | manage.py
| | poals
| | | settings.py
| | | urls.py
| | | wsgi.py
| | | __init__.py
| | requirements.txt
| db_backup.bat
| docker-compose.yml
| letsencrypt
| | certbot.bat
| | src
| | | docker-compose.yml
| | | letsencrypt-site
| | | | index.html
| | | nginx.conf
| | test_certbot.bat
| nginx
| | Dockerfile
| | nginx.conf
```

Currently, the LTI settings are in the `PYLTI_CONFIG` argument in `poals` > `poals` > `settings.py`. The `PYLTI_CONFIG` information should match the values set in the LMS (manner of setting in the LMS varies).

The Course instance is created by a POALS administrator user. To be able to connect with edX in particular, the context value of the created instance should be set to the one provided by edX, which can be obtained from the uniform resource locator (URL) (sample: course-v1:TokyoTechX+Phys101x+1T2016). When creating a unit on an LMS, the module is provided as a custom parameter for the POALS LTI provider.

For the stand-alone version, the Learner instance can be created by the POALS administrator user and assigned a User instance associated with it. For the LTI provider version, when accessed via the LTI consumer for the first time, a Learner instance is created using its Anonymous ID, which for the case of edX is request.user.first_name (with slight differences from the Anonymous ID information we can download from the edX site).

As can be seen from the file directory, the questionnaires used for this research are also coded into POALS.

B

ETHICS REVIEW AND QUESTIONNAIRES

> *"Well, now that we have seen each other," said the Unicorn, "if you'll believe in me, I'll believe in you. Is that a bargain?"*
>
> –Lewis Caroll, *Through the Looking Glass*

This research was approved by Tokyo Tech's ethical research review committee in June 2019. Ethics modification reports were submitted in August 2019, December 2019, and April 2020 to include new participant pools. Another modification report was submitted in May 2021 to add the POALS Analytics Dashboard user study.

B.1 ETHICS STATEMENT

As written in the JSPS Kakenhi proposal (funded with Grant Number JP20H01719)

While the merits of the research will be explained to the subjects at the start of each cycle, they may at some point decide that they may not be able to sustain the effort required to continue with the experiment. The subjects will be free to leave the experiment anytime they wish or refrain from answering questions if they feel mentally burden by it. To ensure research subjects' right to do so, they will be informed of their rights prior to their participation in this study.

Precautions will also be made to ensure that the subjects' personally identifiable information is not leaked. Anonymized IDs will be used to store learner performance data. Feedback from course developers (teachers, teaching assistants, and other persons involved in course development) will be gathered through anonymized surveys.

Contact information of the researcher including cellphone and e-mail address will be provided at the beginning of each cycle. Should a respondent come up with a question or concern, they are free to send their message to the researcher, and the researcher will get back to them via the channel they state is convenient for them. Should there be a request from a researcher´s subject to remove a data record, such request will be granted.

All the data will be saved in a password-protected database installed in password-protected machines. Data analysis tools such as R will be used for this study. As such, regular backups to secure databases installed in secure machines will be conducted. Data will be kept for the period of ten years after the research is concluded. Should the research team be requested to publish the research data, a new set of Anonymized IDs will be created to make the data not linkable to the participants.

All data will not be made available to persons outside of the research team (including the teachers) unless in aggregated form, in which the learners will no longer be individually identifiable. The learner's personal information, aside from those gathered in surveys, will remain in the edX platform and will not be accessible to the researchers. Only the learners will be able to see the entirety of their personal data, and they will not be able to access other learner's data.

B.2 RESEARCH DESCRIPTION

(Description to be given to those cooperating with the research regarding Personalized Online Adaptive Learning System) When not specified, the same English text as presented to the students will be presented to the educators. No Japanese translation will be provided to the educators.

About Personalized Online Adaptive Learning System Research

【Personalized Online Adaptive Learning System】の研究について

This course will be used as an experiment for the research titled "Personalized Online Adaptive Learning System." We would like to request your consent in participating in this experiment. Please read the research details below and respond to the consent form and applicable surveys. Thank you.

このコースは、「Personalized Online Adaptive Learning System」と題した研究の実験のために行われます。この実験への参加に同意していただければと思います。以下に記述されている研究内容をよく読み、同意書と該当するアンケートに回答してください。よろしくお願い致します。

B.2.1 *Research summary*・研究概要について

Online education is a cost-effective way to democratize access to education; but learners in an online classroom setting are prone to isolation, disengagement, fatigue, and shallow learning. In this research, we will use prompts to help learners develop metacognitive skills that will help them regulate their learning, thus making them more motivated to have a deeper understanding of learned concepts. We would also be using artificial intelligence and diagnostic test results to optimize the learner's path through the exercises to lessen fatigue while still ensuring mastery. Natural language processing and machine learning techniques will be used to obtain course quality feedback that the learners might not be able to communicate through the usual channels, making their experiences matter in future course developments.

オンライン教育は、教育へのアクセスを民主化する費用対効果の高い方法です。しかし、学習者はオンラインクラスの中で、孤立、離脱、疲労、および浅い学習をする傾向にあります。この研究では、プロンプトを使用して、学習者が自身の学習を調節するのに役立つメタ認知能力を身につけるための手助けをし、学習した概念をより深く理解できるようにします。また、人工知能と診断テストの結果を用いて、学習者の学習過程を最適化し、理解度を確保しながら疲労を軽減します。自然言語処理と機械学習技術を使用することによって、従来の方法では取得できない可能性のあるコースの質に関する学習者のフィードバックを取得し、そこで得られた経験を今後のコース開発に生かしていきます。

B.2.2 *Significance and goals of research*・研究の意義と目的について

Massive open online courses (MOOC) had only started gaining popular attention in 2012, hence it is a relatively new education delivery method. At this stage, improving online learning experience especially in MOOC-format is still an active research area. Some attempts include displaying prompts to help learners develop good study habits, change the learning paths to allow learners to sufficiently master content as quickly as possible, or to get feedback from learners by analyzing their posts in discussion forums. The point of this research is to investigate possible improvements for several strategies in improve a learner's performance on an online learning platform.

「Massive open online course (MOOC)」は、2012年に注目を集め始めたばかりで、比較的新しい教育方法です。現段階では、特にMOOC形式でのオンライン学習体験の改善は、依然として研究が盛んな分野です。すでに行われた試みとしては、学習者が良い学習習慣を身に付けるための手助けを行うプロンプトを表示すること、学習経路を変更して学習者ができるだけ早く内容を十分に理解できるようにすること、またはディスカッションフォーラムでの投稿を分析して学習者からフィードバックを得ること、などが挙げられます。この調査の目的は、オンライン学習プラットフォームでの学習者のパフォーマンスを改善するために提案された戦略の改善可能な点を調査することです。

B.2.3 *Research methods*・研究方法について

For students In this experiment, you would be interacting with a system embedded in the edX platform. The system will involve the use of metacognitive prompts and/or different methods of displaying quizzes personalized based on your progress. You would additionally be asked to answer questionnaires to: gather demographic data (gender, language proficiency, etc.), measure your metacognitive awareness, and rate your experience using the system.

この実験の参加者には、edXプラットフォームに組み込まれたシステムと対話してもらいます。このシステムには、メタ認知的プロンプトの使用や、進行状況に基づいて個別化されたクイズを表示するためのさまざまな方法が含まれています。さらに参加者には、人口統計データ（性別、言語能力など）の収集、メタ認知意識の測定、システムによる経験の評価を行うために、アンケートに回答してもらいます。

The experiment, including the giving of consent and the review of the experiment details, will be conducted throughout the entire course in the form of assignments to allow you to complete the activities at your own pace. For Tokyo Institute of Technology students, the experiment is expected to be conducted outside class hours. The materials will be delivered online; hence you should be able to access it anytime and anywhere at your convenience.

この実験（実験参加の同意及び実験内容の確認を含む）は、コース全体を通して課題という形で行われるため、自分のペースで学習を終えることができます。実験は授業時間外に実施される予定です。資料はオンライン上で提供されるため、いつでもどこでもアクセスできます。

The inputs you provide in the course will be stored in a database managed and only accessible to the research team. From the standpoint of protection of personal information, participants will only be identified using anonymized IDs, hence no personal information will be made available to the research team. To further ensure anonymity, the data will be re-anonymized after the class has ended and grades are issued.

このコースで参加者が提供する情報は、管理されているデータベースに保存されます。このデータベースには、研究チームのみがアクセスできるようになっています。個人情報保護の観点から、参加者は匿名化されたIDのみによって識別されるため、研究チームが参加者の個人情報を利用することはできません。匿名性をさらに確保するために、コースが終了し、成績が公開された後に、再度データを匿名化します。

Because the research data is not accessible outside of the research team (which does not include your class instructors), we guarantee that your performance on the metacognitive prompting tasks will not affect your grade or receipt of edX certificates in any manner. You can freely choose to participate or not participate, or hold any opinions about the research whether in agreement or disagreement of the research goals without receiving any negative repercussions. We hope though that participating in this research will help you build metacognitive skills that can contribute to better academic performance.

研究データは研究チーム以外からはアクセスできないため（担当教師は含まれない）、メタ認知的プロンプトの課題における参加者の結果が成績に一切影響しないことを保証します。学生はこの研究への参加・不参加に関して自由に決めることができ、研究目標に賛

成であろうと反対であろうと研究に関して自由に意見を述べることが出来ます。ただし、この研究に参加することで、参加者は授業中の態度や成果の向上に貢献できるメタ認知能力を構築できるようになると我々は考えています。

For educators In this study, you would be asked questions exploring your attitudes, strategies, opinions, and pain points with regards to educational technology trends. You will then be presented a prototype that you will rate based on how you might use it in your own classrooms. The user study, including the giving of consent and the review of the study details, may take around fifteen to thirty minutes to complete. The inputs you provide will be stored in a server managed and only accessible to the research team. From the standpoint of protection of personal information, participants will only be identified using anonymized IDs, hence no personal information will be made available to the research team. Because the research data is not accessible outside of the research team, we guarantee that your responses will not inadvertently affect you any manner. You can freely choose to participate or not participate, or hold any opinions about the research whether in agreement or disagreement of the research goals without receiving any negative repercussions. We hope though that participating in this research will help you gain insights you may use in your classrooms.

B.2.4 *Storage of data and their use in other research*・情報の保管と、他の研究への利用について

The collected anonymized data will be stored and managed after undergoing the second pass of anonymization, making it impossible to identify the sources of the data. The data will only be used for basic research in improving online learning experiences. Data will not be used in any other research.

収集された匿名データは、さらなる匿名化を経て保存および管理されるため、データソースを特定することは不可能です。これらのデータは、オンライン学習体験を改善するための基礎研究にのみ使用されます。データは他の研究では一切使用されません。

B.2.5 *Forecasting results (merits and demerits)*・予測される結果（メリットとデメリット）について

For students You may or may not be requested to answer questionnaires and prompt questions while participating in this research, which would involve a minimal amount of time and effort. At most, you would be required to take three surveys (each not taking more than ten minutes), diagnostic and assessment tests (which would still be required in class even if you choose not to participate in the experiment), and prompt questions for each quiz question. Aside from these, no other demerits are anticipated to come from this experiment. As a matter of fact, we even hope that participating in this experiment will help you improve your grades because the additional question prompts are designed to help you in answering the actual quiz questions. Additionally, we look forward to you developing your metacognitive skills through these question prompts, which can be helpful for you as learners even in other domains. If you are a student of Tokyo Institute of Technology and you believe that you are being adversely impacted by this experiment, please reach out to Tokyo Institute of Technology's Harassment Office (contact details listed below).

Contact details no longer written in this dissertation.

この研究に参加している間、参加者はアンケートや簡単な質問に回答するよう要求される場合とされない場合がありますが、たとえ要求されてもそれほど手間はかかりません。参加者は最大で3つのアンケート（それぞれ10分以内）、診断テストと評価テスト（実験に参加しない場合でもコースで必要）、各クイズに答える前に用意された簡単な質問に回答するよう要求されます。これらとは別に、この実験から他のデメリットが生じるとは考えられていません。実際のところ、追加の簡単な質問は、実際のクイズへの回答を手助けするためにつくられているため、この実験に参加することによって参加者は自身の成績を伸ばすことができると私たちは期待しています。さらに、これらの簡単な質問を通じて、参加者がメタ認知能力を開発すること私たちは期待しています。この能力は、参加者が他の分野・領域で学習するときに役立つはずです。東京工業大学の学生で、自身がこの実験によって悪影響を受けていると思われる場合は、東京工業大学のハラスメント窓口（下記の連絡先）にご連絡ください。

For educators There is no personal and/or social merits and demerits to the individuals that will be involved in the study.

B.2.6 *Cooperation with the research is voluntary and retraction of consent is possible at any time*・研究協力の任意性と撤回の自由について

For students You have the complete freedom to participate or not participate in this research. Furthermore, if you no longer wish to cooperate even after having previously given consent, as soon as a request for retraction is received, the further experiment will be canceled and data whose sole purpose is for research will be destroyed. The retraction form will be introduced during the class orientation and can be accessed while the class is still open. The retraction will not penalize you in any way; in particular, the retraction will not affect your grade in the class. However, since the data is re-anonymized after grades release, the data becomes non-traceably anonymized and thus cannot be destroyed. Retraction of consent after grade release is not possible.

この研究への参加・不参加は完全に自由です。さらに、既に同意を示した後で研究への参加をキャンセルしたくなった場合には、そ

の旨を申し出ていただいた時点で実験は中止され、全てのデータは破棄されます。撤回申出書は、コース初回に行われるオリエンテーションの中で説明され、コース開講中であればいつでもアクセスすることができます。撤回をすることによるペナルティ（コースの成績に影響するなど）は一切ありません。得られたデータは成績が公開された後に再匿名化され、追跡不可能となるため、それらを破棄することはできません。つまり、成績が公開された後の同意の撤回はできません。

For educators You have the complete freedom to participate or not participate in this research. Furthermore, if you no longer wish to cooperate even after having previously given consent, as soon as a request for retraction is received, the further study will be canceled and data whose sole purpose is for research will be destroyed. The retraction form will be introduced at the end of this form. The retraction will not penalize you in any way.

B.2.7 *Expenses*・費用について

The research subjects will bear absolutely no supplementary expenses for the tests and analysis that accompany the research. There is no remuneration for the participants.

研究に伴う測定・解析によって研究対象者が負担する付加的な費用は一切ありません。また、実験への協力に対する謝礼もありません。

B.2.8 *Compensation for adverse health effects*・健康被害の補償について

No adverse health effects are anticipated. Should problems arise, please do not hesitate to contact the designated person for this research (contact details below).

Contact details no longer written in this dissertation.

実験による健康への悪影響は予想されていません。問題が発生した場合は、指定された担当者に連絡してください（下記の連絡先）。

B.2.9 *Protection of personal information*・個人情報の保護について

Because the name of the research subject is anonymized, personal information regarding the research subject can in absolutely no way be leaked outside of the research team's control.

研究対象の名前は匿名化されているため、研究対象に関する個人情報が研究チームの管理外に漏洩することは絶対にありません。

B.2.10 *Publication of the research results*・研究成果の公表について

Research results may be publicized through academic associations in educational and computational fields such as the Japan Society for Educational Technology; committees of specialists; international meetings; and in educational and computational journals. In such cases as well, absolutely no identifiable information specific to participants are released.

研究成果は、日本教育工学会などの教育および計算分野の学会、専門委員会、国際会議、教育及び計算ジャーナルを通じて公表される可能性があります。そのような場合でも、実験参加者を識別できるような情報は絶対に公開されません。

B.3 QUESTIONNAIRES FOR THE METACOGNITIVE TUTOR

B.3.1 *Demographics Questionnaire*

Demographic data was collected before the intervention. Below are the questions used and their corresponding answer options for the demographics questionnaire:

- What is your age? 年齢
  - (Integer input)
- What is your current degree level? 学位
  - Bachelors・学士
  - Masters・修士
  - Doctoral・博士
  - Others・その他
- What is your current year level? 年生
  - (Integer input)
- Which country are you from? 国
  - (Django countries)[218]
- What is your current proficiency level in English? 英語の能力
  - No proficiency・習熟度なし
  - Elementary proficiency・挨拶レベル
  - Limited working proficiency・日常会話レベル
  - Professional working proficiency・ビジネス初級
  - Full professional proficiency・ビジネス上級
  - Native or bilingual proficiency・母国語またはバイリンガルレベル

[218] C. Beaven. (Apr. 2017). *django-countries version 4.4, [Online]. Available: https://github.com/SmileyChris/django-countries/commits/v4.4.*

- What is your current proficiency level in Japanese? 日本語の能力
  - No proficiency・習熟度なし
  - Elementary proficiency・挨拶レベル
  - Limited working proficiency・日常会話レベル
  - Professional working proficiency・ビジネス初級
  - Full professional proficiency・ビジネス上級
  - Native or bilingual proficiency・母国語またはバイリンガルレベル
- Do you have any working experience? 実務経験
  - None なし
  - Less than a year 1年未満
  - One to five years 1年から5年
  - More than five years 5年以上
- What is your gender? 性
  - Male・男
  - Female・女
  - Other・その他

B.3.2 *Metacognitive Measurement*

Metacognitive measurement was done before and after the interventions.

B.3.2.1 *Goal-oriented studying, Active studying, Meaningful and memorable studying, Explain to understand, and Self-monitor (GAMES)*

No changes were made to the GAMES[219] questionnaire. This questionnaire was only administered during the pilot study to determine which offline measurement tool for metacognition will be used.

[219] *M. D. Svinicki,* Learning and motivation in the postsecondary classroom. *Anker Publishing Company, 2004.*

> The GAMES questions are not listed here since the original author holds copyright.

B.3.2.2 *Metacognitive Awareness Inventory (MAI)*

Below are the questions for the updated MAI[220] questionnaire used for this research:

[220] *G. Schraw and R. S. Dennison, "Assessing metacognitive awareness,"* Contemporary Educational Psychology, *vol. 19, no. 4, pp. 460–475, 1994.*

1. I ask myself periodically if I am meeting my goals・目標を達成しているかどうかを定期的に自分に聞いています。
2. I consider several alternatives to a problem before I answer・私は答える前に、問題のいくつかの他の候補を検討します。

3. I try to use strategies that have worked in the past・私は過去にうまくいった方法を適用します。
4. I pace myself while learning in order to have enough time・私は十分な時間があるように、学習しながら自分のペースを調整します。
5. I understand my intellectual strengths and weaknesses・私は自分の知的な長所と短所を理解しています。
6. I think about what I really need to learn before I begin a task・私は課題を始める前に、自分が本当は何を学ぶ必要があるのかを考えます。
7. I know how well I did once I finish a test・私は試験を終えた後、自分がどれだけうまくできたかがだいたい分かります。
8. I set specific goals before I begin a task・私は課題を始める前に具体的な目標を設定します。
9. I slow down when I encounter important information・私は重要な情報にを得ると、作業のペースを下げます。
10. I know what kind of information is most important to learn・私はどのような情報を学ぶことが最も重要かを知っています。
11. I ask myself if I have considered all options when solving a problem・私は問題を解決するときにすべての可能性がある回答を検討したかどうかを自問します。
12. I am good at organizing information・私は情報を整理するのが上手です。
13. I consciously focus my attention on important information・私は意識的に重要な情報に注意を払っています。
14. I have a specific purpose for each strategy I use・私は使用する戦略ごとに目的を明らかにしています。
15. I learn best when I know something about the topic・私はトピックについて、何かを知っているとき、最もよく勉強します。
16. I know what the teacher expects me to learn・私は先生が何を学ぶことを期待しているかを知っています。
17. I am good at remembering information・私は情報を覚えるのが得意です。
18. I use different learning strategies depending on the situation・私は状況に応じて異なる学習方法を適用します。

19. I ask myself if there was an easier way to do things after I finish a task・私は課題を完了した後、より簡単な方法がありはしないかと自問します。
20. I have control over how well I learn・私は学び方をうまく管理できます。
21. I periodically review to help me understand important relationships・私は重要な関係を理解できるように定期的に見直しをします。
22. I ask myself questions about the material before I begin・私は学習する前に、使用する教材に疑問を持ちます。
23. I think of several ways to solve a problem and choose the best one・私は問題を解法をいくつか考え、その中から最良の方法を選択します。
24. I summarize what I've learned after I finish・私は全て終わった後に、何を学んだのかをまとめます。
25. I ask others for help when I don't understand something・私はわからないことがあったとき、他の人に聞きます。
26. I can motivate myself to learn when I need to・私は必要なときには、自ら学ぶようにすることができます。
27. I am aware of what strategies I use when I study・私は勉強するときに、どのようにすればよいのかわかります。
28. I find myself analyzing the usefulness of strategies while I study・気が付いたら私は勉強方法の有用性を分析しています。
29. I use my intellectual strengths to compensate for my weaknesses・私は自分の弱点を補うため頭を使います。
30. I focus on the meaning and significance of new information・私は新しい情報の意味に注目します。
31. I create my own examples to make information more meaningful・私は情報をよりわかりやすくするために、独自の例を考えることがあります。
32. I am a good judge of how well I understand something・私はどれだけ理解しているかをだいたい判断できます。
33. I find myself using helpful learning strategies automatically・私は役に立つ学習方法をいつのまにか見つけることがあります。
34. I find myself pausing regularly to check my comprehension・私は理解度をチェックするために、定期的に復習しています。

35. I know when each strategy I use will be most effective・私はどの勉強方法が効果的か知っています。
36. I ask myself how well I accomplish my goals once I'm finished・私は何か物事を終えたときに、目標をどの程度達成できたか振り返ります。
37. I draw pictures or diagrams to help me understand while learning・私は勉強している間、絵や図を描いて内容をより理解できるようにしています。
38. I ask myself if I have considered all options after I solve a problem・私は問題を解いたあと、そのほかの解答方法を考えます。
39. I try to translate new information into my own words・私は新しい情報を自分の言葉で解釈します。
40. I change strategies when I fail to understand・私は理解できなかったとき、勉強方法を変えます。
41. I use the organizational structure of the text to help me learn・私は文章を改善して、内容をより理解しやすくしています。
42. I read instructions carefully before I begin a task・私は課題に着手する前に、説明書を注意深く読みます。
43. I ask myself if what I'm reading is related to what I already know・私は読んでいることが、すでに知っていることとの関連があるかを考えます。
44. I reevaluate my assumptions when I get confused・私は混乱してしまったとき、仮定がただしかったかどうか考え直します。
45. I organize my time to best accomplish my goals・私は自分の目標を達成するために時間の使い方を調整します。
46. I learn more when I am interested in the topic・私はトピックにに興味を持ったとき、より学ぼうとします。
47. I try to break studying down into smaller steps・私は学習全体ををより小さなステップに分けて行います。
48. I focus on overall meaning rather than specifics・私は個々の意味づけよりも全体的な意味を重視します。
49. I ask myself questions about how well I am doing while I am learning something new・私は何か新しいことを学んでいる間、自分がどれだけ勉強しているか振り返っています。
50. I ask myself if I learned as much as I could have once I finish a task・私は課題を終えた時、できる限り多くのことを学べたかどうか振り返っています。

51. I stop and go back over new information that is not clear・私は内容があまり明確でない時、もう一度見直します。

52. I stop and reread when I get confused・私は困ったとき、もう一度読み直します。

53. Are there study behaviors, not listed here, that you engage in? If so, specify them here.
    - (Open response)

Options for the close response questions include:

- Yes・はい
- No・いいえ

Changes from the original include:

- Having binary choice instead of selecting from a range of 0 to 100.
- Adding Japanese translation.
- Adding an open response item.

B.3.3 *Modified Technology Acceptance Model (TAM) for the Metacognitive Tutor*

Software evaluation was conducted after the intervention. Below are the questions for the updated TAM[221] questionnaire used for this research:

[221] *F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," PhD thesis, Massachusetts Institute of Technology, 1985.*

1. I think that I would like to see this software more frequently・このソフトウェアをもっと頻繁に見たいと思います。

2. Using the software improved my performance・ソフトウェアを使用することによって、パフォーマンスが向上しました。

3. Using the software increased my productivity in my studies・ソフトウェアを使用することによって、研究の生産性が向上しました。

4. Using the software enhanced my effectiveness in studying・ソフトウェアを使用することによって、勉強の効率が向上しました。

5. Using the system made it easier to do my schoolwork・このシステムを使用することによって、学校の勉強がはかどるようになりました。

6. I find the software useful・このソフトウェアは有用だと思います。

7. Learning to operate the software was easy for me・ソフトウェアの操作を学ぶことは簡単でした。
8. My interaction with the software was clear and understandable・ソフトウェアとのやり取りは明確で理解しやすいものでした。
9. I find the software easy to use・このソフトウェアは使いやすいと思います。
10. What do you like best about this software?
    - (Open response)
11. What do you like least about this software?
    - (Open response)

The original TAM is a Likert scale with seven options. This was reduced to just five and all labeled to lessen ambiguity. Options for the close response questions include:

1. Strongly Agree・強く同意します
2. Agree・同意します
3. Undecided・どちらでもない
4. Disagree・反対します
5. Strongly Disagree・強く反対します

Other changes from the original include:

- Adding Japanese translation.
- Adding an open response item.
- Rewording to fit school context.
- Removing unrelated items.
  - Accomplishing a task quickly.
    - (Current software actually introduces new activity, so the task cannot be completed faster.)
  - Making the software accomplish a desired task easily.
    - (Current software is single-function.)

B.4 USER STUDY FOR THE ANALYTICS DASHBOARD

Educators were invited for a user study of the POALS Analytics Dashboard. The user study was conducted via Google Forms.

B.4.1 *Written Interview*

Before introducing the educators to the POALS Analytics Dashboard, they were first given a set of questions to give them a context on the user study that they are undergoing.

- Tell me about your regular interaction with your students
- Describe to me your experience teaching in an online classroom, if any
- How often do you engage with your students?
- How do you engage with your students?
- How would you gauge how well your students are doing without looking at assessments? What are the things or information you need to be able to do so? Walk me through the process.
- Can you recall a situation when you are unclear whether your students are learning something or not? What did you do about it?
- Are you familiar with learning analytics dashboards?
  - Yes
  - No
  - Maybe
- What trends on learning outcomes assessment and student engagement are you familiar with?
- What do you think about the trends you have mentioned?
- What do you think your colleagues think about the trends you have mentioned?
- What do you think your students will think about the trends that you have mentioned?
- What are some problems in this space (actively looking out for student learning) that you think must be addressed?
- What is the most frustrating thing about understanding and adjusting to student learning?
- How will you solve these issues?

B.4.2 *Introducing POALS*

After the written interview, the educators were first introduced to the POALS Metacognitive Tutor then the POALS Analytics Dashboard. Below are the text used for the POALS introduction:

The Personalized Online Adaptive Learning System. Perhaps this is not a surprise for you, but several research has shown that online learning can be challenging. To help students and teachers succeed in online learning, we developed the Personalized Online Adaptive Learning System or POALS. Here, I will be introducing you to two components of POALS: the Metacognitive Tutor for the students and the Analytics Dashboard for the teachers.

Metacognitive Tutor. The reason why most students find it hard to learn online is that they do not have sufficient skills to learn independently. The Metacognitive Tutor develops the students' independent learning skills by practicing them to always ask themselves how well they understand a lesson before rushing into their activities. The Metacognitive Tutor works by adding preparation and evaluation phases for problems they need to answer.

In the Preparation Phase, the students are shown the question that they need to answer. Then they are ask what they need in order to successfully answer the problem and rate themselves how prepared are they. At this point, they have the option to review the lesson first if they realize they are not yet prepared, or they can proceed to answering the problem.

After the Preparation Phase, the students can proceed to answering the problem. After submitting their answers, they are moved to the Evaluation Phase.

In the Evaluation Phase, the students are shown how aware are they of their knowledge levels and whether they are pessimistic or optimistic about their views on their ability. They are also given pointers on how to grow as independent learners based on their awareness and outlook. They are finally asked to reflect on their problem solving experience.

Analytics Dashboard. For the teachers, the problem in online learning environments is that they cannot directly observe their students, so it is a little harder to gauge whether the students are actually learning or not. Even if the teacher actively asks the students how are they doing, not everyone would be willing to respond. Worse, those who choose to keep quiet are sometimes the ones most needing attention.

We created the Analytics Dashboard to use the inputs of Metacognitive Tutor and applied computational techniques to create a webpage that will allow the teacher to see how are the students feeling, what are the topics the students are most familiar with, and the relationship of the course modules and the student scores. The Analytics

Dashboard shows data for the entire class (not per module or activity).

In the sentiment analysis portion of the Analytics Dashboard, you will see a graph similar to below. In the bar chart, you will see the overall feelings of the class based on their inputs in the Metacognitive Tutor (higher score means generally happier students). The donut gauge shows the percentage of responses that are negative, neutral, or positive.

In the word cloud portion of the Analytics Dashboard, you will see the topics most frequently mentioned by the students in the Metacognitive Tutor. Why is this important? By knowing the topics the students discussed, you will have an idea which topics are sticking with the students. Usually, you will know from the students' grades which topics are they understanding well because their grades will be good. But what if their grades are not that good? Seeing a topic appear on the word cloud related to a module were the average grade is low can indicate that maybe the module leads several students to misunderstanding. This will then allow you to make adjustments to the lesson to clarify misunderstandings.

In the similarity network portion of the Analytics Dashboard, you will see a point for each of the module you have. When you place the mouse over a module, you will see the average scores of your students in the module. Modules related to each other are connected by a line, and the degree of relationship of the modules is shown by the thickness of the line (the thicker the line is, the more related the modules are). The relationships are computed based on the words you used in the lessons for the said modules (e.g., video transcripts or text in presentation slides). Ever had the experience where your students had low grades in Module 3, only because they already forgot what was taught in Module 1 which is necessary for understanding Module 3? This visualization can help you diagnose signs that your students are starting to forget important information from previous modules.

B.4.3 *Modified Technology Acceptance Model (TAM) for the Analytics Dashboard*

After answering the interview questions, the educators are shown the POALS Analytics Dashboard screenshots with short explanations. Then, the educators are asked to evaluate the POALS Analytics Dashboard using the modified TAM. Just like for the Metacognitive Tutor, the TAM was modified for the Analytics Dashboard to a Likert scale with five options. Options for the close response questions include:

1. Strongly Agree
2. Agree

3. Neutral
4. Disagree
5. Strongly Disagree

All questions in the original TAM are reworded and used for the Learning Analytics Dashboard.

- The tool can help me assess my students' learning climate more quickly
- The tool can improve my ability to respond to my students' needs
- The tool can increase my work productivity
- The tool can enhance my teaching effectiveness
- The tool can make it easier to do my work
- I think the tool can be useful in my work
- I think I can learn how to maximize the tool easily
- I think the tool can give me the information I need without much effort
- The tool and the information it provides is understandable
- The tool gives me flexibility to decide what is best for my students
- I can imagine how will I be interacting with the tool on a regular basis
- I think the tool will be easy to use

B.4.4 *Open Responses*

In the end, the educators are encouraged to give more insights on the POALS Analytics Dashboard.

- What do you think are strengths of the POALS Analytics Dashboard?
- What do you think are shortcomings of the POALS Analytics Dashboard?
- Do you have any further suggestions for the POALS Analytics Dashboard?
- Do you have other information you want to share?

BIBLIOGRAPHY

B. Acar Sesen and E. Ince, "Internet as a source of misconception," *Turkish Online Journal of Educational Technology-TOJET*, vol. 9, no. 4, pp. 94–100, 2010.

Adaptive Instructional Systems (C/LT/AIS) P2247 Working Group. (2021). Adaptive instructional systems (c/lt/ais) p2247.1, [Online]. Available: https://sagroups.ieee.org/2247-1/.

K. Agustianto, A. E. Permanasari, S. S. Kusumawardani, and I. Hidayah, "Design adaptive learning system using metacognitive strategy path for learning in classroom and intelligent tutoring systems," in *AIP Conference Proceedings*, AIP Publishing LLC, vol. 1755, 2016, p. 070 012.

V. Aleven and K. R. Koedinger, "Limitations of student control: Do students know when they need help?" In *International Conference on Intelligent Tutoring Systems*, Springer, 2000, pp. 292–303.

——, "An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor," *Cognitive Science*, vol. 26, no. 2, pp. 147–179, 2002.

L. W. Anderson, B. S. Bloom, *et al.*, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

R. Azevedo, "Reflections on the field of metacognition: Issues, challenges, and opportunities," *Metacognition and Learning*, Jun. 2020.

R. S. d Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing," in *International Conference on Intelligent Tutoring Systems*, Springer, 2008, pp. 406–415.

R. S. Baker, S. M. Gowda, A. T. Corbett, and J. Ocumpaugh, "Towards automatically detecting whether student learning is shallow," in *International Conference on Intelligent Tutoring Systems*, Springer, 2012, pp. 444–453.

S. Baškarada and A. Koronios, "A philosophical discussion of qualitative, quantitative, and mixed methods research in social science," *Qualitative Research Journal*, 2018.

L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

C. Beaven. (Apr. 2017). django-countries version 4.4, [Online]. Available: https://github.com/SmileyChris/django-countries/commits/v4.4.

K. Berthold, M. Nückles, and A. Renkl, “Do learning protocols support learning strategies and outcomes? the role of cognitive and metacognitive prompts,” *Learning and Instruction*, vol. 17, no. 5, pp. 564–577, 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

O. Blomberg, “Conceptions of cognition for cognitive engineering,” *The International Journal of Aviation Psychology*, vol. 21, no. 1, pp. 85–104, 2011.

W. R. Boot, D. J. Simons, C. Stothart, and C. Stutts, “The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects,” *Perspectives on Psychological Science*, vol. 8, no. 4, pp. 445–454, 2013.

V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

A. L. Brown, “Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings,” *The Journal of the Learning Sciences*, vol. 2, no. 2, pp. 141–178, 1992.

M. Brown, “Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data,” *Teaching in Higher Education*, vol. 25, no. 4, pp. 384–400, 2020.

M. K. J. Carlon, A. D.D. C. Asa, N. Keerativoranan, T. Nagahama, and J. S. Cross, “Topic modeling in MOOCs: What was to be discussed, what the instructor discussed, and what the learners discussed,” in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Submitted.

M. K. J. Carlon and J. S. Cross, “Challenges of developing a metacognitive tutor on Open edX,” in *Fourteenth Asia-Oceania Top University League on Engineering Student Conference*, 2019, Details in Section A.1.

——, “A review of quantitative offline measurement tools for computer-based metacognitive tutoring effectiveness assessment,” in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Dec. 2020, pp. 258–264. DOI: `10.1109/TALE48869.2020.9368470`, Details in Sections 1.2, 2.1.

——, “Developing learner metacognitive skills on an online environment,” in *The 2020 Annual Spring Conference of Japan Society for Educational Technology*, 2020, Details in Section 2.3.1.

——, “Learning analytics dashboard prototype for implicit feedback from metacognitive prompt responses,” in *29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education*, 2021, Details in Section 4.8.

——, "Open response prompts in an online metacognitive tutor," in *The 2021 Annual Spring Conference of Japan Society for Educational Technology*, 2021, Details in Section 2.3.2.

——, "POALS Analytics Engine: A student affect dashboard," in *The Eighth UK Japan Engineering Education League Workshop*, 2021, Details in Chapter 4.

——, "Development of open-response prompt-based metacognitive tutor for online classrooms," in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Submitted, Details in Sections 2.2, 2.3.3.

——, "Knowledge tracing for adaptive learning in a metacognitive tutor," *Open Education Studies*, Submitted, Details in Chapter 3.

M. K. J. Carlon, M. R. Gaddem, C. A. Hernández Reyes, T. Nagahama, and J. S. Cross, "Investigating mechanical engineering learners' satisfaction with a revised monozukuri MOOC," in *European MOOCs Stakeholder Summit 2021*, 2021.

M. K. J. Carlon, N. Keerativoranan, and J. S. Cross, "Content type distribution and readability of MOOCs," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 401–404.

M. K. J. Carlon, C. Seng, and J. S. Cross, "Countering negative Matthew effect in undergraduate research with metacognition and digital learning," in *The 2020 Annual Fall Conference of Japan Society for Educational Technology*, 2020, Mentioned in Chapter 5.

E. Chang and S. M. Lee, "Mediating effect of goal adjustment on the relationship between socially prescribed perfectionism and academic burnout," *Psychology in the Schools*, vol. 57, no. 2, pp. 284–295, 2020.

S.-H. Chang, M.-L. Chen, Y.-K. Kuo, and Y.-C. Shen, "A simulation-based LED design project in photonics instruction based on industry–university collaboration," *IEEE Transactions on Education*, vol. 54, no. 4, pp. 582–589, 2011.

K. S. Chew, S. J. Durning, and J. J. Van Merriënboer, "Teaching metacognition in clinical decision-making using a novel mnemonic checklist: An exploratory study," *Singapore Medical Journal*, vol. 57, no. 12, p. 694, 2016.

T. Chothia, S. Holdcroft, A.-I. Radu, and R. J. Thomas, "Jail, hero or drug lord? Turning a cyber security course into an 11 week choose your own adventure story," in *2017 {USENIX} Workshop on Advances in Security Education ({ASE} 17)*, 2017.

L. Cid, A. Pires, C. Borrego, P. Duarte-Mendes, D. S. Teixeira, J. M. Moutão, and D. Monteiro, "Motivational determinants of physical education grades and the intention to practice sport in the future," *PLoS One*, vol. 14, no. 5, e0217218, 2019.

W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

R. Cobos, S. Gil, A. Lareo, and F. A. Vargas, "Open-DLAs: An open dashboard for learning analytics," in *Proceedings of the third (2016) ACM conference on learning@ scale*, 2016, pp. 265–268.

C. L. Colbeck, A. F. Cabrera, and P. T. Terenzini, "Learning professional confidence: Linking teaching practices, students' self-perceptions, and gender," *The Review of Higher Education*, vol. 24, no. 2, pp. 173–191, 2001.

Columbia Center for Teaching and Learning. (Feb. 2018). django-lti-provider version 0.3.3, [Online]. Available: https://github.com/ccnmtl/django-lti-provider/commits/0.3.3.

P. D. Converse, E. Steinhauser, and J. Pathak, "Individual differences in reactions to goal-performance discrepancies over time," *Personality and Individual Differences*, vol. 48, no. 2, pp. 138–143, 2010.

A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1994.

L. F. Cranor, "A framework for reasoning about the human in the loop," in *Proceedings of the 1st Conference on Usability, Psychology, and Security*, Advanced Computing Systems Professional and Technical Association, 2008, pp. 1–15.

J. Crawford, K. Butler-Henderson, J. Rudolph, B. Malkawi, M. Glowatz, R. Burton, P. Magni, and S. Lam, "COVID-19: 20 countries' higher education intra-period digital pedagogy responses," *Journal of Applied Learning & Teaching*, vol. 3, no. 1, pp. 1–20, 2020.

J. S. Cross, N. Keerativoranan, M. K. J. Carlon, Y. H. Tan, Z. Rakhimberdina, and H. Mori, "Improving MOOC quality using learning analytics and tools," in *2019 IEEE Learning With MOOCS (LWMOOCS)*, IEEE, 2019, pp. 174–179.

P. Cunningham, H. Matusovich, and S. Blackowski, "Teaching metacognition: Helping students own and improve their learning," in *Workshop presented at the American Society for Engineering Education Annual Conference and Exposition*, 2018.

Daigo. (Dec. 2020). bert-base-japanese-sentiment, [Online]. Available: https://huggingface.co/daigo/bert-base-japanese-sentiment.

M. T. Damgaard and H. S. Nielsen, "Nudging in education," *Economics of Education Review*, vol. 64, pp. 313–342, 2018.

F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," PhD thesis, Massachusetts Institute of Technology, 1985.

H. Delić and S. Bećirović, "Socratic method as an approach to teaching," *European Researcher. Series A*, no. 10, pp. 511–517, 2016.

A. Demetriou, A. Efklides, M. Platsidou, and R. L. Campbell, "The architecture and dynamics of developing mind: Experiential structuralism as a frame for unifying cognitive developmental theories," *Monographs of the Society for Research in Child Development*, pp. i–202, 1993.

C. Demetriou, B. U. Ozer, and C. A. Essau, "Self-report questionnaires," *The Encyclopedia of Clinical Psychology*, pp. 1–6, 2014.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

B. Deonovic, M. Yudelson, M. Bolsinova, M. Attali, and G. Maris, "Learning meets assessment," *Behaviormetrika*, vol. 45, no. 2, pp. 457–474, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

D. H. Dolmans, S. M. Loyens, H. Marcq, and D. Gijbels, "Deep and surface learning in problem-based learning: A review of the literature," *Advances in Health Sciences Education*, vol. 21, no. 5, pp. 1087–1112, 2016.

W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *Proceedings of the 20th International Conference on World Wide Web*, 2011, pp. 577–586.

S. Doroudi and E. Brunskill, "The misidentified identifiability problem of Bayesian knowledge tracing," *International Educational Data Mining Society*, 2017.

A. L. Duckworth. (Apr. 2013). Grit: The power of passion and perseverance. T. Talks, Ed., [Online]. Available: https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance.

H. Ebbinghaus, "Memory: A contribution to experimental psychology," *Annals of Neurosciences*, vol. 20, no. 4, p. 155, 2013.

EdX Inc. (2021). Open edX XBlock tutorial, [Online]. Available: https://edx.readthedocs.io/projects/xblock-tutorial/en/latest/.

R. Elmore, *Modes of learning framework*, edX, Ed., 2020. [Online]. Available: https://www.edx.org/course/leaders-of-learning.

E. Farrow and J. D. Moore, "Beetle-Grow: An effective intelligent tutoring system to support conceptual change," in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, 2016, pp. 331–332.

D. Feistauer and T. Richter, "How reliable are students' evaluations of teaching quality? A variance components approach," *Assessment & Evaluation in Higher Education*, vol. 42, no. 8, pp. 1263–1279, 2017.

P. Ferguson, "Student perceptions of quality feedback in teacher education," *Assessment & Evaluation in Higher Education*, vol. 36, no. 1, pp. 51–62, 2011.

B. A. Fernie, U. Y. Kopar, P. L. Fisher, and M. M. Spada, "Further development and testing of the metacognitive model of procrastination: Self-reported academic performance," *Journal of Affective Disorders*, vol. 240, pp. 1–5, 2018.

J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," *American Psychologist*, vol. 34, no. 10, p. 906, 1979.

R. Fogarty, *The mindful school: how to teach for metacognitive reflection*. ERIC, 1994.

L. D. Frazier, B. L. Schwartz, and J. Metcalfe, "The MAPS model of self-regulation: Integrating metacognition, agency, and possible selves," *Metacognition and Learning*, pp. 1–22, 2021.

C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change*, vol. 114, pp. 254–280, 2017.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 329–338.

S. Gallagher and J. Palmer, "The pandemic pushed universities online. The change was long overdue," *Harvard Business Review*, 2020.

C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in *International Conference on Intelligent Tutoring Systems*, Springer, 2004, pp. 668–677.

E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning," *arXiv preprint arXiv:2011.05231*, 2020.

F. Gutiérrez, K. Seipp, X. Ochoa, K. Chiluiza, T. De Laet, and K. Verbert, "LADA: A learning analytics dashboard for academic advising," *Computers in Human Behavior*, vol. 107, p. 105 826, 2020.

D. Halpern, S. Tubridy, H. Y. Wang, C. Gasser, P. O. Popp, L. Davachi, and T. M. Gureckis, "Knowledge tracing using the brain," *International Educational Data Mining Society*, 2018.

D. F. Halpern, *Thought and knowledge: An introduction to critical thinking*. Psychology Press, 2013.

M. Hariri-Akbari, B. Shokrvash, F. Mahmoodi, F. Jahanjoo-Aminabad, B. Yousefi, and F. Azabdaftari, “Conversion of extrinsic into intrinsic motivation and computer based testing (CBT),” *BMC Medical Education*, vol. 18, no. 1, pp. 1–8, 2018.

N. T. Heffernan and C. L. Heffernan, “The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching,” *International Journal of Artificial Intelligence in Education*, vol. 24, no. 4, pp. 470–497, 2014.

Highcharts. (2021). Interactive JavaScript charts for your webpage, [Online]. Available: https://www.highcharts.com/.

K. Higuchi, “KH Coder 3 reference manual,” *Kioto (Japan): Ritsumeikan University*, 2016.

E. S.-c. Ho *et al.*, “Characteristics of East Asian learners: What we learned from PISA,” *Educational Research Journal*, vol. 24, no. 2, p. 327, 2009.

C. Hodges, S. Moore, B. Lockee, T. Trust, A. Bond, *et al.*, “The difference between emergency remote teaching and online learning,” *Educause Review*, vol. 27, pp. 1–12, 2020.

X. Hong, P. Wong, D. Liu, S.-U. Guan, K. L. Man, and X. Huang, “Lifelong machine learning: Outlook and direction,” in *Proceedings of the 2nd International Conference on Big Data Research*, 2018, pp. 76–79.

L.-L. Hsu and S.-I. Hsieh, “Factors affecting metacognition of undergraduate nursing students in a blended learning environment,” *International Journal of Nursing Practice*, vol. 20, no. 3, pp. 233–241, 2014.

E. Huang, H. Valdiviejas, and N. Bosch, “I’m sure! Automatic detection of metacognition in online course discussion forums,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7.

IMS Global Learning Consortium. (Apr. 2019). Learning tools interoperability core specification 1.3, [Online]. Available: http://www.imsglobal.org/spec/lti/v1p3/.

S. Jagadeesan and J. Subbiah, “Real-time personalization and recommendation in Adaptive Learning Management System,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2020.

I. Jahnke, Y.-M. Lee, M. Pham, H. He, and L. Austin, “Unpacking the inherent design principles of mobile microlearning,” *Technology, Knowledge and Learning*, vol. 25, no. 3, pp. 585–619, 2020.

H. Jang, “Supporting students’ motivation, engagement, and learning during an uninteresting activity.,” *Journal of Educational Psychology*, vol. 100, no. 4, p. 798, 2008.

Japan Cabinet Office. (nd). Society 5.0, [Online]. Available: https://www8.cao.go.jp/cstp/english/society5_0/index.html.

D. A. Joyner, "Peripheral and semi-peripheral community: A new design challenge for learning at scale," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 313–316.

D. A. Joyner and C. Isbell, "Master's at scale: Five years in a scalable online graduate degree," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 2019, pp. 1–10.

D. Joyner, "The CHI of teaching online: Blurring the lines between user interfaces and learner interfaces," in *Designing for the User Experience in Learning Systems*, Springer, 2018, pp. 81–102.

G. D. Kader and M. Perry, "Variability for categorical variables," *Journal of Statistics Education*, vol. 15, no. 2, 2007.

A. Kamenetz. (May 2020). Colleges faces student lawsuits seeking refunds after coronavirus closures. NPR, Ed., [Online]. Available: https://www.npr.org/2020/05/29/863804342/colleges-face-student-lawsuits-seeking-refunds-after-coronavirus-closures.

Y. Kanazawa, "Emotion as "deeper" than cognition: Theoretical underpinnings and multidisciplinary lignes de faits to the Emotion-Involved Processing Hypothesis (EIPH)," 国際学研究= *Journal of International Studies*, vol. 9, no. 1, pp. 185–206, 2020.

A. Kassambara, *rstatix: Pipe-friendly framework for basic statistical tests*, R package version 0.6.0, 2020. [Online]. Available: https://CRAN.R-project.org/package=rstatix.

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

R. F. Kizilcec and D. Goldfarb, "Growth mindset predicts student achievement and behavior in mobile learning," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 2019, pp. 1–10.

R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado, "Recommending self-regulated learning strategies does not improve performance in a MOOC," in *Proceedings of the third (2016) ACM Conference on Learning@ Scale*, 2016, pp. 101–104.

N. Kornell, "Metacognition in humans and animals," *Current Directions in Psychological Science*, vol. 18, no. 1, pp. 11–15, 2009.

V. Kovanovic, M. Scheffel, N. Pinkwart, and K. Verbert, "LAK 2020 program chairs' welcome," in *ACM International Conference Proceeding Series*, 2020, p. iii.

M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

E. R. Lai and M. Viering, "Assessing 21st century skills: Integrating research findings," *Pearson*, 2012.

J. Lappeman, R. Clark, J. Evans, L. Sierra-Rubia, and P. Gordon, "Studying social media sentiment using human validated analysis," *MethodsX*, vol. 7, p. 100 867, 2020.

R. B. Larson, “Controlling social desirability bias,” *International Journal of Market Research*, vol. 61, no. 5, pp. 534–547, 2019.

C. V. Le, Z. A. Pardos, S. D. Meyer, and R. Thorp, “Communication at scale in a MOOC using predictive engagement analytics,” in *International Conference on Artificial Intelligence in Education*, Springer, 2018, pp. 239–252.

Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International Conference on Machine Learning*, PMLR, 2014, pp. 1188–1196.

A. Linden and J. Fenn, “Understanding Gartner's hype cycles,” *Strategic Analysis Report N° R-20-1971. Gartner, Inc*, vol. 88, 2003.

H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

H. W. Marsh and L. A. Roche, “Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility,” *American Psychologist*, vol. 52, no. 11, p. 1187, 1997.

W. Matcha, D. Gašević, A. Pardo, *et al.*, “A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective,” *IEEE Transactions on Learning Technologies*, vol. 13, no. 2, pp. 226–245, 2019.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, “Japanese morphological analysis system ChaSen version 2.0 manual,” *NAIST Technical Report*, 1999.

W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

A. M. Mellis and W. K. Bickel, “Mechanical Turk data collection in addiction research: Utility, concerns and best practices,” *Addiction*, vol. 115, no. 10, pp. 1960–1968, 2020.

P. Nedungadi and M. Remya, “Incorporating forgetting in the personalized, clustered, Bayesian knowledge tracing (PC-BKT) model,” in *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, IEEE, 2015, pp. 1–5.

D. J. Nicol and D. Macfarlane-Dick, “Formative assessment and self-regulated learning: A model and seven principles of good feedback practice,” *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218, 2006.

T. Nixton and M. Yudelson. (2013). Functions for fitting Bayesian knowledge tracing (BKT) models from data, [Online]. Available: https://github.com/IEDMS/REDM/tree/master/bkt.

P. Norvig, *English letter frequency counts: Mayzner revisited or ETAOIN SRHLDCU*, Retrieved February 23, 2021 from http://norvig.com/mayzner.html, nd.

K. Ohtani and T. Hisasaka, "Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance," *Metacognition and Learning*, vol. 13, no. 2, pp. 179–212, 2018.

D. F. Onah, J. E. Sinclair, and R. Boyatt, "Exploring the use of MOOC discussion forums," in *Proceedings of London International Conference on Education*, 2014, pp. 1–4.

Oregon State University. (2021). Make better to-do lists, [Online]. Available: https://success.oregonstate.edu/learning/better-lists.

P. Ortiz and D. F. Harrell, "Chimeria: Grayscale MOOC: Towards critical self-reflection at scale," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 2019, pp. 1–4.

C. Papageorgiou and A. Wells, "An empirical test of a clinical metacognitive model of rumination and depression," *Cognitive therapy and research*, vol. 27, no. 3, pp. 261–273, 2003.

L. Pappano. (Nov. 2012). The year of the MOOC. T. N. Y. Times, Ed., [Online]. Available: https://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html.

Z. A. Pardos, Y. Bergner, D. T. Seaton, and D. E. Pritchard, "Adapting Bayesian knowledge tracing to a massive open online course in edX," in *Sixth International Conference on Educational Data Mining*, International Educational Data Mining Society, 2013, pp. 137–144.

Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a Bayesian networks implementation of knowledge tracing," in *International Conference on User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 255–266.

——, "KT-IDEM: Introducing item difficulty to the knowledge tracing model," in *International Conference on User Modeling, Adaptation, and Personalization*, Springer, 2011, pp. 243–254.

Z. A. Pardos, S. Tang, D. Davis, and C. V. Le, "Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework," in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 2017, pp. 23–32.

J. Piaget, "Cognitive development in children: Piaget," *Journal of Research in Science Teaching*, vol. 2, no. 3, pp. 176–186, 1964.

C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 505–513.

R. F. Pierson. (May 2013). Every kid needs a champion. T. Talks, Ed., [Online]. Available: https://www.ted.com/talks/rita_pierson_every_kid_needs_a_champion.

ProZ.com Translation Services, *Word count of translation into English from Japanese*, Retrieved February 22, 2021 from https://www.proz.com/forum/japanese_%E6%97%A5%E6%9C%AC%E8%AA%9E/14521-word_count_of_translation_into_english_from_japanese.html, nd.

J. Reich, "Two stances, three genres, and four intractable dilemmas for the future of learning at scale," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 3–13.

——, "Ed tech's failure during the pandemic, and what comes after," *Phi Delta Kappan*, vol. 102, no. 6, pp. 20–24, 2021.

S. P. Reise, A. T. Ainsworth, and M. G. Haviland, "Item response theory: Fundamentals, applications, and promise in psychological research," *Current Directions in Psychological Science*, vol. 14, no. 2, pp. 95–101, 2005. DOI: 10.1111/j.0963-7214.2005.00342.x.

M. A. Revilla, W. E. Saris, and J. A. Krosnick, "Choosing the number of categories in agree–disagree scales," *Sociological Methods & Research*, vol. 43, no. 1, pp. 73–97, 2014.

E. Ribarsky, "Choose your own adventure: Examining social exchange theory and relational choices," *Communication Teacher*, vol. 27, no. 1, pp. 29–32, 2013.

R. Rochmad, M. Kharis, A. Agoestanto, M. Z. Zahid, and M. Mashuri, "Misconception as a critical and creative thinking inhibitor for mathematics education students," *Unnes Journal of Mathematics Education*, vol. 7, no. 1, pp. 57–62, 2018.

I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger, "Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking," *Metacognition and Learning*, vol. 2, no. 2, pp. 125–140, 2007.

X. Rong, *deepnet: Deep learning toolkit in r*, R package version 0.2, Mar. 2014. [Online]. Available: https://CRAN.R-project.org/package=deepnet.

C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

D. Sale, "Metacognitive capability: The superordinate competence for the twenty-first century," in *Creative Teachers*, Springer, 2020, pp. 77–129.

G. Schraw, "Measuring self-regulation in computer-based learning environments," *Educational Psychologist*, vol. 45, no. 4, pp. 258–266, 2010.

G. Schraw and R. S. Dennison, "Assessing metacognitive awareness," *Contemporary Educational Psychology*, vol. 19, no. 4, pp. 460–475, 1994.

G. Sedrakyan, J. Malmberg, K. Verbert, S. Järvelä, and P. A. Kirschner, "Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation," *Computers in Human Behavior*, vol. 107, p. 105 512, 2020.

R. Sellers, A. Wells, and A. P. Morrison, "An experimental manipulation of negative metacognitive beliefs in non-clinical paranoia: Effects on intrusions and state anxiety," *Journal of Experimental Psychopathology*, vol. 9, no. 3, jep–062 117, 2018.

C. Seng, M. K. J. Carlon, and J. S. Cross, "Information literacy training effectiveness on Cambodia's province-based undergraduates," *International Journal of Comparative Education and Development*, 2020.

——, "Research self-efficacy of Cambodian undergraduate students at province-based universities," *International Journal of Sociology of Education*, vol. 9, no. 2, pp. 155–190, 2020.

R. Setiawan, W. Budiharto, I. H. Kartowisastro, and H. Prabowo, "Finding model through latent semantic approach to reveal the topic of discussion in discussion forum," *Education and Information Technologies*, vol. 25, no. 1, pp. 31–50, 2020.

D. Shah. (Dec. 2020). The second year of the MOOC: A review of MOOC stats and trends in 2020 - class central, [Online]. Available: https://www.classcentral.com/report/the-second-year-of-the-mooc/.

O. Skrypnyk, S. Joksimovic, V. Kovanovic, D. Gasevic, and S. Dawson, "Roles of course facilitators, learners, and technology in the flow of information of a cMOOC," *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 3, 2015.

A. Sloman, *Varieties of metacognition in natural and artificial systems*, 2011.

P. Smolen, Y. Zhang, and J. H. Byrne, "The right time to learn: Mechanisms and optimization of spaced learning," *Nature Reviews Neuroscience*, vol. 17, no. 2, p. 77, 2016.

G. Song, "Sentiment analysis of Japanese text and vocabulary learning based on natural language processing and SVM," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2021.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

E. Strickland. (Feb. 2021). OpenAI’s GPT-3 speaks! (Kindly disregard toxic language), [Online]. Available: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business.

K. Sun, A. H. Mhaidli, S. Watel, C. A. Brooks, and F. Schaub, “It’s my data! Tensions among stakeholders of a learning analytics dashboard,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

S. Sutipitakwong and P. Jamsri, “Pros and cons of tangible and digital wireframes,” in *2020 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2020, pp. 1–5.

M. D. Svinicki, *Learning and motivation in the postsecondary classroom.* Anker Publishing Company, 2004.

K. Sweeny and J. A. Shepperd, “The costs of optimism and the benefits of pessimism.,” *Emotion*, vol. 10, no. 5, pp. 750–753, 2010. DOI: 10.1037/a0019016.

J. Sweller, J. J. van Merriënboer, and F. Paas, “Cognitive architecture and instructional design: 20 years later,” *Educational Psychology Review*, vol. 31, no. 2, pp. 261–292, 2019.

P. Tarricone, *The taxonomy of metacognition*. Psychology Press, 2011.

W. L. Taylor, “"cloze procedure": A new tool for measuring readability,” *Journalism Quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

D. Tempelaar, B. Rienties, and Q. Nguyen, “Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application,” *PLOS One*, vol. 15, no. 6, e0233977, 2020.

Texas A & M University. (nd). Course design, [Online]. Available: https://cte.tamu.edu/Instructional-Resources/Course-Design.

R. H. Thaler, “From cashews to nudges: The evolution of behavioral economics,” *American Economic Review*, vol. 108, no. 6, pp. 1265–87, 2018.

The Office of Research Integrity. (nd). The lab, [Online]. Available: https://ori.hhs.gov/content/thelab.

S. Tobias and H. T. Everson, “Knowing what you know and what you don’t: Further research on metacognitive knowledge monitoring. Research Report No. 2002-3,” *College Entrance Examination Board*, 2002.

Tohoku University. (Jan. 2021). BERT base Japanese (IPA dictionary, whole word masking enabled), [Online]. Available: https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking.

M. Üce and I. Ceyhan, "Misconception in chemistry education and practices to eliminate them: Literature analysis.," *Journal of Education and Training Studies*, vol. 7, no. 3, pp. 202–208, 2019.

United Nations Educational, Scientific and Cultural Organization. (Jun. 2020). Education: From disruption to recovery, [Online]. Available: https://en.unesco.org/covid19/educationresponse.

United Nations Sustainable Development Goals. (2017). Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all, [Online]. Available: https://unstats.un.org/sdgs/report/2017/goal-04/.

K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.

L. Vangsness and M. E. Young, "More isn't always better: When metacognitive prompts are misleading," *Metacognition and Learning*, pp. 1–22, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

M. Victori and W. Lockhart, "Enhancing metacognition in self-directed language learning," *System*, vol. 23, no. 2, pp. 223–234, 1995.

P. Virtanen, H. M. Niemi, and A. Nevgi, "Active learning and self-regulation enhance student teachers' professional competences," *Australian Journal of Teacher Education*, vol. 42, no. 12, p. 1, 2017.

W. Wan and C. Y. Johnson. (May 2020). Are we near to having a vaccine for Covid-19? W. Post, Ed., [Online]. Available: https://www.washingtonpost.com/health/2020/05/27/coronavirus-endemic/.

Web Courseworks. (2021). eLearning hype curve predictions, [Online]. Available: https://webcourseworks.com/elearning-predictions-hype-curve/.

M. Wen, D. Yang, and C. Rose, "Sentiment analysis in MOOC discussion forums: What does it tell us?" In *Educational data mining 2014*, Citeseer, 2014.

J. Wilkowski, D. M. Russell, and A. Deutsch, "Self-evaluation in advanced power searching and mapping with Google MOOCs," in *Proceedings of the first ACM Conference on Learning@ Scale*, 2014, pp. 109–116.

S. Won, L. C. Hensley, and C. A. Wolters, "Brief research report: Sense of belonging and academic help-seeking as self-regulated learning," *The Journal of Experimental Education*, vol. 89, no. 1, pp. 112–124, 2021.

M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," in *International Conference on Artificial Intelligence in Education*, Springer, 2013, pp. 171–180.

F. M. Zanzotto, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, 2019.

J. Zhao, S. Bhatt, C. Thille, D. Zimmaro, and N. Gattani, "Interpretable personalized knowledge tracing and next learning activity recommendation," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 325–328.

L. Zhao and C. Ye, "Time and performance in online learning: Applying the theoretical perspective of metacognition," *Decision Sciences Journal of Innovative Education*, vol. 18, no. 3, pp. 435–455, 2020.

Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, "MOOC dropout prediction using FWTS-CNN Model based on fused feature weighting and time series," *IEEE Access*, vol. 8, pp. 225 324–225 335, 2020.

M. Zhu, C. J. Bonk, and M. Y. Doo, "Self-directed learning in MOOCs: Exploring the relationships among motivation, self-monitoring, and self-management," *Educational Technology Research and Development*, pp. 1–21, 2020.

B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educational Psychologist*, vol. 25, no. 1, pp. 3–17, 1990.

ACKNOWLEDGEMENTS

When Professor Jeffrey Cross asked me at the start of this journey why do I feel qualified to pursue a doctoral degree, I had many insecurities: no substantial research experience and subpar master's degree performance, just to name a few. I was very sure he would decline at some point. I am grateful that he decided to believe in me and graciously provided mentorship and an environment to help me succeed. Indeed, this work would be impossible without him.

There are other people I need to thank, too.

The primary data collection for this research was supported by the Tokyo Tech Electrical Engineering Literacy teaching team: Professor Akira Chiba, Professor Takayuki Iwasaki, and Professor Khanh Tran Gia. The members of my assessment committee – Professor Junichi Takada, Professor Kayoko Nohara, Professor Akihito Yamashita, Professor Naoaki Okazaki, and Professor Daniel Berrar – have all provided valuable feedback on my research. Their critiques had prompted me to attempt new approaches, investigate different lines of thought, and clarify my presentation to prevent errors and misunderstandings.

The professors, students, and staff at the Cross lab and the Online Education Development Office then and now have provided me multiple opportunities to test out my ideas and take leadership in many ways. Every discussion, brainy or silly, made challenges easier to tackle. Please continue thinking outside the box and having each other's back.

いつもウィズダム株式会社にお世話になっております。皆様のご支援がなければ、自信を持って博士号を取得することはできませんでした。本当に感謝しております。

I am fortunate to have friends willing to extend their expertise and kindness. They participated in interviews, reviewed my work, provided consultation, and encouraged me. They helped me navigate graduate school, and they continue rooting for me as I explore career opportunities. Finally, this section would not be complete without giving thanks to my family. I can do the dumbest thing on earth, and to them, I will still be the shining light of genius. My husband, Mike, had always been with me through every ups and downs. His sacrifices and the motivation he gives me are beyond measure.

The Japan Society for Promotion of Science (JSPS) supported this work via the Grants-in-Aid for Scientific Research (Kakenhi) Grant Number JP20H01719.

And thank you, reader, for reading. This has been a labor of love, so I hope you recover from the cringe.