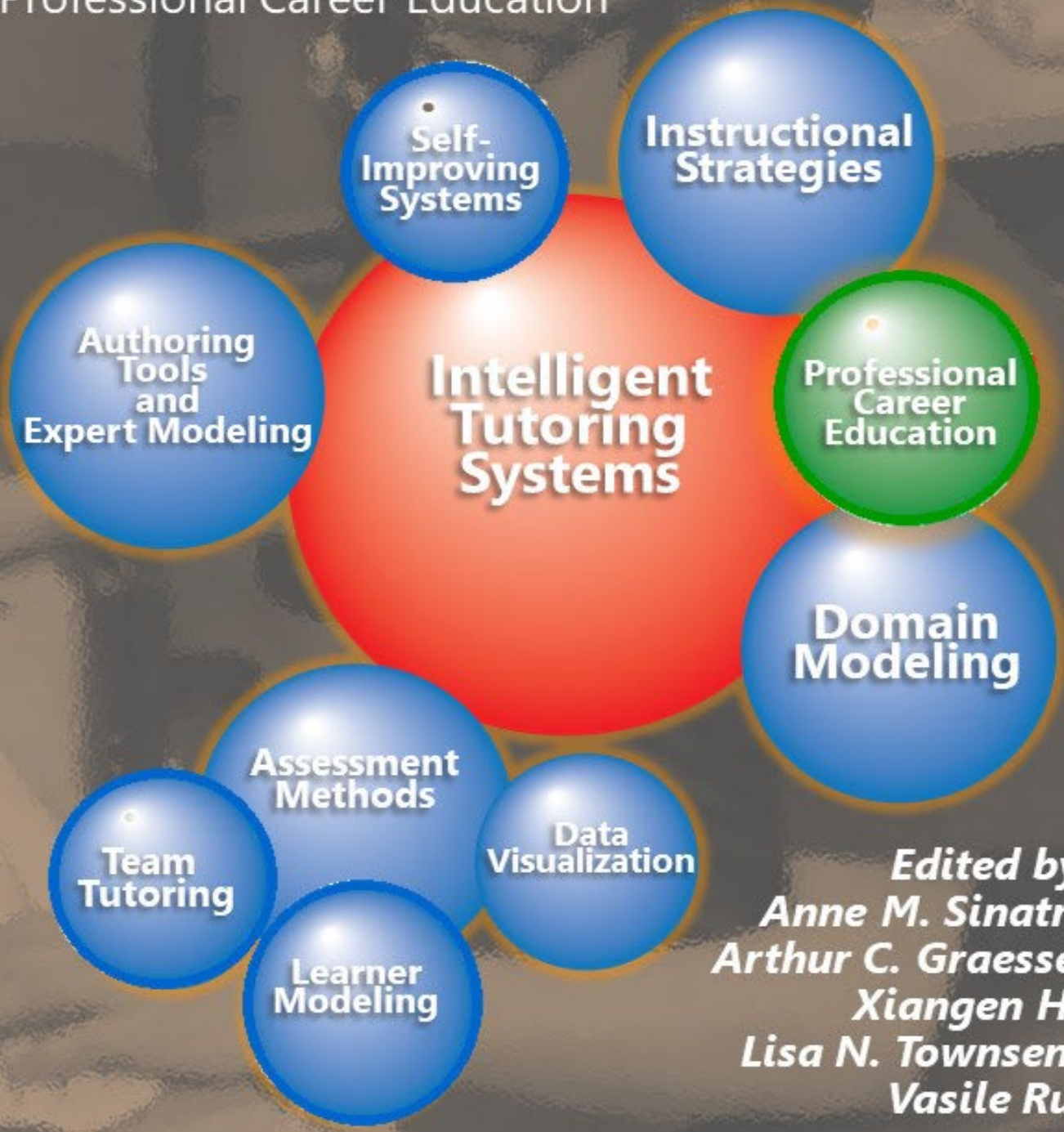
Design Recommendations for Intelligent Tutoring Systems

Volume 11
Professional Career Education



Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Lisa N. Townsend
Vasile Rus

A Book in the Adaptive Tutoring Series

Design Recommendations for Intelligent Tutoring Systems

Volume 11
Professional Career Education

Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Lisa N. Townsend
Vasile Rus

A Book in the Adaptive Tutoring Series

Printed in the United States of America
First Printing, September 2023

US Army Combat Capabilities Development Command - Soldier Center
Simulation and Training Technology Center
Orlando, Florida

International Standard Book Number:
978-0-9977258-5-8

Special thanks to Jody Cockroft, University of Memphis, for her efforts in coordinating the workshop that led to this volume.

Dedicated to current and future scientists and developers of adaptive learning technologies

CONTENTS

INTRODUCTION TO PROFESSIONAL CAREER EDUCATION APPLICATIONS

Anne M. Sinatra[1], Arthur C. Graesser[2], Xiangen Hu[2], Lisa N. Townsend[1], and Vasile Rus[2], Eds.

[1]U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center
[2]University of Memphis Institute for Intelligent Systems

The *Design Recommendations for Intelligent Tutoring Systems* series has covered many different topics over the past ten years. Those topics have ranged from general components of intelligent tutoring systems (ITSs) (Learner Modeling, Instructional Management, Authoring Tools, Domain Modeling) to advanced elements (Assessment Methods, Team Tutoring, Self-Improving Systems, Data Visualization, Competency Based-Scenario Design). Our most recent previous volume included a series of Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analyses on all the initial topics as well as overviews of ITSs in general and the Generalized Intelligent Framework for Tutoring (GIFT) software (Sottilare et al., 2012; Sottilare et al., 2017; Goldberg & Sinatra, 2023).

Each book in the *Design Recommendations for Intelligent Tutoring Systems* series has been associated with an Expert Workshop on the same topic. These workshops are part of a cooperative agreement (W911NF-18-2-0039) between US Army Combat Capabilities Development Command (DEVCOM) Soldier Center and University of Memphis. One of the goals of the expert workshops is to learn more about ITS capabilities that are being developed, and how these approaches, as well as lessons learned, could enhance the GIFT software (GIFT is freely available at https://www.GIFTtutoring.org). Invited experts in industry, academia, and government discuss the expert workshop topic, their applicable work, and suggestions for improving GIFT in what is usually a two day event. Both the University of Memphis and GIFT Teams participate in the workshop, help to guide discussion, and ask questions that will provide insight into current challenges in GIFT.

The expert workshop associated with this current book was held virtually in October 2022, and included presentations about both general approaches and specific applications to professional education in ITSs. Additionally, the University of Memphis team that participated in the workshop included Arthur C. Graesser, Xiangen Hu, Vasile Rus, and Jody Cockroft. The US Army DEVCOM Soldier Center team who participated in the workshop included Benjamin Goldberg, Gregory Goodwin, Anne M. Sinatra, Randall Spain, and Lisa N. Townsend.

The current volume and the expert workshop that was associated with it, branched out in a new direction and rather than addressing specific components of an ITS or types of features/approaches that could be included in ITSs, it focused on how to apply an ITS for specific types of training. The specific focus was on ITSs for Professional Career Education. This topic area was selected, as in general, ITS research tends to be focused on K-12 or college education, and in many cases on domains such as algebra or physics. However, for the military, and for industry, trainees are adult learners and domains tend to be more active, applied, and experiential. This workshop provided an opportunity for discussion of specific examples of applied training that occurs with ITSs, as well as discussion of general approaches and considerations for applied professional education in ITSs.

Sections of the Book

This book is organized into two sections:

I. Standards and General Approaches

II. Specific Applications

Section I includes chapters that discuss general approaches and techniques that can be used with ITSs for Professional Education.

Section II includes specific applications and use cases that ITSs have been used for in professional education including worker training, commercial pilots, suicide prevention training, teaching, and medical education.

References

Goldberg, B., & Sinatra, A.M. (2023). Generalized Intelligent Framework for Tutoring (GIFT) SWOT analysis. In Sinatra, A.M., Graesser, A.C., Hu, X., Goodwin, G., & Rus, V. (Eds.), Design Recommendations for Intelligent Tutoring Systems, Volume 10: Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems, pp. 9-26. US Army DEVCOM Soldier Center.

Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). *The Generalized Intelligent Framework for Tutoring (GIFT)*. Orlando, FL: U.S. Army Research Laboratory Human Research & Engineering Directorate (ARL-HRED).

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017.

Acknowledgements

Research was sponsored by the U.S. Army Combat Capabilities Development Command (DEVCOM) – Soldier Center (SC) and was accomplished under Cooperative Agreement Number W911NF-18-2-0039. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM SC or the US Government.

SECTION I – STANDARDS AND GENERAL APPROACHES

Vasile Rus[1], Lisa N. Townsend[2], Arthur C. Graesser[1], and Anne M. Sinatra[2], Eds.

[1]University of Memphis Institute for Intelligent Systems
[2]U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center

CHAPTER 1 – INTRODUCTION TO STANDARDS AND GENERAL APPROACHES

Lisa N. Townsend[1], Vasile Rus[2], Anne M. Sinatra[1], and Arthur C. Graesser[2]
U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center[1]; University of Memphis[2]

Core Ideas

The chapters in this section address the critical need for operational advanced technologies, such as artificial intelligence (AI) and machine learning (ML), to maintain a globally enduring strategic advantage in training and readiness. This collection of chapters cover a broad range of ideas linked by an overarching theme: the need to develop AI-driven systems to support the acquisition of higher-level skills in professional development and training, including the ability to perform well in teams. Standards are currently playing a growing role in the development of an AI-driven, competency-based, working and learning ecosystem.

Important focus areas include determining how well training addresses a required skillset and automating links to effective training opportunities to learn new skills. These areas can be used to inform and enhance both advances in Intelligent Tutoring Systems (ITSs) and the proficiencies that the Generalized Intelligent Framework for Tutoring (GIFT) amplifies. Identifying which skills a person has already mastered and which ones the person needs to build can create an advantage in obtaining a highly skilled, adaptable workforce in an ever-changing operational environment. These learner skills and other data elements contribute to a larger data management infrastructure that can be quickly used to inform education and training decisions. The infrastructure is growing more complex as it accumulates data from multiple sources, across different technologies and platforms, all integrated to advise the best learning opportunities, at the right time, aligned to the individual (or team).

Individual Chapters

The chapter by *Goodell* and *Hu* argues for the need of scaled life-long talent development given the shortcomings of traditional educational systems that cannot keep pace with accelerated developments in AI and its impact on various professions. The authors discuss the importance of developing soft skills that help the learner adapt to potentially changing work. The authors assert that new models for scaled life-long talent development should be built on three pillars: learning engineering, adaptive learning technologies, and standards.

Lester, Min, Rowe, Smith, and *Spain* discuss how AI technologies can be used to develop team training systems. Teamwork is essential in complex enterprises whether in the workplace or the military. The authors discuss the need to translate research-based team training strategies into AI-driven team training systems, as well as the need for developing team assessments that are reliable and actionable (i.e., can inform effective feedback). Another important need is for AI-driven generation of competency-based team scenarios.

Sottilare describes different approaches to professional development and how ITS research and capabilities may impact the professional development process. AI and ML have been applied to determine ideal times for interventions during intelligent tutoring, but the type of learning that occurs in professional development may differ from the scope of traditional ITSs. The chapter describes the current state of ITSs and

opportunities for them to be utilized for continuing education. There are discussions of ITS effectiveness, credibility, affordability, engagement, and accessibility in the context of professional development.

The chapter by *Robson, Kelsey, Egerton, Nasir,* and *Glover* highlights the *SkillSync*™ project's alignment service, a web application that connects companies to college development programs and training providers, thereby offering upskilling opportunities to workers. This AI-enabled service computes an *alignment score,* the degree to which a set of training materials covers a prioritized list of desired skills and a program of instruction that meets the upskilling needs of the company. This ability to fill real-time learner gaps and reconfigure content is part of the outer loop in an ITS model that is responsible for selecting learning experiences on a set of desired skills.

Milham and *Smith* articulate the vision of a career-long learning ecosystem through application of Total Learning Architecture (TLA) standards. The desired end state requires interoperability of various software systems in order to exchange, understand, and use data as well as to manage lifelong learning and to support a data-driven organization. This TLA Data Strategy involves an ecosystem comprised of learning record providers and/or learning record consumers, leading to an individual event-driven architecture. Learning experiences encountered across a career and performance experiences are linked with competencies, credentials, and outputs aligned to multiple career path options and career milestones through a conceptual model, which is explored through a professional career education use case.

CHAPTER 2 – STANDARDS FOR INTELLIGENT TUTORING FOR THE CONVERGENCE OF LEARNING AND WORKING

James Goodell[1] and Xiangen Hu[2]
Quality Information Partners[1]; University of Memphis[2]

Introduction

Artificial intelligence is changing the nature of human endeavor and productivity. Traditional professional career education systems cannot adapt fast enough to meet the pace of this change and the need for life-long learning. This chapter proposes that new models of scaled life-long talent development can be built on three foundational pillars: 1) learning engineering, 2) adaptive learning technologies, including intelligent tutoring systems (ITSs), and 3) technical standards.

The Challenge

The nature of working and learning is changing. The service economy that followed the industrial age is being replaced with an "intelligence augmentation economy" in which teams of human workers and intelligent agents are working and learning together (Craig & Goodell, 2022).

Professional career education has traditionally preceded and provided primary foundational knowledge for a person's career followed by continuing education and on the job training, as secondary and supplemental, to maintain skills and certifications required for some careers. The changing nature of work increasingly depends on new knowledge that did not exist during a person's pre-career education, while ubiquitous access to factual knowledge renders traditional education models and some job tasks obsolete. Instead, *learning ability* itself, is among the top competencies required for the modern workforce. Workers must dynamically develop new skills and capabilities in response to an ever-changing business environment to support businesses that compete through innovation and agility in a rapidly changing world.

According to the third edition of the World Economic Forum's *Future of Jobs Report 2020* (2020), the top five skills needed for 2025 will be:

- analytical thinking and innovation,
- active learning,
- complex problem solving,
- creativity, and
- leadership and social influence.

If this is true, then developing these soft skills should be a new priority for pre-career professional education. This shift in priority for pre-career professional education would mean a "retooling" of the current academic and professional training institutions.

In-career continuing education also requires "retooling" to match the pace, scale, and agility or learning and development, and changing nature of work, so that organizations have the human-capacity to thrive in a changing world and to defend against new threats to their existence.

The challenge is to adapt current systems and models of learning to support, at scale, the new nature of work in which humans and intelligent agents work and learn together.

ITS technologies are one of three pillars that we propose can support new models of scaled life-long work-embedded talent development. We see ITSs increasingly used in work-embedded learning in which intelligent agents are both assisting with work tasks and assisting with learning.

Components embedded in on-the-job intelligence augmentation systems can monitor task performance and learning behaviors similar to what happens in ITSs (Tang et al., 2021) and continuously update learner models, so as to detect the difficulties of learners/workers, opportunities to optimize work processes, and to facilitate learning.

State of the Field and Supporting Research

The Learning Analytics and Knowledge (LAK), Artificial Intelligence in Education (AIED), Educational Data Mining (EDM) and other research communities have developed models, methods, and technologies (Siemens & Baker, 2012) that can be applied as algorithms within ITSs. Figure 1 shows some of the methods for prediction, classification, knowledge inference, and behavior detection in part of a learning analytics process model as presented in the "Data Analysis Tools" (Czerwinski, Domadia et al., 2022) chapter of *Learning Engineering Toolkit*. Researchers continue to innovate on and test variations of these methods in various learning contexts (Zhang et al., 2021). As larger training data sets become available machine learning has the potential to predict, classify, infer, and detect with even greater precision.

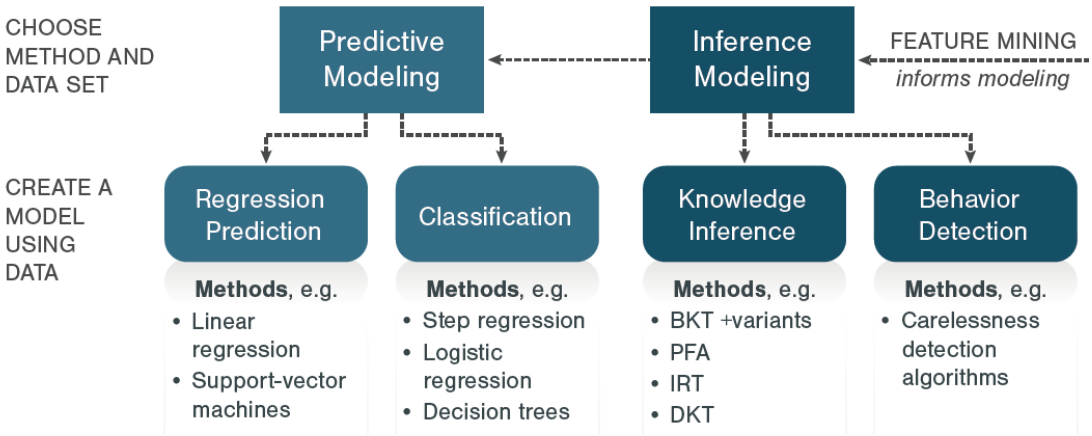


Figure 1. Learning analytics methods from learning analytics process model for learning engineering. (CC BY Jim Goodell and Steve Ritter; Attribution 2.0 Generic License: https://creativecommons.org/licenses/by/2.0/) Figure reprinted with permission from Czerwinski, Domadia et al. (2022). Data Analysis Tools. In Learning Engineering Toolkit, 367. Routledge.

Other researchers and other research communities have begun to develop and test standardized components for systems that support learning as a means to mature the practice of learning engineering (Saxberg, 2022; Goodell & Thai, 2020). An award-winning discussion paper suggests that the IEEE Standards Association recognize learning engineering as a new field of engineering (Goodell, Jay et al., 2022).

Research has advanced the state-of-the-art of ITSs. As documented in this Design Recommendations for Intelligent Tutoring Systems series and continuing research at the Soldier Center (Goldberg et al., 2021) the Generalized Intelligent Framework for Tutoring (GIFT) framework is being applied to work-embedded

and team training contexts. These systems are ready to take on the challenge of scaled work-embedded intelligence augmentation.

Discussion

New models of scaled life-long talent development can be built on three foundational pillars:

> **Learning engineering** - the process and practice that applies the learning sciences using human-centered engineering design methodologies and data-informed decisions making (Goodell & Kolodner, 2022). Like in other domains, we need science to discover truths about learning, but we need engineering to create scalable solutions to problems using science as one tool in that endeavor.
>
> **Adaptive learning technologies** (including ITSs) provide a scalable platform for deeply contextualized adaptation to optimize learning for every learner or team of professional learners.
>
> **Standards** - technology standards, process standards, practice standards, and engineering design patterns allow learning engineers to develop solutions from a systems perspective with access to reusable components that enable scaling of complex systems.

Learning Engineering

We propose learning engineering as one of the pillars for ITSs applied to future professional career education because the scale and complexity of the challenge calls for a greater level of rigor and scale than currently exist. We see the maturation of learning engineering as a multidisciplinary practice as a critical success factor in responding to this challenge.

According to IEEE ICICLE Learning engineering is an **iterative process** and practice that:

1. applies the learning sciences,
2. uses human-centered engineering design methodologies, and
3. uses data-informed decision-making

to support learners and their development. (Institute of Electrical and Electronics Engineers IC Industry Consortium on Learning Engineering, 2019).

Learning engineering is most often done as a team process in which team members bring expertise from a variety of professional specialties to *do* learning engineering as a coordinated effort, rooted in a shared understanding and vocabulary. The team members collectively have broad knowledge of engineering processes as well as of learning science, computer science, data science, instructional design, artificial intelligence and machine learning, pedagogy, and andragogy, human-centered user experience design, product testing, and the development of policies, regulations, and standards. The exact competencies required by a learning engineering team depends on the challenge and the contextual factors in the problem space (Goodell, 2022).

Learning engineering is a process. Learning engineering is a repeatable process intended to iteratively design, test, adjust, and improve conditions for learning. As shown in Figure 2 (Kessler et al., 2022), it starts with defining the *challenge* and the contextual factors surrounding that challenge, including understanding the learners, learning environment, team and resource constraints (Kessler et al., 2022).

Depending on the nature of the challenge the learning engineering team may begin a *creation* phase of iterative design and development, an *implementation* phase, or an *investigation* phase.

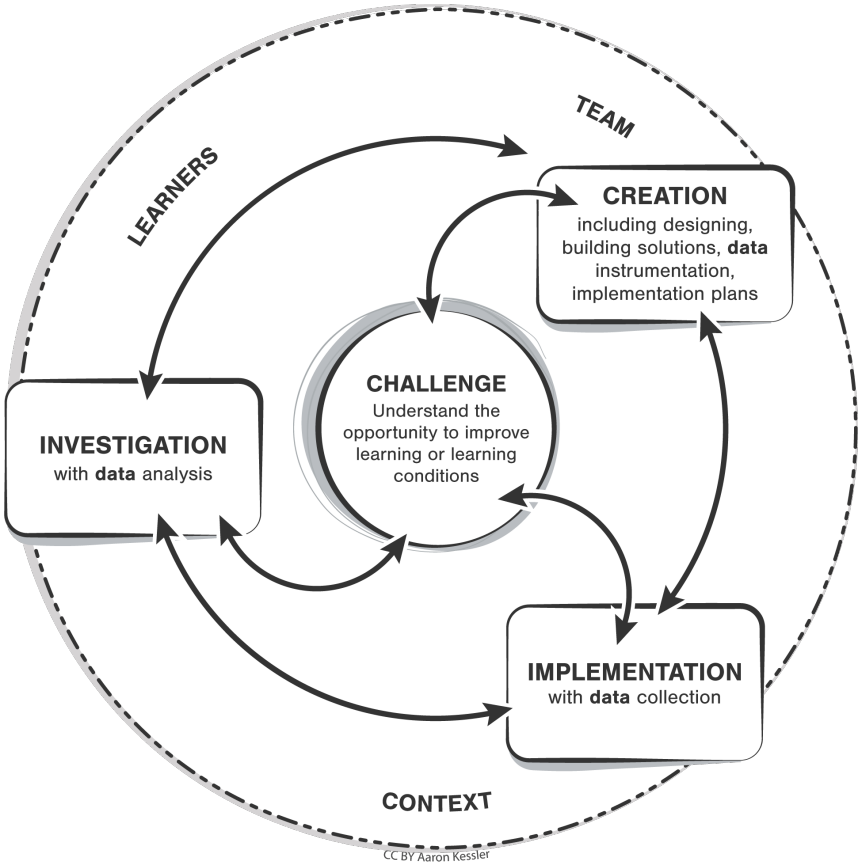


Figure 2. The learning engineering process. (CC BY Aaron Kessler; Attribution 2.0 Generic License: https://creativecommons.org/licenses/by/2.0/) Figure reprinted with permission from Kessler et al. (2022). Learning Engineering is a Process. In Learning Engineering Toolkit, 31. Routledge.

GIFT implementations use the learning engineering iterative process, applying the learning sciences, using human-centered engineering design methodologies, and data-informed decision-making. Instrumented data informs feedback within scenarios deployed with GIFT systems and to inform future iterative improvements in the system and content. In 2019, Design Recommendations for Intelligent Tutoring Systems: Volume 7 presented recommendations for data-informed self-improving systems (Sinatra et al., 2019).

Learning engineering is engineering. While different engineering domains apply different scientific discoveries and have domain-specific standards and practices, there are common principles that apply across domains. One of the benefits of applying engineering principles to learning contexts, and ITSs specifically, is the ability to develop complex systems that can be deployed at scale and operate within engineered tolerances.

Scalability of complex systems is achieved in part by breaking them into modules, with interfaces between those modules. In other fields of engineering, such as electronics engineering, standardized components are used that offer predictable functionality and tolerances within specified operating conditions. As the electronics industry matured, more complex components such as standard integrated circuits supplemented

basic components, such as resistors, capacitors, and transistors, to support rapid design of more complex products.

Scaled interoperability between modules is made possible by using standard interfaces. For learning systems, the industry has begun to develop standards for interfaces, such as Experience API (xAPI) IEEE 9274.x and standard specifications components, such as IEEE 1484.20.3 Standard for Reusable Competency Definitions. However, the fields of learning engineering for ITSs are immature compared to electronics engineering.

Learning engineering is human-centered. Learning engineering requires human-centered design that starts by understanding the challenge from the learners' perspectives and then creates solutions through research-based iterative design. Learning engineering's human-centered perspective draws from several fields including human-centered design, design thinking, universal design for learning, learning experience design, and design-based research. (Thai et al., 2022).

Learning engineering is data-informed. Data-informed decision-making is an essential and integral part of learning engineering that includes instrumentation— designing, developing, and implementing the data collection—and analytics analysis and use of data within the learning solution to inform iterative improvements to the learning solution (Czerwinski, Goodell et al., 2022).

Learning engineering is ethical. There are ethical considerations at each stage of the learning engineering process (Schoenherr, 2022) Engineering professions adopt codes of ethics. Artificial Intelligence (AI)/Machine Learning (ML) and ITS technologies bring their own set of ethical considerations such as data privacy and algorithm bias. As does the very context of the challenge, the blurring of lines between learning and working, and the short and long-terms interests and behaviors of workers/learners and employers in the use of these technologies.

Adaptive Learning Systems

The second pillar for responding to the intelligence augmentation economy is adaptive systems at the intersection of working and learning. GIFT (Sottilare et. al., 2012; Sottilare et al., 2017) and other ITS research projects, have led the way on developing the viability of adaptive learning across many learning contexts (Nye et al., 2014). However, these scientific discoveries are not yet being transferred to the field as scaled innovations for professional career education and workforce development.

De facto standard data formats, open data sets of ITS log data, and open tools have advanced foundational research in the learning science community (Stamper et al., 2010). Other researchers have developed learning analytics methods that measure learning, and to detect productive vs. unproductive learning behaviors (Barrett, 2022; Baker; 2005; Baker, 2006).

The authors recognize an opportunity to advance the process and practice of learning engineering in general and specifically for professional career education by moving ITS and learning analytics research findings into standard "components" for engineered learning systems.

Standards

While reusable software code from frameworks like GIFT and from other research projects exist, the industry has not developed comprehensive standards for ITS components. Systems and ecosystems in the workplace have already been transformed with data and technology standards for cloud-based, AI-enhanced, instrumented, process improvement. New standards are being developed (HR Open, 2022) for systems that support skills-based hiring and advancement (US Chamber of Commerce Foundation, 2022),

with the potential to provide macro-adaptive feedback for corporate and government policy while empowering workers with new pathways and insights. In connection with the Advanced Distributed Learning Initiative, research has informed standards development in areas such as competency definitions, learning experience instrumentation, and learning experience delivery.

For data instrumentation, xAPI Profiles (Advanced Distributed Learning Initiative, 2020) serve to standardize and constrain interfaces between system modules for semantic interoperability applied to specific pedagogical and learning-context.

GIFT research and similar initiatives have led the way for proving the effectiveness of ITSs and learning engineering processes on a limited scale. For scaled impact we recommend further research toward technology transfer that specifies new modules that by design can become standard components and standard interfaces used for scaled engineering of ITSs.

The new nature of work in which humans and intelligent agents work and learn together will require a faster pace of learning and faster pace of ITS engineering that the authors believe can only be possible with standard reusable components for more rapid and predictable design and development of systems.

A key enabler for advancement of ITS engineering could be a standardized module for plug-and-play inference-making such as for inferring a learner/worker's competence level on a task and whether they are engaged in productive learning/working behaviors.

A set of inference-making components could be standardized with (1) an xAPI Profile defining inputs, (2) a standard for encoding a "system of equations" and context parameters for inference-making, and (3) a standard for outputs via xAPI to nodes in a learner model graph. Output standards could include data serialization formats for updating probability matrices mapped to competency definitions and context tags based on existing standards. Other interface standards could address updates to prior inference data in the learner model graph. Additional standardized modules could define inference-based triggering for immediate next step branching of the learner experience, e.g., based on a detected misconception. And component standards could define rules for feedback, prompts, and branching.

Standardized modules and standard engineering design patterns for ITSs, that allow for flexibility in what algorithms are used to control adaptations, can support faster development and iteration in response to the changing nature of working and learning.

Recommendations for GIFT and Intelligent Tutoring Systems

GIFT is already being applied to work-embedded learning through scenario-based training, such as in STEEL-R (Goldberg et al., 2021). A key to the advancement of learning engineering, so that we can move beyond isolated examples toward broader and scaled impact, is the modularization of learning systems. Modularization and standardized components can enable scaled and rapid production of new systems adapted to new learning contexts.

New research could prove the feasibility of a standardized module for plug-and-play inference-making—a module that allows different inference algorithms to be plugged into ITSs, producing standardized outputs, such as learner model data. We recommend research be done in partnership with developers of two or more ITS platforms to demonstrate interoperability across systems and across two or more learning contexts.

Conclusions

ITSs have an important role to play in equipping the learning in the "intelligence augmentation economy" in which teams of learner/workers paired with teams of AI agents get work done while continuously learning about the job and each other. This new nature of working and learning requires more agile work-embedded learning that assumes both human actors and AI agents are learning from each other and from the context of the work. It requires a scale and pace of developing adaptive learning systems and experiences faster than current processes, resources, and models allow. However, principles and rigor of engineering can be applied—modularization, standardization, control theory, data-driven process improvement, etc.—so that ITSs can provide a scaled response and facilitate new models of life-long professional learning.

References

Advanced Distributed Learning Initiative (2020). xAPI Profile Server to Launch in 2021. ADL Initiative, US Department of Defense. https://adlnet.gov/news/2020/12/02/xAPI-Profile-Server-to-Launch-in-2021.

Baker, R. S. "Designing intelligent tutors that adapt to when students game the system." PhD diss. Carnegie Mellon University, 2005. http://reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-05-104.pdf

Baker, R., Corbett, A., Koedinger, K., Evenson, S., Roll, I., Wagner, A., Naim, M., Raspat, J., Baker, D., & Beck, J.. (2006). Adapting to When Students Game an Intelligent Tutoring System. 4053. 392-401. 10.1007/11774303_39.

Barrett, M., Czerwinski, E., Goodell, J., Jacobs, D., Ritter, S., Sottilare, R., & Thai, K.-P. (2022). Learning Engineering Uses Data (Part 2): Analytics. In *Learning Engineering Toolkit*, 175–198. Routledge. https://doi.org/10.4324/9781003276579-10

Craig, S. & Goodell, J. (2022). What Does an Emerging Intelligence Augmentation Economy Mean for HF/E? Can Learning Engineering Help? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, In press.

Czerwinski, E., Domadia, T., Craig, S., Goodell, J., and Ritter, S. (2022). Data Analysis Tools. In *Learning Engineering Toolkit*, 367. Routledge.

Czerwinski, E., Goodell, J., Ritter, S., Sottilare, R., Thai, K.-P., & Jacobs, D. (2022). Learning Engineering Uses Data (Part 1): Instrumentation. In *Learning Engineering Toolkit*, 153–174. Routledge. https://doi.org/10.4324/9781003276579-9

Goldberg, B., Owens, K., Gupton,K., Hellman, K., Robson, R., Blake-Plock, S. & Hoffman, M. (2021). Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2021.

Goodell, J. (2022). Introduction. In *Learning Engineering Toolkit*, 5–25. Routledge. https://doi.org/10.4324/9781003276579-3

Goodell, J., Jay, M., Olaniyi, N., & Rogers, J. (2022, July). Should IEEE Establish Learning Engineering as a New Engineering Profession?. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 44-45). IEEE.

Goodell, J., & Kolodner, J. (Eds.). (2022). Learning Engineering Toolkit: Evidence-Based Practices from the Learning Sciences, Instructional Design, and Beyond (1st ed.). Routledge. https://doi.org/10.4324/9781003276579

Goodell, J., & Thai, K.-P. (2020, July). A learning engineering model for learner-centered adaptive systems. In *International Conference on Human-Computer Interaction* (pp. 557-573). Springer, Cham.

HR Open Standards Consortium. (2022, September) Active Standards Projects. HR Open Standards Consortium. https://www.hropenstandards.org/events/active-standards-projects

Institute of Electrical and Electronics Engineers IC Industry Consortium on Learning Engineering (2019). https://sagroups.ieee.org/icicle/

Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2022). Learning Engineering is a Process. In *Learning Engineering Toolkit*, 31. Routledge.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427-469.

Saxberg, B. (2022). Tools from a component-based trade. Discussion thread in learning engineering google group. https://groups.google.com/g/learning-engineering/c/DaSAyzLvCPA

Schoenherr, J.R. (2022). Learning Engineering is Ethical. In *Learning Engineering Toolkit*, 201–228. Routledge. https://doi.org/10.4324/9781003276579-11

Siemens, G. & Baker, R. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*. Association for Computing Machinery, New York, NY, USA, 252–254. https://doi.org/10.1145/2330601.2330661

Sinatra, A.M., Graesser, A.C., Hu, X., Brawner, K., and Rus, V. (Eds.). (2019). Design Recommendations for Intelligent Tutoring Systems: Volume 7 - Self-Improving Systems. Orlando, FL: US Army CCDC. ISBN 978-0-9977257-7-3. Available at: https://gifttutoring.org/documents

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The Generalized Intelligent Frame-work for Tutoring (GIFT). Orlando, FL: Army Research Laboratory.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT).

Stamper, J., Koedinger, K., Baker, R. S. J. d., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A Data Analysis Service for the Learning Science Community. In Intelligent Tutoring Systems (pp. 455–455). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13437-1_112

Tang, Y., Li, Z., Wang, G., & Hu, X. (2021). Modeling learning behaviors and predicting performance in an intelligent tutoring system: a two-layer hidden Markov modeling approach. Interactive Learning Environments, 1–13. https://doi.org/10.1080/10494820.2021.2010100

Thai, K.-P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning Engineering is Human-Centered. In *Learning Engineering Toolkit*, 83–123. Routledge. https://doi.org/10.4324/9781003276579-7

US Chamber of Commerce Foundation. (2022) T3 Innovation Network Skills-Based Hiring and Advancement Brief. US Chamber of Commerce Foundation. https://www.uschamberfoundation.org/sites/default/files/USCCF_T3NetworkSBHAExplainer_ViewOnly.pdf

World Economic Forum (2020). The Future of Jobs Report 2020. https://www.weforum.org/reports/the-future-of-jobs-report-2020

Zhang, J., Das, R.,Baker, R. & Scruggs, R. (2021). Knowledge Tracing Models' Predictive Performance when a Student Starts a Skill. Educationaldatamining.org. Retrieved November 21, 2022, from https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_126.pdf

CHAPTER 3 – THE FUTURE OF AI-DRIVEN TEAM TRAINING

James Lester[1], Wookhee Min[1], Jonathan Rowe[1], Andy Smith[1], and Randall Spain[2]
North Carolina State University[1],
U.S. Army Combat Capabilities Development Command - Soldier Center,
Simulation and Training Technology Center[2]

Introduction

With the increasingly important role of teamwork in the twenty-first century workplace, team training has emerged as a critical focus for a broad range of professional settings. Team training has been shown to improve team effectiveness (Salas et al., 2008), and training technologies specifically designed to improve team performance show great promise. Because of significant advances in artificial intelligence (AI) in recent years, the next generation of team training systems will utilize AI to create team-based adaptive coaching, feedback, and assessment, which will be essential for meeting the U.S. Army's vision for providing competency-based training to collective units. While recent years have seen the development of intelligent tutoring systems that can deliver robust adaptive learning experiences, most research and development in intelligent tutoring systems has centered on supporting individual learning rather than on team training. In contrast, AI-driven team training will leverage emerging AI technologies to support the acquisition of core competencies that enable teams to function efficiently and effectively.

AI-driven team training offers considerable potential for training team members on how to communicate and coordinate with one another, engage in effective backup behaviors, and develop leadership competencies to facilitate team cohesion and shared mental models (Sottilare, Burke et al., 2018). A critical objective of team training is enhancing *team effectiveness,* which focuses on both team performance outcomes as well as teamwork processes that produce effective outcomes (Salas et al., 2005). Team training scenarios can be specifically designed to train teamwork skills by creating scenarios that require trainees to coordinate actions, communicate effectively, and manage conflicts. Team training scenarios can also be designed to train leadership skills to maximize team cohesion and efficacy. An essential requirement of team training scenarios is ensuring they exercise teamwork processes and coordination mechanisms that contribute to overall team effectiveness.

AI is advancing rapidly with core capabilities in natural language processing, computer vision, and machine learning becoming increasingly powerful (Zhang et al., 2021). The capacity to understand, generate, and translate language, to summarize documents, to engage in spoken language dialogue, to recognize objects and human activities, to generate images, and to accurately answer questions about documents and videos has far surpassed expectations of only a few years ago. These developments are profoundly changing the technology landscape and creating the opportunity to fundamentally re-envision how team training is designed and delivered over the next decade. In this chapter, following stage-setting remarks about team training competencies, we introduce two key families of AI-driven team training functionalities: competency-based scenario generation for team training and automated assessment of team training.

Training Team Competencies

Teams play a critical role in the workplace, including the military. The complex nature of military operations often requires multiple units to engage in coordinated and interdependent actions to achieve mission success. In these situations, it is critical that team members provide periodic updates to one another, communicate clearly and concisely, provide and request assistance when needed, and provide guidance to

team members (Salas et al., 1995; Salas et al., 1997; Salas et al., 2015; Smith-Jentsch et al., 1998). The U.S. Army's principles of Mission Command highlight the importance of building these teamwork competencies through tenets such as "Build cohesive teams through mutual trust" and "Create shared understanding" (U.S. Army, 2012).

While team training researchers have made substantial progress towards developing effective training strategies to improve team performance, these efforts have not been fully realized in AI-driven training systems that support team training. AI-driven team training systems are intelligent tutoring systems that guide learning experiences by tailoring instruction and recommendations based on the goals, needs, and preferences of each learner in the context of domain learning objectives (Sottilare, Barr et al., 2018). A critical feature of AI-driven team training systems is the capability to scaffold trainee learning. By leveraging recent advances in AI and machine learning, team-based intelligent tutoring systems are envisioned to replicate the capabilities of effective human instructors, including monitoring and tracking Soldiers' evolving competency states, assessing and diagnosing problems, and providing support and assistance as needed (TRADOC, 2017). An important challenge for team training researchers is to develop models that can be instantiated in team-based intelligent tutoring systems to support teamwork development and team performance (Johnston et al., 2018).

Translating theories of team process and performance into AI-based agents that can replicate the behaviors and insights of an effective instructor poses significant challenges. Teamwork is multifaceted, which makes it difficult to devise team coaching models that can address the behavioral, cognitive, and affective aspects of teamwork. Team performance measures should diagnose team processes, but teamwork skills are not readily quantifiable, and it can be challenging to distinguish between individual deficiencies and team-level deficiencies. Team assessments should be reliable, and they should also be actionable by guiding specific team performance feedback, but existing technologies do not yet meet the capabilities of human instructors for assessing and diagnosing a team's communication and coordination behaviors, their level of cohesion and mutual trust, or the richness of a team's shared mental model. Finally, team training systems should be able to generate scenarios including a rich set of events that will prompt team members to engage in the team processes required for successful performance. In the next section, we will discuss how advances in AI can address several of these challenges.

AI-Driven Team Scenario Generation

Competency-based team scenario generation is a particularly promising example of how AI advances can be leveraged to create highly customized team training experiences. Competency-based scenario generation for team training offers considerable potential for creating synthetic training experiences optimized for teams, tasks, contexts, and stress levels. Competency-based scenario generators are a form of interactive narrative generators that create interactive narrative experiences in which learners solve problems and complete activities in synthetic training environments that are tailored to the Soldiers' specific performance needs, competency states, and inherent training needs of their units. Interactive narrative generators can dynamically shape training experiences, story events, characters, and settings to enhance active learning and promote student agency (Wang et al., 2018). Competency-based scenario generators will leverage recent advances in machine learning and utilize data-driven approaches to support competency-driven training.

Recent years have seen a broad range of computational techniques that hold significant potential for competency-based scenario generation (Folsom-Kovarik et al., 2019), including genetic algorithm techniques (Folsom-Kovarik et al., 2018) as well as machine learning approaches based on dynamic Bayesian networks (e.g., Lee et al., 2014), deep generative models (e.g., Park et al., 2019), and reinforcement learning (e.g., Rowe & Lester, 2015; Wang et al., 2017; 2018).

For example, researchers have investigated a scenario variation tool that leverages genetic algorithms with novelty search, for scenario generation (Folsom-Kovarik & Brawner, 2018). This tool iteratively searches a space of prospective solutions through mutation or crossover operations with a particular focus on novelty and can generate a range of training scenario events with varying levels of instructional support and difficulty in challenges, which can then be dynamically tailored based on individual trainees' and a group of trainees' needs.

As another example, deep generative models have been utilized for procedural content generation (PCG), which can provide a framework for generating scenarios that meet instructor-specified objectives, while significantly reducing development costs (Awiszus et al., 2021; Luo et al., 2016; Park et al., 2019). Our previous work has investigated PCG based on a multistep deep convolutional generative adversarial network (DCGAN), a type of deep generative model, to create novel educational game levels (Park et al., 2019). Findings suggested that with only a small reduction in the novelty of the generated levels, the resulting multistep generator exhibits significantly enhanced performance by generating a higher percentage of solvable levels compared to the generator trained only on human-authored levels. Deep generative models and conditional variants of those (Torrado et al., 2020) hold potential to dynamically create novel, adaptable scenarios by having the generated scenarios conditioned on individual trainees' competencies as well as a group of trainees' competencies.

Reinforcement learning (RL) has emerged as a particularly powerful form of machine learning that has direct applicability to problems framed in terms of sequential decision making, including automated scenario generation tasks. Competency-based scenario generation can be formalized as a RL task by conceptualizing a scenario generator as an agent that focuses on adapting key dimensions of an exemplar training scenario to achieve instructor-specified objectives for training over time. Decisions about how to adapt different elements of a training scenario (e.g., terrain, unit location, unit behavior, time of day, mission objective) are each modeled as a Markov decision process (MDP). The MDP's state is encoded as a feature vector that summarizes the learner's current state, or in the case of dynamic scenario adaptation, the history of the learner's interaction with the generated scenario thus far. Actions represent the set of possible adaptations the generator can enact to augment a particular dimension of the exemplar scenario. The reward function encapsulates measures of trainee performance that the scenario generator seeks to optimize. The solution to an RL-based scenario generation problem is a policy, or mapping between states and actions, that governs how the scenario generator produces new scenarios that differ from the selected exemplar scenario. RL provides a systematic process for automated scenario adaptation, gradually improving its policy over time as more trainees interact with the scenario generator. Ideally, RL-based scenario generation models are induced using data from learner interactions with scenarios in a simulation-based training environment. However, synthetic data can also be utilized to bootstrap initial investigation into the particular RL formalization of a scenario generation model, including the state representation, action set, and reward model that have been chosen to formalize scenario generation decisions (Wang et al., 2018).

Key to RL-based scenario generation is a *scenario adaptation library*, which enumerates potential transformations that can be applied to a "parent" or exemplar scenario to generate different "child" scenarios. For a given exemplar scenario, this includes determining what types of elements can be adapted, how those elements can be adapted, and when the elements can be adapted to produce a new scenario that is qualitatively different while still aligned with the learning objectives of the starting scenario and the learning domain. Using the Scenario Adaptation Library, RL-based scenario generation can produce a wide variety of training scenarios that can then be deployed with trainees, as well as simulated learners, to evaluate generated scenarios' effectiveness with respect to performance outcomes. These evaluations are then used to refine and improve the model.

In RL terminology, scenario adaptations correspond to the actions in an MDP. They cumulatively define the space of possible generatable scenarios to address instructor-provided learning objectives. In previous

work, we investigated a deep RL framework for personalizing problem-solving scenarios in a narrative-centered learning environment for middle school science education called Crystal Island (Wang et al., 2018). Specifically, we utilized policy gradient RL methods to induce scenario adaptation policies for controlling learning-related adaptable events related to non-player character behavior, pedagogical feedback, and embedded assessments. The scenario adaptation policies were trained using synthetic data from a bipartite player simulation model trained on player action sequences and player outcomes. Results suggested that properly configured deep RL-based narrative planners can significantly outperform linear RL-based interactive narrative planning techniques. Notably, this work focused on generating narrative adaptation policies to support individual students' learning in a middle school science context. AI-driven competency-based team scenario generation will enable team training systems to dynamically craft synthetic learning experiences that support optimal team training and dynamically respond to a broad range of team competencies in institutional training settings in the US Army.

AI-Driven Team Assessment

Because assessment is key to effective training, creating AI-driven team assessment frameworks holds considerable promise for diagnosing a team's strengths and deficiencies and prescribing coaching, feedback, and remediation. Traditional approaches for team assessment have relied on administering external assessments, requesting individuals to provide self-reports, or having instructors rate team performance using checklists. Expanding these methods and leveraging advances in AI-driven assessment methods can significantly increase insight into the actions and behaviors that influence and impact team performance. For example, stealth assessment frameworks that integrate authentic problem-solving scenarios in synthetic training scenarios hold significant promise for unobtrusively measuring teamwork competencies (Min et al., 2020; Shute et al., 2021). Stealth assessment techniques have shown promising results for unobtrusively measuring students' problem solving (Zhao et al., 2015), creativity (Shute & Rahimi, 2021), and computational thinking skills (Min et al., 2020).

Stealth assessment frameworks are rooted in evidence-centered design (ECD), which offers a methodological framework to assess learners' focal knowledge, skills, and abilities by analyzing data drawn from learners' interactions with a training and learning environment (Mislevy et al., 2003). An important step in ECD is evidence modeling, which focuses on identifying behaviors and actions learners take within a simulated environment that can be used to infer competencies of interest. Although the design of evidence rules and statistical models is often created through the collaborative work of domain experts and assessment designers (Mislevy & Riconscente, 2011), more recent work has investigated data-driven approaches to automatically devising evidence models using machine learning techniques such as deep neural networks (Min et al., 2020), hybrid methods that effectively leverage predictive capacity yielded by a range of stealth assessment models (Henderson et al., 2020), and generative zero-shot learning (Henderson et al., 2022). For instance, Henderson et al. (2022) utilized a generative zero-shot learning approach to generalize stealth assessment models for new domains in a game-based learning environment. Results indicated that the zero-shot learning approach was able to effectively model competency states even for unseen levels and scenarios in the game for which no prior data and competency labels existed. These results highlight the promise of using machine learning techniques to support improved prediction of student competency.

A critical step to devising a robust stealth assessment framework for team training is to produce fine-grained evidence about actions and behaviors enacted by team members during team training events or exercises that can be utilized to accurately assess team performance. Team communication data, team movement patterns, and individual team-member actions and sequences can be gathered during synthetic training events and utilized to develop stealth assessment models for team competency and skill development. Team science researchers have made important advancements in formulating unobtrusive measures in simulation-

based training to support assessment of team coordination, back up behaviors, and team communication. (Decostanza et al., 2018; Folsom-Kovarik & Sinatra, 2020; Gilbert et al., 2018; McCormack et al., 2018; Spain et al., 2021).

Because many teamwork and team decision-making behaviors can be assessed by monitoring a team's verbal communication, a significant task for the team training research community is to develop natural language recognition and processing capabilities that can be included in a team stealth assessment framework to automatically assess team performance. Over the past few years, we have been devising natural language processing methods to analyze team communication, such as those employed by the Team Communication Analysis Toolkit (TCAT), which automatically analyzes team communication data, categorizes it into dialogue classifications schemes, and provides summary statistics of critical team communication features (Spain et al., 2022). In recent TCAT work, we created a multi-party dialogue analysis framework with conditional random fields and deep learning models to analyze speaker intent and team communication directional flow (Min et al., 2021). By analyzing team discourse during training episodes from live capstone training exercises, TCAT models capture key characteristics of team dialogue communication that can inform stealth assessment of team competencies such as information exchange, closed-loop communication, and backup behaviors.

Additional natural language processing-based approaches are continuing to be investigated by the team tutoring research community to identify predictors of high and low performing teams, to identify when critical incidents occur by analyzing team member speech, and to predict team performance (Foltz, 2018). For example, McCormack et al., (2020) explored methods to assess team cohesion by analyzing inclusive vs. exclusive language (e.g., we, us, our vs. I, me, mine) captured from team speech recordings during synthetic training events. More recently, Folsom-Kovarik et al. (2022) developed an intent recognition and speech classification system that automatically analyzes team communication data to identify markers of information exchange, communication quality, and concise communication. Given the critical role communication plays in effective teams, a particularly promising approach to creating computational models of teamwork assessment is integrating analysis of spoken team communication (spoken language dialogue models including prosodic analysis), semantic analysis that extracts both utterance-level and dialogue-level semantics, and analysis of behavioral trace data captured from training simulation systems to provide a granular multi-dimensional account of team communication.

AI-driven team assessment will thus be able to assess team communication content, quality, and information exchange features, and provide insights into team processes and cognitive states that will be used to dynamically inform team tutoring policies in intelligent tutoring systems for teams.

Conclusions and Recommendations for GIFT

With increasingly powerful AI technologies spanning natural language processing, computer vision, and machine learning, AI-driven team training systems will soon be able to provide team training capabilities including 1) real-time team training scenario generation that leverages emerging machine learning-based frameworks to deliver customized data-driven training scenarios that are tailored to individual teams to develop robust team competencies, and 2) AI-driven team assessment that utilizes stealth assessment frameworks leveraging natural language processing and machine learning to dynamically and reliably evaluate team competencies. Further, as computer vision continues to make significant advances, particularly in human activity recognition, team assessment frameworks will become increasingly multimodal and integrate computer vision into team assessment pipelines to support integration of team activity and spoken team communication data for evaluating team performance.

Given these developments, we recommend that future work on team training in the Generalized Intelligent Framework for Tutoring (GIFT) support three sets of capabilities. First, GIFT should provide scenario generation functionalities for automatically generating team training scenarios. This will entail supporting both team interaction data collection (which will be used to train scenario generation policies), machine learning capabilities (including reinforcement learning) to machine-learn the generation policies, and integration with synthetic training environments. Second, GIFT should provide team training assessment capabilities. These should support team competency diagnostics that are actionable and reliable. This will entail supporting student model representations and inference methods that utilize state-of-the-art natural language processing, computer vision, and machine learning methods to effectively distinguish between individual and team-level competencies. Finally, GIFT should provide team feedback capabilities that seamlessly integrate with the scenario generation and team assessment capabilities. This will entail enabling team feedback to be driven by real-time multi-dimensional team assessment, with feedback being "narratively embedded" in the generated team training scenarios playing out in synthetic training environments.

Acknowledgments

The research described herein has been sponsored by DEVCOM Soldier Center under cooperative agreement W912CG-19-2-0001. The statements and opinions expressed in this chapter do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Awiszus, M., Schubert, F., & Rosenhahn, B. (2021, August). World-GAN: a generative model for Minecraft worlds. In *2021 IEEE Conference on Games* (pp. 1-8). IEEE.

DeCostanza, A. H., Gamble, K. R., Estrada, A. X., & Orvis, K. L. (2018). Team measurement: Unobtrusive strategies for intelligent tutoring systems. In *Research on Managing Groups and Teams: Vol. 19. Building Intelligent Tutoring Systems for Teams* (pp. 101–130).

Folsom-Kovarik, J.T., & Brawner, K. (2018). Automating variation in training content for domain-general pedagogical tailoring. In *Proceedings of the Sixth Annual GIFT User Symposium* (pp. 75-86). U.S. Army Research Laboratory.

Folsom-Kovarik, J. T., Roque, A., & Sinatra, A. M. (2022). Addressing team process with automated speech act assessments. In *Proceedings of the Tenth Annual GIFT User Symposium* (pp. 139-146). US Army Combat Capabilities Development Command–Soldier Center.

Folsom-Kovarik, J. T., Rowe, J., Brawner, K., & Lester, J. (2019). Toward automated scenario generation with GIFT. *Design Recommendations for Intelligent Tutoring Systems: Volume 7 - Self-Improving Systems*, 109-118.

Folsom-Kovarik, J. T., & Sinatra, A. M. (2020). Automating assessment and feedback for teamwork to operationalize team functional resilience. In *Proceedings of the Eighth Annual GIFT Users Symposium* (pp. 126-135). US Army Combat Capabilities Development Command–Soldier Center.

Foltz, P. W. (2018). Automating the assessment of team collaboration through communication analysis. *Design Recommendations for Intelligent Tutoring Systems: Volume 6 - Team Tutoring*, 179-186.

Gilbert, S. B., Slavina, A., Dorneich, M. C., Sinatra, A. M., Bonner, D., Johnston, J., ... & Winer, E. (2018). Creating a team tutor using GIFT. *International Journal of Artificial Intelligence in Education*, *28*(2), 286-313.

Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., ... & Lester, J. (2022). Enhancing stealth assessment in game-based learning environments with generative zero-shot learning. In *Proceedings of the Fifteenth International Conference on Educational Data Mining* (pp. 171-182). International Educational Data Mining Society.

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid

stealth assessment framework. In *Proceedings of the Thirteenth International Conference on Educational Data Mining* (pp. 92-103). International Educational Data Mining Society.

Johnston, J. H., Burke, C. S., Milham, L. A., Ross, W. M., & Salas, E. (2018). Challenges and propositions for developing effective team training with adaptive tutors. In Joan Johnston, Robert Sottilare, Anne M. Sinatra, C. Shawn Burke (ed.), *Building Intelligent Tutoring Systems for Teams* (pp. 75-97). Emerald Publishing Limited.

Lee, S. Y., Rowe, J., Mott, B., & Lester, J.C. (2014). A supervised learning framework for modeling director agent strategies in educational interactive narrative. *IEEE Transactions on Computational Intelligence and AI in Games*, 6, 203-215.

Luo, L., Yin, H., Cai, W., Zhong, J., & Lees, M. (2016). Design and evaluation of a data-driven scenario generation framework for game-based training. *IEEE Transactions on Computational Intelligence and AI in Games*, *9*(3), 213-226.

McCormack, R. K., Kilcullen, T., Sinatra, A. M., Brown, T., & Beaubien, J. M. (2018). Scenarios for training teamwork skills in virtual environments with GIFT. In *Proceedings of the Sixth Annual GIFT Users Symposium* (pp. 189-198). US Army Research Laboratory.

McCormack, R., Case, A., Howard, D., Logue, J., Kay, K., & Sinatra, A. M. (2020). Teamwork training in GIFT: Updates on measurement and audio analysis. In *Proceedings of the Eighth Annual GIFT Users Symposium* (pp. 155-162). US Army Combat Capabilities Development Command–Soldier Center.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E.,& Lester, J. C. (2020). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325.

Min, W., Spain, R., Saville, J. D., Mott, B., Brawner, K., Johnston, J., & Lester, J. (2021). Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In *Proceedings of International Conference on Artificial Intelligence in Education* (pp. 293–305). Springer, Cham.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 75-104). Routledge.

Park, K., Mott, B. W., Min, W., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2019). Generating educational game levels with multistep deep convolutional generative adversarial networks. In *Proceedings of the 2019 IEEE Conference on Games* (pp. 345-352). IEEE.

Rowe, J., & Lester, J. (2015). Improving student problem solving in narrative centered learning environments: A modular reinforcement learning framework. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence in Education* (pp. 419-428). Springer, Cham.

Salas, E., Bowers, C.A., & Cannon-Bowers, J.A. (1995). Military team research: Ten years of progress. *Military Psychology, 7,* 55-75.

Salas, E., Cannon-Bowers, J. A., & Johnston, J. H. (1997). How can you turn a team of experts into an expert team?: Emerging training strategies. *Naturalistic Decision Making, 1*, 359-370.

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors, 50(3)*, 540-547.

Salas, E., Sims, D.E. & Burke, C.S. (2005). Is there a "big five" in teamwork? *Small Group Research*, *36*(5), 555–599.

Salas, E., Tannenbaum, S. I., Kozlowski, S. W., Miller, C. A., Mathieu, J. E., & Vessey, W. B. (2015). Teams in space exploration: A new frontier for the science of team effectiveness. *Current Directions in Psychological Science*, *24*(3), 200-207.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141.

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998). Measuring team-related expertise in complex environments. *Making decisions under stress: Implications for individual and team training*, *1*, 61-87.

Sottilare, R., Barr, A., Robson, R., Hu, X., & Graesser, A. (2018). Exploring the opportunities and benefits of standards for Adaptive Instructional Systems (AISs). In *Proceedings of the Adaptive Instructional Systems*

Workshop in the Industry Track of the 14th International Intelligent Tutoring Systems (ITS) Conference (pp. 49-53).

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education*, *28*(2), 225–264.

Spain, R., Min, W., Saville, J., Brawner, K., Mott, B., & Lester, J. (2021). Automated assessment of teamwork competencies using evidence-centered design-based natural language processing approach. In *Proceedings of the Ninth Annual GIFT Users Symposium* (pp. 140–149). US Army Combat Capabilities Development Command–Soldier Center.

Spain, R., Min, W., Saville, J., Emerson, A., Pande, J., Brawner, K., & Lester, J. (2022). Leveraging advances in natural language processing to support team communication analytics in GIFT. In *Proceedings of the Tenth Annual GIFT Users Symposium* (pp. 147–156). US Army Combat Capabilities Development Command–Soldier Center.

Torrado, R. R., Khalifa, A., Green, M. C., Justesen, N., Risi, S., & Togelius, J. (2020). Bootstrapping conditional gans for video game level generation. In *2020 IEEE Conference on Games* (pp. 41-48). IEEE.

TRADOC (2017). *The U.S. Army Learning Concept for Training and Education: 2020-2040*. Retrieved from: https://adminpubs.tradoc.army.mil/pamphlets/TP525-8-2.pdf

U.S. Army, (2012). *Army Doctrine Reference Publication (ADRP) 6-0, Mission Command*. Washington, DC: Headquarters, Department of the Army.

Wang, P., Rowe, J., Min, W., Mott, B., & Lester, J. (2017). Interactive narrative personalization with deep reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 3852-3858).

Wang, P., Rowe, J., Min, W., Mott, B., & Lester, J. (2018). High-fidelity simulated players for interactive narrative planning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 3884-3890).

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & Perrault, R. (2021). The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

Zhao, W., Shute, V., & Wang, L. (2015). Stealth assessment of problem-solving skills from gameplay. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

CHAPTER 4 – OPTIMIZING INTELLIGENT TUTORING SYSTEM DESIGN FOR PROFESSIONAL DEVELOPMENT

Robert A. Sottilare
Soar Technology, Inc.

Introduction

Professional development involves the development of credentials (e.g., degrees, certifications) that indicate a level of proficiency for individuals in an occupational field relative to their peers or a professional standard (Speck & Knipe, 2005). Professionals in various fields have common sets of knowledge and skills that are expected to be acquired to perform at various levels of effectiveness. In earning and maintaining essential knowledge and skills, professionals seek to achieve levels of certification through academic experiences (formal coursework), technical conference participation, training, and informal learning (e.g., reading and viewing instructional videos) that are situated to the field of practice (e.g., medicine or engineering). Professional development experiences include collaborative development, apprenticeship, and individual initiatives for self-improvement and growth.

While informal learning activities have been shown to be valuable in expanding knowledge (Le Clus, 2011), an essential element of the professional development certification process is evidence-based assessment (Hunsley & Mash, 2007). In the United States, licensing is provided to qualified candidates who have demonstrated proficiency through testing and experience standards established by a professional board (e.g., Florida Board of Professional Engineers). Intelligent Tutoring Systems (ITSs) are gaining traction as developmental tools leading to certification or professional licensing. *How might ITS research and emerging capabilities influence the process of professional development?*

In general, ITS research is an important stream of investment that consistently produces new methods, concepts, and prototypes that improve ITS assessment, recommendations, and interventions. While there are many approaches to assessing learning, new methods are being developed to improve the timeliness and accuracy of learning assessments. While new ITS technologies emerge regularly, new concepts for optimizing the selection of ITS interventions for learners continue to accelerate learning. As documented in the literature, no ITS technology has a greater impact on learning than artificial intelligence and machine learning (AI/ML) tools and methods which influence the accuracy of instructional assessments, the relevance of recommendations and interventions, and the continuous improvement of ITS policies.

Our research over the last four years has been focused on exploiting AI/ML techniques to enhance ITS processes to accurately predict learner states (e.g., performance, proficiency, emotions) that influence learning efficiency and effectiveness, recognize learner and team readiness, rapidly develop and test intelligent agents that support simulation-based training, improve the engagement of learner interface design and optimize the selection of tutor interventions.

At SoarTech, our technical approaches have included causal modeling with a focus on understanding root causes of learning outcomes by testing and eliminating potential root causes in simulation-based training, innovation of methods to recognize events or trends, and construction of data pipelines leading to more accurate ensemble (compound) machine learning models that include neural network, clustering, and decision-tree solutions. We have applied AI/ML methods to guide subject matter experts (SMEs) in identifying and associating appropriate interventions in context with learner and simulated environmental conditions for both real-time feedback and after-action review (AAR) processes.

ITSs for professional development might differ from ITSs for other purposes or instructional domains in the way that skills developed during training transfer to conduct tasks required every day in the operational environment. In other words, we believe there is a higher emphasis on the effectiveness and efficiency in which knowledge is acquired and applied on a regular basis at work. For this reason, we developed this chapter to concentrate on the salient characteristics of ITSs and how they might be applied to the professional development space. The assumption is that efficiency is coveted in a professional development environment and that discovery learning is not a practical alternative.

In centering our narrative on the salient characteristics of ITSs, our goal is to more easily identify opportunities to apply AI/ML to automate ITS processes and reduce ITS author workload while also making ITSs more practical solutions in the professional development domain. Previously, Sottilare and Gilbert (2011) identified five salient ITS characteristics. Our goal in this chapter is to extend this narrative to consider some unique aspects of ITSs for professional development sponsored by large organizations where a disciplined process fosters a pipeline of professional skills across the organizational structure. We discuss this goal in more detail in the next section of this chapter.

Goals and Scope

Given our primary goal of identifying the salient characteristics of a functional ITS, we selected four attributes to examine. First, ITSs for professional development must be validated as ***effective and credible*** in teaching operational tasks and concepts so that they are considered useful by the organization, the instructors and the learners. ITSs and their content must also be relevant and focused on learning and growth to promote the development of essential skills in a professional curriculum. The second characteristic, ***affordability***, illustrates the need to efficiently guide learners to domain proficiency with a reasonable return-on-investment (ROI) for the sponsoring organization. The third salient characteristic, ***engagement*** must enable the system to build rapport with learners, so they continue to be ITS users and come back again and again to learn new content or refresh prior knowledge. Finally, given the context of professional development, any ITS applied in this domain must be ***easily accessible*** to offer flexibility and accommodate the busy schedules of learners with day jobs.

A major challenge for ITSs that support professional development is that the topics of instruction must be specific enough to support the development of useful operational skills. In addressing this challenge, an ITS for professional development should consider how each of these salient characteristics will be shaped during the design process and how they will influence deployment of instruction to trainees, maintenance by learning engineers and instructional designers, and evaluations. Another goal is for these systems to be self-improving in that they consider outcomes in the policy maintenance process. ITS designers should identify and consider requirements for current and future training needs to craft individualized training and development plans. Finally, learning engineers should consider how AI/ML methods support accurate prediction of learner states, recognition of events and trends, development of recommendations, and optimal selection of interventions. Given the close coupling of learners and ITSs, many ITS processes must be able to support real-time interaction. Next, we examine how salient characteristics posed in this chapter might influence the quality of professional development.

State of the Field and Supporting Research

In this section, we begin to scrutinize the literature and evaluate the influence of ITS salient characteristics with respect to learning outcomes (e.g., knowledge and skill acquisition, transfer of learning). We also compare outcomes to those of expert human tutors as a baseline for evaluating the effectiveness, credibility, affordability, engagement, and accessibility of ITS designs for professional development. Revisiting the

question posed in our introduction, *how might ITS research and emerging capabilities influence the process of professional development*, we specifically examine how ITS research and development principles influence the five salient characteristics posed above.

ITS Effectiveness and Credibility

In reviewing the literature, VanLehn (2011) noted that the relative effectiveness of ITSs and expert human tutors were roughly equal in STEM domains - science (e.g., physics), technology (e.g., computer programming), engineering (e.g., strength of materials), and mathematics (e.g., calculus, trigonometry and algebra). Since 2011, ITSs have been more broadly applied to military occupational training that includes the learning of cognitive, psychomotor and team-based tasks (Sinatra, 2022). Instructional models like INSPIRE (Lepper & Woolverton, 2002) highlight the importance of tutors that are intelligent, nurturing, Socratic, progressive, indirect, reflective, and encouraging. While we agree that the INSPIRE model's salient characteristics are important considerations in ITS design, we would also add that they are largely dependent on the accuracy of learner assessments and the quality and completeness of domain content. It would be difficult to produce an effective tutoring session (human-led or machine-based) without an accurate model of the learner's progress toward learning objectives. To illustrate this point, we examine the tutoring process described by Graesser et al. (1995) and Person et al. (2003) which describes five steps in which the learner model including learner states and goals are a central element:

1. Interact with the learner and update the learner model as changes in learner states (performance, emotions) occur
2. Select instructional strategies (recommendations) based on learner states and learning science principles
3. Select instructional tactics (actions) based on learning science principles (policies) and context
4. Repeat steps 1-3 until learning goals are achieved
5. Document achievements for use in subsequent instructional experiences to improve the learner and instructional models

Just as knowledgeable, trustworthy, and reliable are important characteristics for human tutors, they may be even more important to the design and effectiveness of credible machine-based tutors. ITS designs that have incomplete or incorrect domain knowledge are often quickly dismissed by learners. The ITS knowledge base should be as complete as possible and include sufficient content, subject matter expertise, and knowledge of the learner to support assessments and provide relevant interventions (e.g., feedback, support, direction, reflections). Credibility can be enhanced over time by providing a mechanism for users (e.g., learners, training coordinators) to flag missing or incorrect knowledge. A last, but important point, is that ITS interface design should facilitate learning and not detract from it (Corbett et al., 1997).

ITS Affordability

Given a professional development context, large organizations should focus on the development of essential skills in an efficient manner. A mix of formal training, informal learning, and job experience should be evaluated to efficiently guide learners to various domain proficiency levels based on their role/position and level of responsibility. From an affordability perspective, it might not make investment sense to attempt to train a novice to the proficiency level of an expert. Each organization should consider the investment of time needed to examine the transfer of training skills to operational/work tasks.

Another consideration in developing an efficient and affordable adaptive training program is the cost of curating/managing content, and building, deploying and maintaining essential ITSs. Content curation is

"the process of gathering information relevant to a particular topic or area of interest, usually with the intention of adding value through the process of selecting, organizing, and looking after the items in a collection or exhibition" (Wikipedia, 2022). Typically, content curation requires highly skilled instructional designers, and ITS development requires highly skilled computer scientists with advanced degrees (Choksey, 2004).

To support affordability goals, organizations should evaluate open source curation tools with user friendly interfaces along with open source ITS development toolsets such as the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare et al., 2012; Sottilare et al., 2017). Every organization should independently evaluate their risk and reward for developing/purchasing adaptive training capabilities and consider the best technology in the marketplace to support their goals. We also recommend a review of adaptive instructional standards and recommended practices to understand the costs and benefits of adoption (Sottilare, 2022).

ITS Engagement

Engagement "refers to the degree of attention, curiosity, interest, optimism, and passion that students show when they are learning or being taught, which extends to the level of motivation they have to learn and progress in their education" (Kalyani & Rajasekaran, 2018, p. 23). Regardless of whether we are discussing classroom, online or adaptive instruction, engagement is a critical element of instructional effectiveness. There are many principles associated with engagement and motivation during instruction (e.g., setting clear goals and objectives), but establishing competition and rewards may be the most important. Competition can be established through the implementation of performance standards, badges and certifications associated with training achievements and skill development. Similarly, rewards can be established with various levels of achievement.

ITS Accessibility

Finally, our fourth salient characteristic, accessibility, is needed to provide high availability to content developers, course authors, learners, and other users with demanding jobs and responsibilities. Designing ITSs for accessibility should include user availability to ITS tools on a variety of platforms (e.g., smartphones, tablets, workstations) on both internal networks and remotely. Tools should be available to collaboratively curate content and construct ITSs. Learner records, including group descriptive statistics, should be available to organizational leaders, group supervisors, and learners to assess organizational learning and job readiness. Data collected during training events should also be stored to support construction of learner models (e.g., event performance, competency/proficiency). ITS assessments, recommendations, and instructional decisions should also be recorded and stored for later analysis to support a self-improving ITS learning landscape.

Recommendations for GIFT and Intelligent Tutoring Systems

Based on our discussion of ITS desirable salient characteristics, we provide three key design recommendations for implementing an ITS-based learning landscape for a large organization (e.g., military or large corporation). Our first design recommendation is to create a tool for organizations to assess the cost-benefit and ROI associated with implementing adaptive instruction versus one-size-fits-all online instruction. A well-informed cost-benefit tool would enable organizations to assess the merits of future adaptive training investments. A second ITS design recommendation is to enhance rapport in ITSs through the integration of virtual humans (VHs) as both tutor and peer learner (Graesser et al., 2017). This recommendation should include reconfiguration of VHs to support learner preferences. This recommendation is also tied to our third research recommendation below. Our third design recommendation

is a general call to action to identify opportunities that encourage competition and provide rewards for achievement in areas of organizational need. Rewards can include certifications, bonuses, free lunches or time off.

We also recommend future research to address three major areas of need to enable more effective and efficient ITS-based training. Our first research recommendation is to analyze and create methods to rapidly evaluate the credibility of ITSs by comparing required and available domain knowledge. This research will aid ITS developers in projecting the effectiveness and reliability of their newly created ITSs. Our second research recommendation is to identify mechanisms to distinguish knowledge performance from interface performance. This research would enable ITS developers to assess the influence of interface familiarity in ITSs. For example, a learner with poor interface skills, but high domain knowledge would likely still perform poorly. Our third research recommendation is to develop methods to reduce the authoring skills/time/cost required to integrate VHs with simulation-based training environments (Sottilare et al., 2022). Our team has been evaluating data driven methods that are independent of the simulation environment and the assessment capability (e.g., GIFT).

Conclusions

We close out our discussion on ways to optimize ITS design for professional development with a few obvious conclusions that still merit space in this chapter. ITSs are data driven systems that can be greatly influenced by AI/ML methods that provide insight into learners and the learning process. It is important for ITS designers and researchers to develop a data pipeline through data cleaning and feature engineering processes that enable insights into individual learner states, trends, and events.

Culture is an overused and misunderstood concept, but it is critical to instill an organizational culture that promotes learning and considers the effectiveness of instructional systems in transferring skills from training to operations in the workplace.

References

Choksey, S. D. (2004). *Developing an affordable authoring tool for intelligent tutoring systems* (Doctoral dissertation, Worcester Polytechnic Institute).

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In *Handbook of human-computer interaction* (pp. 849-874). North-Holland.

Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, *76*, 607-616.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, *9*(6), 495-522.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annu. Rev. Clin. Psychol.*, 3, 29-51.

Kalyani, D., & Rajasekaran, K. (2018). Innovative teaching and learning. *Journal of Applied and Advanced Research*, *3*(1), 23-25.

Le Clus, M. (2011). Informal learning in the workplace: A review of the literature. *Australian Journal of Adult Learning*, *51*(2), 355-373.

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In *Improving Academic Achievement* (pp. 135-158). Academic Press.

Person, N. K., Graesser, A. C., Kreuz, R. J., & Pomeroy, V. (2003). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education* (IJAIED), *12*, 23-39.

Sinatra, A. M. (2022, May). The 2022 Instructor's Guide to the Generalized Intelligent Framework for Tutoring (GIFT). In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10)* (p. 21).

Sottilare, R. (2022). Analyzing the Motivation for Adaptive Instructional System (AIS) Standards. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2022, Orlando, Florida.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT). *Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED)*.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT). GIFTtutoring. org, 1-19.

Sottilare, R., & Gilbert, S. (2011). Considerations for tutoring, cognitive modeling, authoring and interaction design in serious games. In Authoring Simulation and Game-based Intelligent Tutoring workshop at the *Artificial Intelligence in Education Conference (AIED)*.

Sottilare, R., Woods, A., Giranda, N., Bertrand, M., Ortiz, E. & Friedman, B. (2022). Facilitating the Integration of Virtual Humans within GIFT. 10th Annual GIFT Users Symposium, Orlando, Florida.

Speck, M., & Knipe, C. (Eds.). (2005). Why Can′ t We Get It Right?: Designing High-Quality Professional Development for Standards-Based Schools. Corwin Press.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.

CHAPTER 5 – ALIGNING TRAINING WITH DESIRED SKILLS: THE OUTER LOOP FOR UPSKILLING

Robby Robson, Elaine Kelsey, Lauren Egerton, Sazzad Nasir, and Kari Glover
Eduworks Corporation

Introduction

The context for this chapter is the *SkillSync*™ project that is part of US National Science Foundation (NSF) Convergence Accelerator (National Science Foundation, n.d.) track on *AI and the Future of Work*. The Skillsync project has developed a web application whose purpose is to connect companies to college professional development programs and other training providers that can offer efficient, equitable, and effective upskilling opportunities to incumbent workers. This app is supported by a set of AI (Artificial Intelligence) services that are more broadly applicable and are available independently of SkillSync. More detailed descriptions of SkillSync, the SkillSync app, and its supporting AI services can be found in additional publications (Lisle & Robson, 2021; Molnar et al., 2022; Robson, Kelsey, Goel, Egerton et al., 2022; Robson, Kelsey, Goel, Rugaber et al., 2022).

Of particular interest in this chapter is an AI-enabled *alignment service* that computes the degree to which a set of course materials cover a prioritized list of desired skills, represented in SkillSync as a number between 0 and 100 called an *alignment score.* In the SkillSync app, the alignment score is used by training providers to formulate and propose a program of instruction that meets the upskilling needs requested by a company. Looked at more generally, the alignment score can be viewed as a tool for determining what content will fill the largest number of learner skill gaps and for identifying when course content might be combined or reconfigured to improve efficiency. As such, we posit that it can be used to manage the outer loop (VanLehn, 2006) of intelligent tutoring systems (ITSs) used in professional development. This is discussed in the recommendations section of this chapter.

Much of this chapter is devoted to explaining how the SkillSync alignment service was developed. Many of these techniques can be applied to other AI-enabled services used in professional development, training, and workforce applications and to ITSs used for these purposes. These potential applications are also discussed in the recommendation section.

Goals and Scope

As the nature of work evolves at an increasing pace, upskilling of incumbent workers becomes an urgent imperative for many employers. Job roles change rapidly, due to changes in demand and new product requirements. Skills that did not exist until recently are needed urgently, and labor market constraints require that much of this new capacity be generated from an employer's existing workforce. Employees are required or desire to develop new capabilities to meet this challenge. Rapid upskilling of the incumbent workforce poses several challenges. Among these are the needs to quickly identify appropriate training resources and to balance the needs of employers for rapid and cost-efficient training with employee preferences for training that includes industry-recognized courses and credentials.

For this reason, courses or materials drawn from pre-existing training and education programs (e.g. community college courses) are a potentially rich source of upskilling training. However, several challenges occur when employers try to access these courses and materials for upskilling of incumbent workers. Specifically, descriptions of training content are often geared toward students or academic audiences, and are almost never tagged with metadata to indicate which workplace skills they train. While competencies and skills may be inferred from reviewing training materials, these are rarely stated explicitly in the language and taxonomies that employers are familiar with. This disconnect in language and frameworks means that matching incumbent workers to appropriate courses for upskilling can involve significant trial and error.

The SkillSync alignment service is designed to bridge this gap, using a set of AI techniques to effectively search course descriptions and content for a set of specific skills required for upskilling incumbent workers. The service is designed to translate between the language and mental models of employers and academic training providers, using a multi-dimensional variant of semantic search to quickly identify the subset of courses and training opportunities that provides the best upskilling match with the least training.

The SkillSync Alignment Service

The SkillSync alignment service takes as input two sets of data: Skills and training opportunities. In this context the term "skills" is used generically and could represent desired or existing knowledge, skills, or abilities as well as job tasks. The skills are the training objectives that are to be provided by training opportunities. As is the case in most professional development and upskilling settings, these opportunities are usually "courses" ranging in length from an hour to the equivalent of a one-term or one-semester course and are not degree programs.

Skills are stored in machine-actionable formats in the Competency and Skills System (CaSS) (ADL, 2020) where they can be retrieved as linked data. The data associated with skills can range from little more than a short description to more detailed descriptions and data that links the skill to other resources and tags skills with elements from concept schemes that define the context of the skill. In SkillSync, metadata also indicates whether it is a skill that can be assumed to be held by a learner (existing skills) or whether it is a skill that needs to be acquired (target skills). For the latter, SkillSync also records its priority on a scale of critical (priority 1), important, desired and nice to have (priority 4).

Training opportunities are also stored in CaSS. The data associated with each training opportunity includes at least a title and short description and may include additional information such as a syllabus, learning outcomes, and catalog information (duration, cost, location, and course materials).

Comparing Training Opportunities to Skills

Given a list of desired skills and a description of a training opportunity (generically called a *course* in this chapter), there are essentially two ways to determine the extent to which a course addresses those skills. The first, which is what an expert human might do, is to examine the description and apply contextual and world knowledge to make a holistic judgment. The second is to identify the skills or learning outcomes associated with the course and to compare these to the desired skills. Associated skills and learning outcomes may be explicitly stated in the course description or may be generated from the description.

Most automated (or semi-automated) approaches use variants of the second method. This has two inherent drawbacks. First, information is likely to be lost by reducing a course to a list of skills or outcomes, including contextual information that is necessary to properly interpret the skills. Second, the list of skills required for a job usually comes from a different taxonomy than the skills or outcomes associated with a course, making it necessary to compare skills from two different taxonomies. Common methods include mapping all skills to a master taxonomy, which facilitates comparisons but introduces a further loss of fidelity, and comparing skills statements based on key words, which is inaccurate. It is also possible to train machine learning (ML) models to make comparisons, and once these are in place, they can be applied to directly compare skills statements to descriptions of training materials, which is what SkillSync does.

SkillSync's Alignment Score

SkillSync's alignment score uses language models that are pre-trained on corpora that include high probability and high frequency multi-word expressions with a given set of domains and that incorporate ontologies with world knowledge. This enables the models to operate on "concepts" rather than terms. The building block of these models are deep neural network (DNN) based transformers (Wolf et al., 2019, 2020) that take sequences of multi-word tokens as input and are trained to predict how the sequence will be completed or when input tokens are randomly masked, with the training objective being to predict masked tokens solely from the surrounding context. The output of these transformers can be viewed as vectors that represent the context and meaning of the input token sequences in a succinct (although non-transparent) way. Thus, given two inputs, say a skills statement and a course description, each model generates two vectors u and v, and distance between u and v serves as a similarity measure.

The pre-trained language models can be fine-tuned to produce different measures (Merchant et al., 2020). SkillSync currently produces four additional measures by stacking a classifier on the top of the transformer models. This classifier learns to weight the components of the vectors differently for different tasks. The fine-tuned models can also be generalized to address questions in other domains of applicability by adopting a "few-shot" approach (Brown et al., 2020) that requires a relatively small amount of labelled data set, e.g., on the order of two-hundred examples. The four measures used in SkillSync (in addition to the basic similarity measure produced by a model trained on a specific domain) are:

1. **Context Similarity:** Rates on a scale of 1 – 5 the degree to which the skill statement and course come from the same domain (including occupation, work setting, industry, and profession).

2. **KSA Similarity:** Rates on a scale of 1 – 5 the similarity between the skill's statement and KSAs (Knowledge, Skills, and Abilities) mentioned in the course description.

3. **Scope:** Tags whether the closest KSA in a course description is (roughly) broader, the same, or narrower than the skill statement (or if there is no overlap at all).

4. **Difficulty:** Tags whether the skill statement is (roughly) more advanced, at the same level, or less advanced than the course (or N/A).

The data provided to labelers includes a single skills statement (which described knowledge, a skill, or an ability, i.e., a KSA) and text that includes a course title and a (short) course description. The labelers are instructed to classify the statement as knowledge, a skill, or an ability and to score the four types of alignment between the statement and course description.

Finally, the basic similarity score, the four additional alignment scores, and the prioritization of desired skills are combined to compute an overall alignment score. The weights given to each component in this

alignment score are determined by trial and error. Figure 1 below shows the high-level steps involved in developing the service that computes this score.

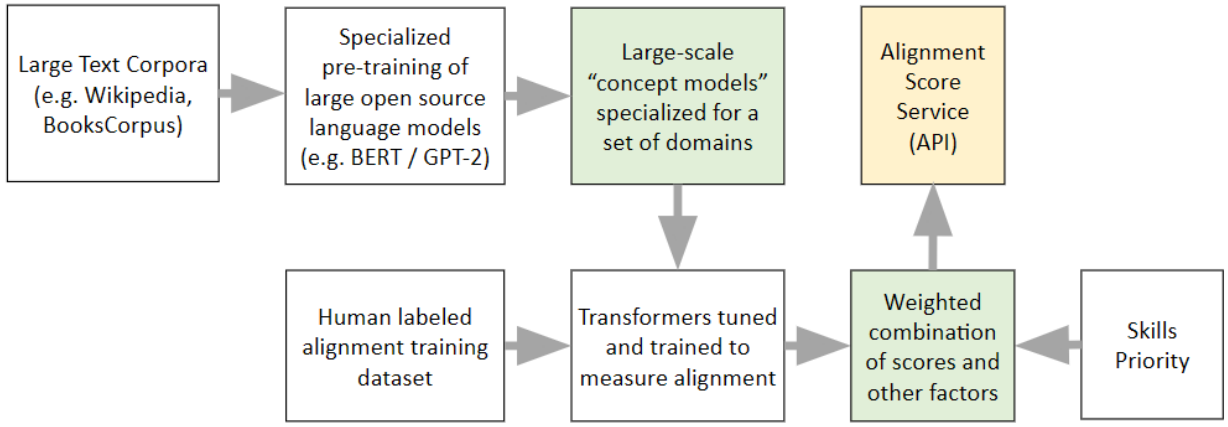


Figure 1. Steps involved in developing the Alignment Score Service

Bias Reduction

It is a goal of the SkillSync project to increase participation by groups that are currently under-served and under-represented in the workforce. An underlying assumption is that this can be achieved in part by focusing on skills rather than formal educational degrees and past work experience, but that is not alone sufficient. Explicit and implicit bias can be introduced into skills-based hiring in many ways, one of which is through language that reflects occupational biases. Such language is often reflected in language models (Bolukbasi et al., 2016; Kirk et al., 2021), so in developing language models the project has taken several steps to reduce such bias. These are not discussed in this chapter, but the topic is raised since we see language models as being increasingly used in ITSs.

Discussion

The Outer Loop for Upskilling

There are two loops in the model of an ITS introduced by VanLehn: An *outer loop* that sequences tutoring activities for each "task" and is macro-adaptive, and an *inner loop* that observes the learner as they step through a solution and is micro-adaptive. As pointed out by VanLehn (VanLehn, 2006, p. 227), "the inner loop can also assess the student's evolving competence and update a student model, which is used by the outer loop to select a next task that is appropriate for the student." This places the state of a learner's skills as part of the learner model used by an ITS and identifies the outer loop as the component of an ITS that is responsible for selecting learning experiences based on a set of desired skills. In this regard, a notional architecture of ITSs used for professional career education is shown in Figure 2. In this variant of the standard model, the "expert model" becomes a "professional model" that contains the skills required for a job or desired for career advancement as well as other relevant data such as credentials and experience. The job of the outer loop is to perform gap analyses, analyze available training opportunities, and make selections.

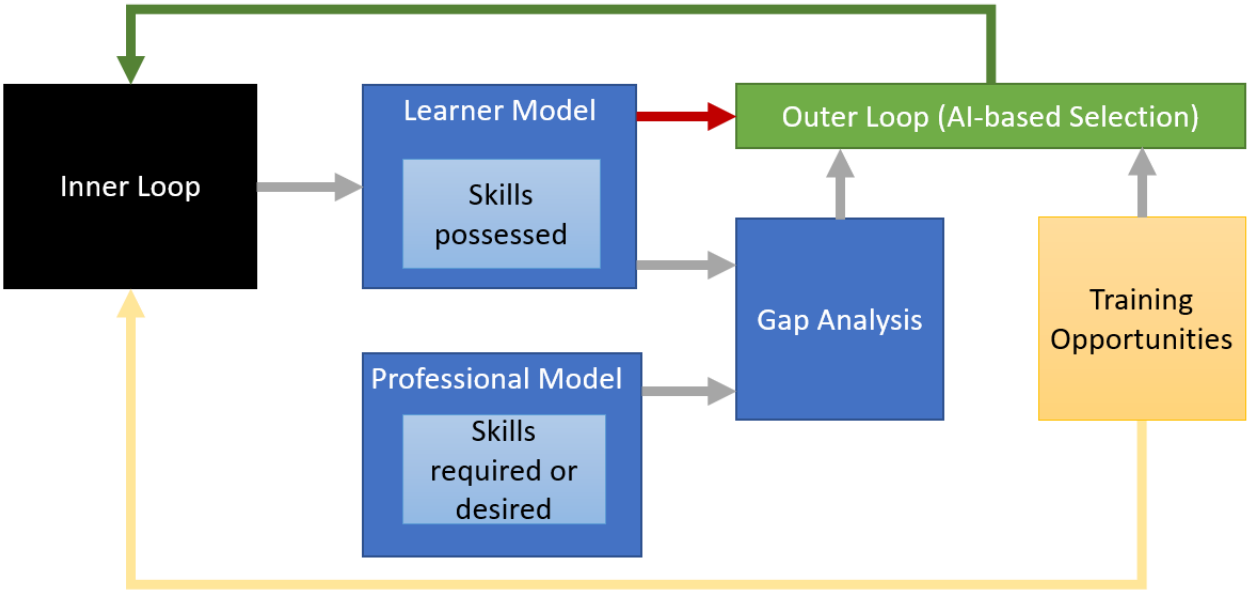


Figure 2. Notional ITS Architecture for Professional Education and Upskilling

In the use case addressed by SkillSync – upskilling incumbent workers – the selection of training opportunities is accomplished by human operators with the aid of machine-generated alignment scores and explicit knowledge of the skills that a cohort of workers can be assumed to possess, but the alignment score and the techniques used to generate it can be modified to select training opportunities within an ITS. The steps required for this purpose are discussed next.

Modifications Required to Implement the Outer Loop for Upskilling

The existing alignment score takes desired skills and priorities into account and can be used to choose a set of training opportunities that provide required skills. This selection is currently aided by a table that shows which skills are "covered" by a course and displays the overall alignment score. Additional metadata, such as the duration of each course, remuneration required (if relevant), course ratings, and time slots are needed to construct a "reward function" that can be used to optimize a selection. In practice, it likely suffices to minimize the overall duration of a set of activities while ensuring maximal coverage.

The drawback of this simplistic approach is that *prerequisites* must also be taken into consideration. To avoid selecting training that is inappropriate for a learner, it is necessary to maintain the current state of learner skills (which is part of the learner model) and to identify the prerequisite skills required by a given course. This problem of prerequisite detection is closely related to the problem of detecting which skills can be acquired through training, and in this case the same underlying language models and fine tuning methods can be applied. As with skills alignment, if prerequisite skills are listed in a course description, they could be compared to learner skills, but there is little point in doing so if the models are good enough. Using pre-trained and tuned language models obviates the need for such explicit listings and enables more complete context-dependent judgments to be made.

Once prerequisite skills are identified, they can be used together with existing skill state estimates in a reward function that can be computed in real time (or near real time) to guide the outer loop of ITSs. This approach uses language models to approximate decisions that would be made by well-informed human tutors, with the potential of reducing the effects of assumptions that can introduce unwanted bias. As an

example of such bias, suppose that an ITS is being developed to put adults on a career pathway in the field of electric vehicles. A human tutor may erroneously feel that skills repairing internal combustion engines are required. Since over 90% of existing auto mechanics are male (Bureau of Labor Statistics, 2022), this requirement introduces a strong male bias into the job opportunity. The ability to reduce gender bias in language models, including reducing bias in the training and tuning process, can help eliminate this type of bias and lead to more gender-neutral outer loop selection processes which, in turn, could reduce the tendency to either send women (in this case) through unnecessary and potentially discouraging remedial training or eliminate them altogether. Of course, other biases may be introduced by AI, but increasing awareness of such biases is leading to research into how to eliminate them (Mehrabi et al., 2019; Roselli et al., 2019; Silberg & Manyika, 2019).

Recommendations for GIFT and Intelligent Tutoring Systems

There is a widely recognized need to upskill the current workforce (Bashay, 2020; Bishop, 2019; Kovács-Ondrejkovic et al., 2019; WEF, 2019). Inasmuch as ITSs have exhibited positive learning effect sizes, it is reasonable to assume that ITSs will be increasingly used for upskilling workers. When this happens, methods will be required to manage the outer loop, i.e., to automate selection of training opportunities, many of which include experiential and largely non-cognitive components. The techniques outlined in this chapter, and implemented in SkillSync, have the potential to provide such methods. In a Generalized Intelligent Framework for Tutoring (GIFT) setting, where for example, GIFT is used to orchestrate a variety of synthetic, semi-synthetic, and live experiential training environments (Goldberg et al., 2021; Robson, Ray et al., 2022), we see these methods as useful for both providing an outer loop wrapper that selects training scenarios as shown in Figure 2 and as the subject of future research.

Conclusions

The Skillsync Alignment Service offers a potential tool for addressing a major challenge in outer loop identification and selection of appropriate training materials and courses. Specifically, it addresses the disparate language, mental models, and frameworks of academic training providers and employers, effectively translating between them to find the subset of materials that represent the most efficient and effective opportunity for upskilling incumbent workers.

References

ADL. (2020). *Competency & Skills System (CaSS)*. Advanced Distributed Learning Initiative. https://adlnet.gov/projects/cass/

Bashay, M. (2020). Digital Skills for an Equitable Recovery: Policy Recommendations to Address the Digital Skill Needs of Workers Most Vulnerable to Displacement. *National Skills Coalition*. https://eric.ed.gov/?id=ED607390

Bishop, M. (2019). Addressing the Employment Challenge: The Use of Postsecondary Noncredit Training in Skills Development. *American Enterprise Institute*. https://eric.ed.gov/?id=ED596261

Bolukbasi, T., Chang, K., & Zou J. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Brown, T., Mann, B., & Ryder, N. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Bureau of Labor Statistics. (2022). *Employed persons by detailed industry, sex, race, and Hispanic or Latino ethnicity*. https://www.bls.gov/cps/cpsaat18.htm

Goldberg, B., Owens, K., Gupton, K., Hellman, K., K. S., Robson, R., Blake-Plock, S., & Hoffman, M. (2021). *Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy*. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL. https://www.gifttutoring.org/attachments/download/4295/21332.pdf

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., & Asano, Y. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*, *34*. https://proceedings.neurips.cc/paper/2021/hash/1531beb762df4029513ebf9295e0d34f-Abstract.html

Kovács-Ondrejkovic, O., Strack, R., Antebi, P., López Gobernado, A., & Lyle, E. (2019). *Decoding Global Trends in Upskilling and Reskilling*. https://www.bcg.com/publications/2019/decoding-global-trends-upskilling-reskilling.aspx

Lisle, M., & Robson, R. (2021). *Supporting companies and colleges as they reskill the workforce of the future*. The EvoLLLution. https://evolllution.com/revenue-streams/workforce_development/supporting-companies-and-colleges-as-they-reskill-the-workforce-of-the-future/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1908.09635

Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What Happens To BERT Embeddings During Fine-tuning? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.14448

Molnar, L., Mehta, R. K., & Robson, R. (2022). Artificial Intelligence (AI), the Future of Work, and the Building of a National Talent Ecosystem. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, 99–103.

National Science Foundation. (n.d.). *Convergence Accelerator*.March 18, 2022, https://beta.nsf.gov/funding/initiatives/convergence-accelerator

Robson, R., Kelsey, E., Goel, A., Egerton, L., Garn, M., Lisle, M., LaFleur, A., Kitchens, J., Northcott, E., & Robson, E. (2022). Making AI work for skills-based training: A case study. *International Training Technology Exhibition & Conference (IT^{2}EC)*.

Robson, R., Kelsey, E., Goel, A., Rugaber, S., Robson, E., Garn, M., Lisle, M., Ray, F., Kitchens, J., & Nasire, S. M. (2022). Intelligent Links: AI-supported connections between Employers and Colleges. *AI Magazine*, *43*(1), 75–82.

Robson, R., Ray, F., Hernandez, M., Blake-Plock, S., Casey, C. Hoyt, W., Owens, K., Hoffman, M., & Goldberg, B. (2022). *Mining Artificially Generated Data to Estimate Competency*. Educational Data Mining , Durham, UK.

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing Bias in AI. *Companion Proceedings of The 2019 World Wide Web Conference*, 539–544.

Silberg, J., & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute*. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.pdf

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*. https://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai16-3-02

WEF. (2019). *Towards a Reskilling Revolution: Industry-Led Action for the Future of Work*. World Economic Forum. https://www.weforum.org/whitepapers/towards-a-reskilling-revolution-industry-led-action-for-the-future-of-work

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S. & Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1910.03771.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S. & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

CHAPTER 6 - USING TLA STANDARDS TO FACILITATE AUTOMATION AND ADAPTATION ACROSS THE HUMAN CAPITAL SUPPLY CHAIN

Laura Milham[1] and Brent Smith[1,2]
ADL Initiative[1], SETA[2]

Background

The United States is engaged in global competition to advance U.S. interests and gain an enduring strategic advantage. Extraordinary challenges and unprecedented opportunities shaped by an increasingly competitive global economy, shifting demographics, and rapidly evolving technologies demonstrate the need for a highly skilled workforce that can think critically and adapt to the ever-changing operational environment. Our nation's ability to maintain military superiority over our adversaries will rely on advanced technologies such as artificial intelligence and machine learning, quantum computing and cybersecurity, automation, and robotics, directed energy and hypersonic systems, among others.

To compete for talent in a globally competitive talent marketplace, the Department of Defense (DoD) must invest in developing existing talent for the skills and jobs of the future. While the pace of technology development is advancing at an increasing rate, the supply of skilled workers has not kept up with demand. This places a heavy burden on the DoD workforce, and indicates the need for improved talent management, including more efficient and effective workforce recruiting, engagement, development, and planning (Defense Business Board, 2022). However, talent data is not currently integrated across the DoD enterprise which results in limited visibility into which skills our personnel have and which ones they are working to build.

The human capital supply chain is a complex network of systems with inherent challenges to accommodating data interoperability. Even within a single organization, the specific composition and arrangement of learning technologies will differ and change over time. This becomes a greater issue when looking across several organizations. The capabilities desired for a DoD learning ecosystem come not from individual components or databases, but from the enterprise-level collection, dissemination, and analysis of learner data that support the planning and controlling of human capital accession, including education and training. The ADL Initiative's Total Learning Architecture (TLA) defines a set of policies, specifications, business rules, and data standards for enabling a defense-wide learning ecosystem where learner data elements can be captured, shared, and interpreted for use by other DoD systems across other DoD functional areas (ADL Initiative, 2021).

DoD established the Enterprise Digital Learning Modernization (EDLM) reform in 2018 to implement the data management infrastructure required to support a TLA-enabled ecosystem of interoperable tools, technologies, and platforms across the services (Sims et al., 2020). This ecosystem uses digital learning technologies, driven by data, to provide more effective, equitable, and efficient learning opportunities across military, civilian, and DoD intel personnel. This, in turn, supports priorities for (a) upskilling and supporting the workforce, (b) enterprise shared services for information technology, and (c) data-centric digital modernization.

Operationalizing the Total Learning Architecture

The vision of a career-long learning ecosystem requires that diverse DoD learning technologies interoperate. Technologically, that means the various software systems need to be able to exchange, understand, and use data from across the enterprise. The TLA Data Strategy provides a common set of data standards and technical specifications designed to be implemented across DoD's education and training community. This overarching strategy ensures that all data resources are designed in a way that they can be used, shared, and moved efficiently across the organization.

The key to managing lifelong learning within the TLA is afforded through interoperable technical standards, linked vocabularies, and a federated catalog that provides pointers to authoritative sources of learner data. Figure 1 provides an overview of the different IEEE standards that comprise the foundation of the TLA data strategy (Smith et al., 2021). This data enables a ledger of learner performance that links learning experiences with competencies, credentials, and different career trajectories a learner may follow. This overarching strategy will ensure that all data resources are positioned in a way that they can be used, shared, and moved efficiently across the organization.

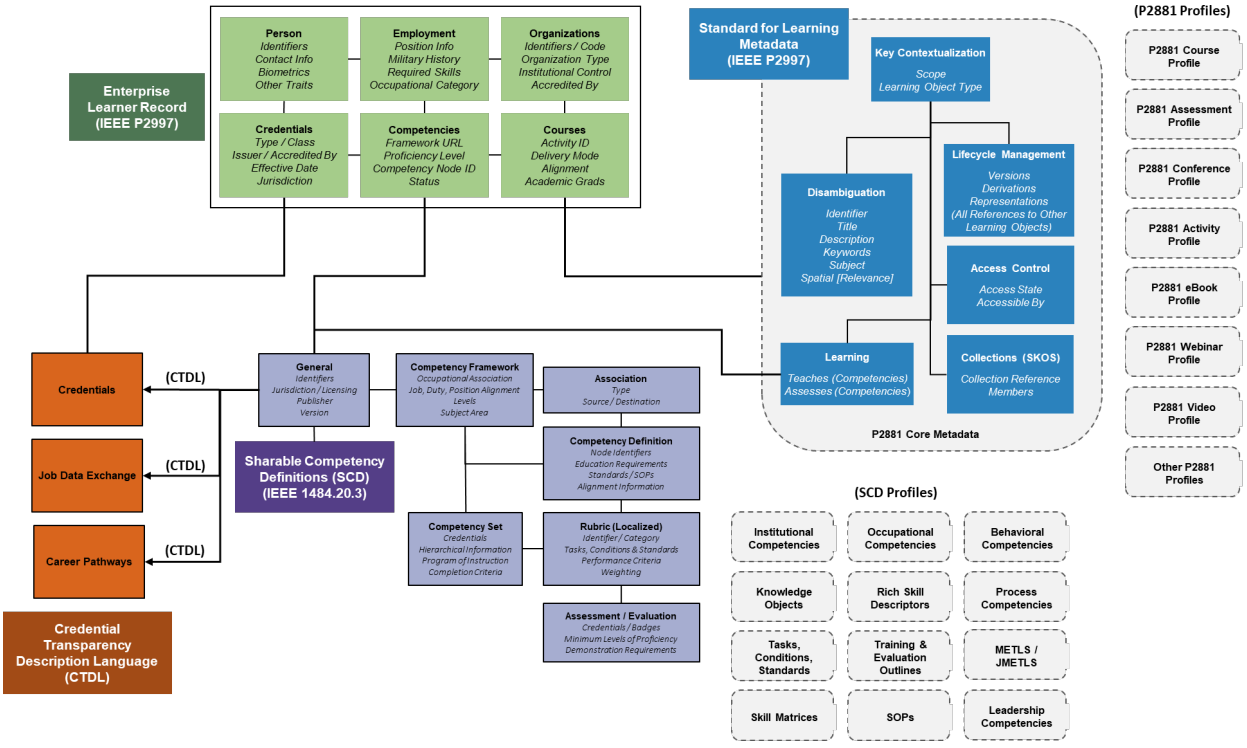


Figure 1. TLA Data Model. The TLA data model (Smith et al., 2021) is comprised of core IEEE standards. Each standard has one or more application profiles, which are schemas that consist of data elements drawn from one or more namespaces, combined, and optimized for a particular application.

The TLA assumes a set of enterprise services and associated infrastructure to ensure semantic interoperability, maintain digital identity for users, and operate within the needs of the interconnected digital world. Every device or service in the ecosystem appears as either a learning record provider (LRP) and/or a learning record consumer (LRC). The resulting architecture is asynchronous, and event driven. From a digital learning perspective, The DoD's Identity, Credentialing, and Access Management (ICAM)

policies are used to link an individual's unique identity to their career-long education and training records created and stored across various DoD schools and training sites.

Within the TLA, the learner is the critical element to defining overall system behavior. The learner is always present and it is the learner that defines the context under which learning services should respond. The Experience API (xAPI) standard is used to track a learner's interactions and performance across different systems (e.g., learning activity, registration system). The standard defines the general structures for creating xAPI statements. The TLA Master Object Model (MOM) is an xAPI profile employed as a conceptual model that describes how to link learners with all the different learning experiences (e.g., learner pathway) encountered throughout their DoD career. The MOM captures the context of these experiences (IEEE P2881) and aligns them with the competencies (IEEE 1484.20.3), credentials (CTDL), and other key performance metrics in the operational environment. The MOM captures the object life cycle of learners executing a single "thread of learning" that culminates in the reporting and evaluation of a learning event.

TLA Use Case: Professional Career Education

When considering how TLA data standards are used to support professional career education, it is helpful to consider how the DoD workforce really learns. People learn through a diverse array of channels and formats. Nearly half of the workforce says the biggest constraint on their learning is time (LinkedIn Learning, 2022). They also report that 86% of learning happens in short bursts of 45 minutes or less. Research shows that people are constantly learning while on the job by searching for resources and reading, listening, watching, or engaging in activities that improve their ability to perform.

Learning is about information transfer, not necessarily application and impact. Skilling is the transfer of knowledge with an intent to bring impact through behaviors and actions on the job. When people can apply knowledge to address specific issues, they are using their skills (Harvard Business, 2019). Learning will always be the cornerstone of professional career education, but right now, building skills is what is most urgent. Skills can be benchmarked, quantified, analyzed, and aligned to an organization's operational objectives. People also learn for personal gain. This motivation is among the most powerful drivers of engagement. Many leaders are still focusing on knowledge retention, course completions, and satisfaction surveys. Ongoing modernization efforts across the DoD training and education components are reimagining their curriculum to better serve the needs of their personnel and to help their organizations to respond more rapidly to changes in the operational environment.

While formal courses still play a huge role in professional career education, the DoD workforce is constantly learning within the flow of work by reading, listening to, and watching learning content that they find through various means. However, there is a disconnect between tracking learner progress through these independent learning activities and the tracking that occurs within the formal programs of instruction available throughout the department. Moreover, most formal training and education programs were not designed for today's learner habits where learning and skill development is part of everyday routines (Udemy Business, 2022). The opportunity is to consider how intelligent tutoring systems (ITSs) can incorporate these independent, unscheduled learning activities to connect skill building to career pathing, professional development, and retention.

These trends and challenges dictate the need for new methods and tools to assess learner proficiency outside of these formal programs of instruction. People want learning integrated into the tools and technologies they already use on a day-to-day basis. These informal learning opportunities change the nature of how learners are assessed. Assessments need to help people understand what skills they need to obtain to reach both personal and business goals. Ideally, assessments become integrated into the flow of work so that skill development becomes part of an individual's daily routine. Self-regulated learning (SRL) allows

individuals to select learning topics based on job demands or other interests and then control and enhance their learning through processes such as goal setting, strategy selection, and monitoring (Fowlkes-Ratcliff, 2022). SRL also has potential to change the nature of how ITSs operate. Workers thrive when they are provided consistent guidance on goals for upskilling, new development opportunities, and just-in-time feedback on their operational performance.

TLA data standards allow the myriad of disparate informal learning systems to exchange data between different organizations. Standards reduce the time spent cleaning and translating data when separate organizations have agreed to exchange data. TLA data standards define entity names, data element names, descriptions, definitions, and formatting rules. Standardized data on its own, has potential to become inflexible and overly constrained in time. TLA standards are not designed to be overly prescriptive in how the data is defined. TLA core standards include a minimum set of data elements for each TLA data pillar.

Application profiles are used to define the alignment between TLA data and the different types of systems (e.g., learning interventions) that are used to support human capital management and talent development. An application profile is a structured template of information that describes a data container. A key strategic concept for a profile is that it contains minimal information requirements to assure a container is sufficiently described for self-identification to support any enterprise query or data sharing need. This information is provided to enable valid enterprise consumption of the associated data.

The xAPI standard uses this approach and numerous xAPI profiles have been created to support the implementation and adoption of xAPI across different media types, learning modalities, and instructional domains. An xAPI Profile server supports the different communities of practice responsible for the creation and maintenance of xAPI profiles.

As an example, Figure 2 shows the IEEE P2881 Standard for Learning Metadata (this is a draft), which includes the P2881 core, a small set of data elements required for every learning asset in an organization. A P2881 course profile expands on the core to define what data should be used to describe a course. P2881 profiles increase the fidelity and granularity of data that can be collected about the myriad of learning resources available within the DoD. Numerous P2881 profiles will be created to describe the specific data that should be collected for the different types of learning resources that are used in the DoD. These might include simulations, serious games, webinars, conference proceedings, mobile applications, among others. This approach makes the governance of the TLA data strategy more flexible by enabling the training and education community to create profiles that best suit their needs without having to modify the standard.

Controlled vocabularies are used to populate each profile's data elements; these are also provided to inform the architectural design patterns applied to develop different types of human capital management systems that consume TLA data. In software engineering, a design pattern is a reusable solution to a common task within a given set of similar contexts. These work in concert with the TLA microservices to allow different systems to publish and subscribe to different types of TLA data. The DoD linked data and schema server will provide a single source of truth (i.e., authoritative data source) for those data definitions and will

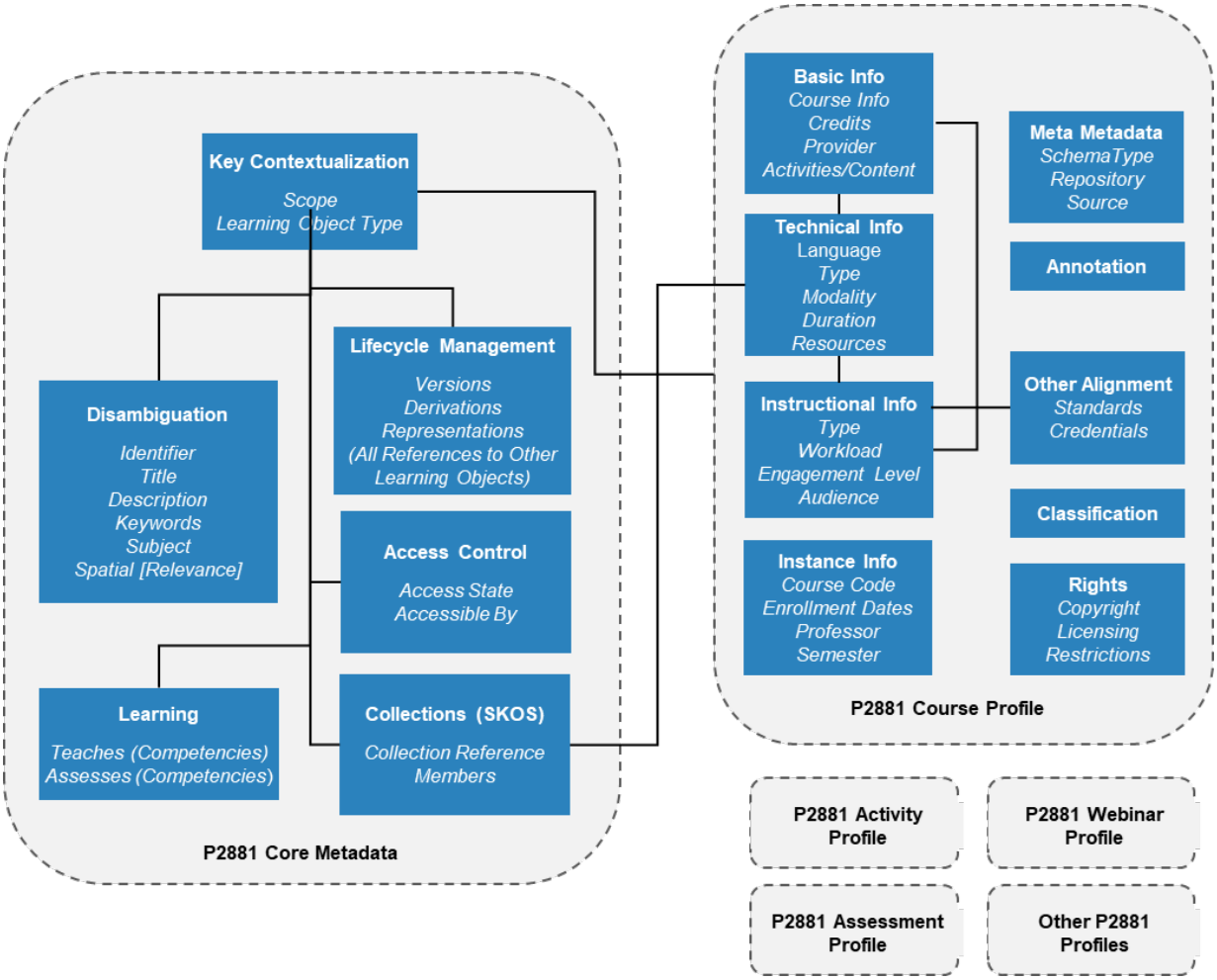


Figure 2. Draft P2881 Standard for Learning Metadata. The IEEE P2881 standard includes a P2881 core and numerous P2881 profiles. P2881 metadata is used to describe different learning experiences and link each learning experience to the knowledge, skills, and other behaviors being taught. These learning outcomes are aligned with competencies, credentials, occupations, and careers.

establish immutable Internationalized Resource Identifiers (IRIs) for each term and schema that all DoD technologies can reference.

Linked Data is essential to preserving the meaning and context of data communicated between TLA conformant systems, without requiring the transmittal of the entire data definition with each data set. It helps abstract the definitions of data elements away from the data sources themselves, which improves data integrity, overall system resiliency, efficiency, and semantic interoperability. Linked Learner Data is a methodology for defining and exposing data vocabularies via published, structured metadata that can be interpreted by humans and machines to enable semantic interoperability. This ensures that different systems use specific terms in the same way. It also helps clarify the relationship among elements, data element formats, and pre-defined assemblages of terms.

TLA Control Loops: Enabling Different Views of the Learner Data.

The TLA MOM conceptual model enables a ledger of lifelong learning that preserves the *chain of evidence* generated by each learner's path through the myriad of learning experiences encountered across their DoD career. The MOM conceptual model includes statements that describe key learner milestones for tracking and managing learner progression from the macro (i.e., career state) level to the micro (i.e., learning experience) level. The TLA's MOM verbs are grouped to represent the learner state within the different systems a student interacts with across the 5 TLA control loops shown in Figure 3 (Smith et al., 2021).

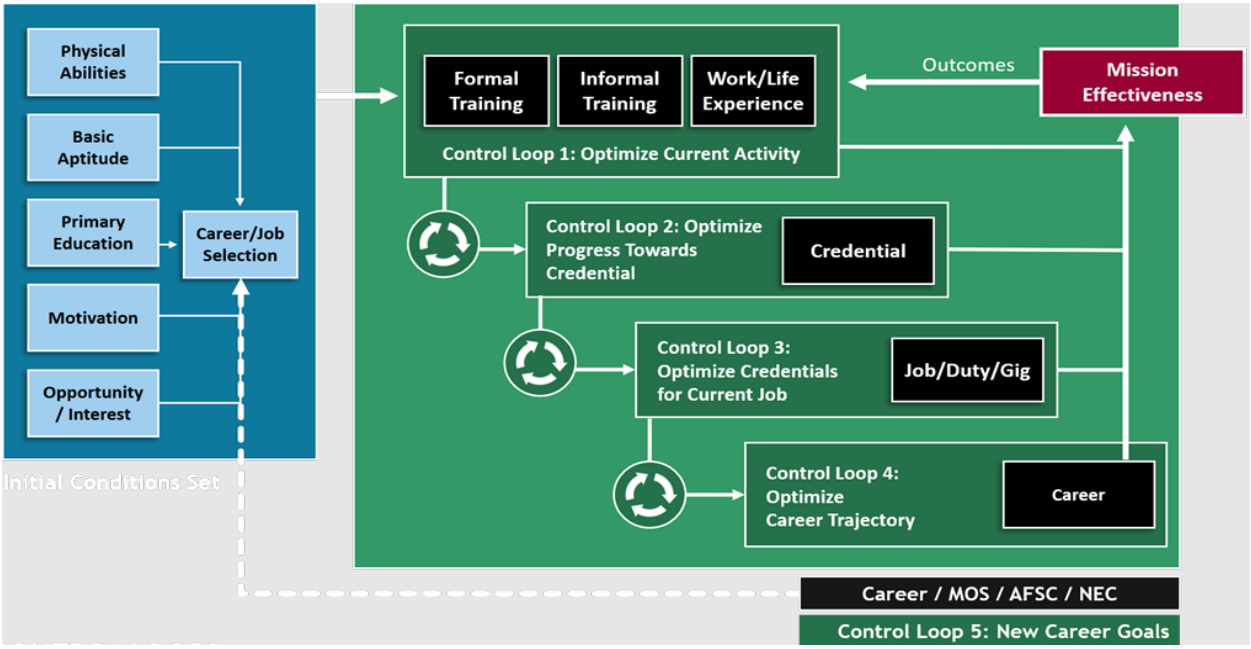


Figure 3. TLA Control Loops. TLA Control Loops operate in parallel but provide a convenient way to limit and categorize data displayed in decision support aids. The TLA MOM profile helps organize these filters.

The TLA control loops were created to show that learning data may be viewed from different perspectives requiring different levels of granularity and fidelity over different time horizons. In other words, the same data collected from a single learning experience may be used in different ways depending on the purpose for which it is being used. For example, a learner may be pursuing a specific job credential required for promotion. They need to participate in one of more courses (e.g., a sequence of learning activities) in

support of their career trajectory. This example can be viewed in the context of control loops 2, 3, and 4. The five control loops in order of ascending time horizons address:

- **Control Loop 1**: Improving a learner's mastery of competencies within the current learning experience (e.g., intelligent tutoring, instructor support). This control loop typically uses the xAPI data stored in a Noisy LRS (learner record store).
- **Control Loop 2**: Optimizing a learner's progress toward a credential. This control loop uses TLA MOM statements and the P2881 Standard for Learning Metadata to optimize the delivery of different learning experiences in pursuit of a Credential (e.g., degree, certificate, license).
- **Control Loop 3**: Prioritizing the pursuit of credentials or activities to meet requirements for a job. This control loop uses the TLA MOM statements and Sharable Competency Definitions to optimize an individual's learning plan in pursuit of a job.
- **Control Loop 4**: Career field management including the planning and execution of education and training goals for an overall career trajectory. This control loop uses all TLA Core data and introduces a new component to the TLA's competency pillar. The Credential Transparency Description Language (CTDL) and Job Data Exchange (JDX) standard is used to decompose position descriptions into their required competencies and credentials.
- **Control Loop 5**: Providing options for supporting post-career transition and retraining to pursue other career goals. This control loop relies on the historical TLA data generated by a learner to identify gaps between the competencies and credentials they currently have and the requirements for the new job/duty/occupation they wish to pursue.

Throughout a career, learning may unfold in a variety of ways from self-regulated learning to formal programs of instruction. The development of application profiles for the different TLA standards need to consider the different views and time horizons. The data generated in control loops 2, 3, 4 and 5 inform individual learning goals. These are organized according to their required competencies and credentials, which are also aligned to jobs, position descriptions, and career milestones (Credential Engine, 2022). Learning activities and experiences are aligned to these competencies and organized to achieve the underlying learner goals. The TLA MOM states provide the mechanism to track the learner as they progress through these formal or informal programs of instruction. In either case, the launching and capture of the TLA MOM's Learning Activity State provides the evidence for demonstration of competency. Other connected TLA systems use this information for a wide range of purposes.

Recommendations for GIFT and Intelligent Tutoring Systems

TLA data standards enable new opportunities for ITSs, automated instructor support tools, and other adaptive instructional systems that span the TLA control loops. The TLA Capability Maturity Model (CMM) shown in Figure 4 was developed to appraise an organization's capabilities in meeting EDLM (Enterprise Digital Learning Modernization) and TLA requirements. The CMM attempts to provide a baseline view of organizational capabilities with the intent of communicating the value of organizational improvement via the associated risks and benefits. CMMs define a multi-level and multi-dimensional path of increasingly organized and systematically more mature processes. "Maturity" refers to the degree of process formality and optimization employed by an organization, from ad hoc practices to formally defined steps, to managed result metrics, to active optimization of processes.

- **CMM Level 1**: Use performance tracking to measure learning outcomes. At the most basic level this means instrumenting your Learning Management System (LMS) to publish xAPI statements and establish a LRS to capture those statements.
- **CMM Level 2**: Federate LRSs and integrate one or more course catalogs. Integrate multiple sources of xAPI into local federated LRSs. Shift from local to enterprise-level Identity, Credentialing, and Access Management (ICAM). Store course catalog data in an organized digital system. Other learning technologies may connect to the LRS from inside or outside the network security boundary.
- **CMM Level 3**: Consolidate course catalogs and use competency-based (outcome-based) learning. Connect local course catalogs and content repositories into a consolidated organizational resource. Use Sharable Competency Definitions to describe the knowledge, skills, abilities, aptitudes, and other behaviors being taught. Use the TLA MOM to drive competency management across disparate pools of learner data. Conformance to TLA core data standards creates interoperability with defense-wide systems.
- **CMM Level 4**: Institutionalize competency-based learning and link learner records into an enterprise learner record. Shareable Competency Definitions are used throughout the organization for courses and learner performance. Competencies and credentials conferred from external organizations are considered in the progression of each learner. Learner proficiency is demonstrated across training, education, and operational systems (performance support tools, on-the-job training, performance reviews). Learner records are aggregated into a local Learner Profile.
- **CMM Level 5**: Add enterprise learner records and integrate with human capital management systems. Local learner records are aggregated into a defense-wide data fabric that includes pointers to the authoritative data sets that comprise an individual's lifelong learning journey. Each learning activity generates xAPI to track learner performance, uses P2881 Metadata to describe instructional resources, and aligns with Sharable Competency Definitions that define the knowledge, skills, abilities, and other behaviors required to meet operational objectives. These data are shared with other defense systems to continuously improve the efficiencies of how we train, educate, and operate.

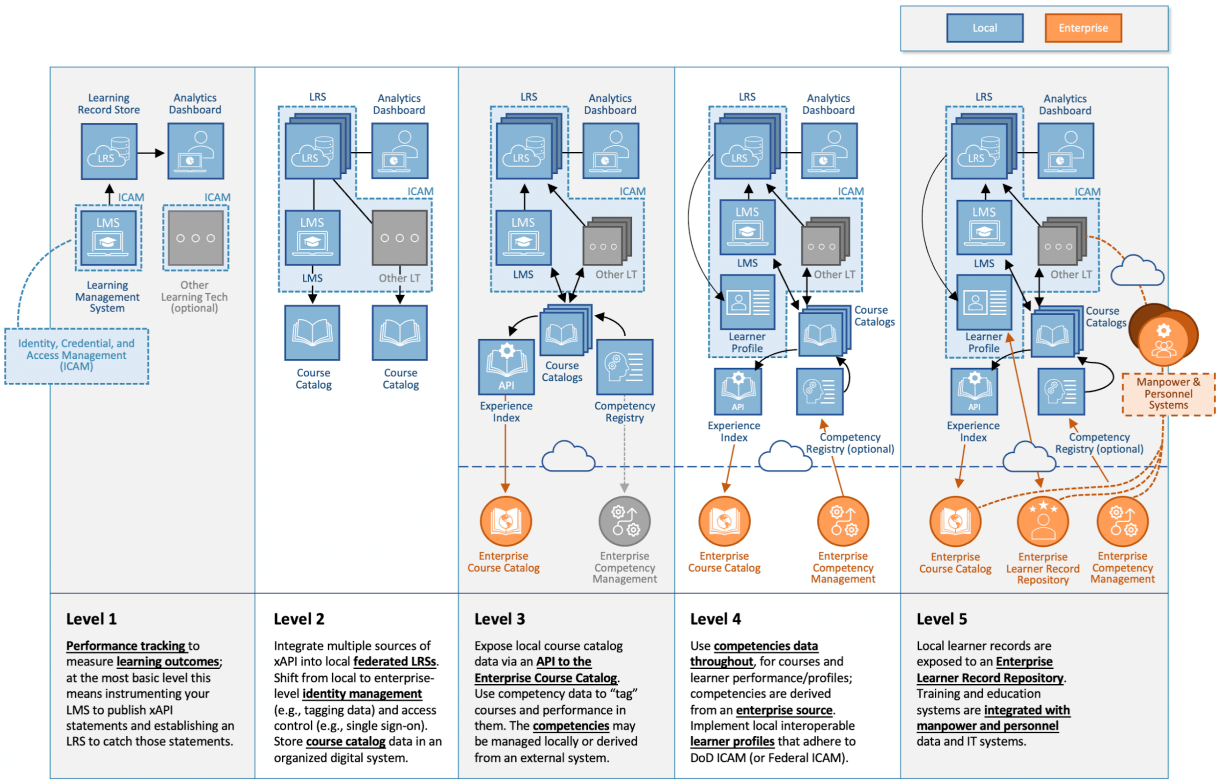


Figure 4. TLA Capability Maturity Model. The ADL Initiative's Capability Maturity Model (CMM) provides a thorough description of maturity, based on the context of policies, instructional design processes, technology infrastructure, and governance of its adherence to the TLA's data interoperability standards.

More importantly, the model can be used by an organization to quantify their current maturity and to highlight areas where improvement is needed to help guide future investments. The CMM allows for the gradual migration of legacy systems to a microservice-based infrastructure of core services that federate data across other technology components. Beyond the technical considerations, the CMM also evaluates the current state of organizational maturity through the lens of the processes, workflows, and incentives that promote learning and enable continuous process improvement. The CMM helps identify an organization's current state and identifies areas for improvement based on organizational priorities.

Conclusions

Data underpins digital modernization and is the fuel of the decision-making process within the DoD. The DoD Data Strategy describes an ambitious approach for transforming the Department into a data-driven organization. The TLA aligns the DoD's training and education community with the broader DoD Data Strategy and other defense-wide initiatives to enable a comprehensive strategy for (a) protecting the privacy and security of learner data, (b) enabling continuous process improvements throughout the continuum of lifelong learning, and (c) establishing a federated data strategy across the human capital supply chain.

Acknowledgement

The appearance of external hyperlinks does not constitute endorsement by the United States Department of Defense (DOD) of the linked websites, or the information, products or services contained therein. The DOD does not exercise any editorial, security, or other control over the information you may find at these locations.

References

2022 Workplace Learning Report - The Transformation of L&D (2022) Linkedin Learning. Available at: https://learning.linkedin.com/content/dam/me/learning/en-us/pdfs/workplace-learning-report/LinkedIn-Learning_Workplace-Learning-Report-2022-EN.pdf (Accessed: January 16, 2023).

2022 Workplace Learning Trends Report. (2022). *Udemy Business*. Available at https://business.udemy.com/2022-workplace-learning-trends-report/?utm_source=direct&utm_medium=direct (Accessed: March 28, 2022)

ADL Initiative (2021). *Total Learning Architecture (TLA) Functional Requirements Document*. ADL Initiative. https://www.adlnet.gov/assets/uploads/2021%20TLA%20Functional%20Requirements%20Document%20w%20SF298.pdf

Credential Engine (2022). *CTDL Pathways, Conditions and Constraints: Essential Reading: Pathways Overview*. Available at https://docs.google.com/document/d/1I5MHBTbZG04N-l6kCyMExdY8HXUXE8oN6RoDNT2KadM/edit# (Accessed: January 16, 2023).

Defense Business Board (2022). *Strengthening Defense Department Civilian Talent Management*. Defense Business Board. https://dbb.defense.gov/Portals/35/Documents/Reports/2022/DBB%20FY22-03%20Talent%20Management%20Study%20Report%2018%20Aug%202022%20-%20CLEARED.pdf

Fowlkes-Ratcliff, J., (2022). Directed Self-Regulated Learning and Learning System Support. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Washington, DC: NTSA.

How the Workforce Learns in 2019. (2019). Harvard Business Publishing. Available at http://www.harvardbusiness.org/wp-content/uploads/2019/10/How-Workforce-Learns-2019-Report.pdf (Accessed: January 16, 2023)

Sims, K., Schatz, S., Brewer, V., Rogers, A., McMahon, S., (2020). Enterprise Digital Learning Modernization: What, Why, and Who Says So? In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Washington, DC: NTSA.

Smith, B., Schatz, S., & Turner, J. (2021). Total Learning Architecture Data Model for Analytics and Adaptation. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Washington, DC: NTSA.

SECTION II – SPECIFIC APPLICATIONS

Anne M. Sinatra[1], Xiangen Hu[2], Arthur C. Graesser[2], and Lisa N. Townsend[1], Eds.

[1]U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center
[2]University of Memphis Institute for Intelligent Systems

CHAPTER 7 – INTRODUCTION TO SPECIFIC APPLICATIONS

Arthur C. Graesser[1], Xiangen Hu[1], Anne M. Sinatra[2], and Lisa N. Townsend[2]
University of Memphis[1]; US Army DEVCOM Soldier Center[2]

Core Ideas

Intelligent Tutoring Systems (ITSs) and other types of adaptive instructional systems are expected to differ among professions. There is no universal approach for designing technologies to support professional career education. Training for some professions may require virtual reality or augmented reality, whereas other professions may regard these technologies as expensive and superfluous. Just-in-time conversational agents may benefit learning and motivation for some populations and professions, but for others it would be sufficient to have an adaptive selection of texts, pictures, and videos. The chapters in this section discuss the use and tests of intelligent technologies in the education of individuals in specific careers or populations.

While reading these chapters it is worthwhile to consider the eight affordances of digital learning technologies that were identified in *How People Learn* volume 2 of the National Academy of Sciences, Engineering and Medicine (2018). The affordances are: interactivity, adaptivity, feedback, choice, nonlinear access, linked representations, open-ended learner input, and communications with other people or agents. These digital affordances rarely exist when individuals read a textbook, watch a video, or listen to a classroom lecture -- which for many decades have been the three most dominant media for education and training. This begs the question of whether the technologies with the various digital affordances have added value. Answers to this question are important for the Generalized Intelligent Framework for Tutoring (GIFT) in achieving the goal of personalized learning.

Individual Chapters

Dorneich, Wu, Gilbert, and *Winer* discuss the role of off-site mixed reality training for many jobs in the workforce. Mixed reality (MR) has virtual and real-world elements in a 3-dimensional scene, thereby incorporating the resources of both virtual reality (VR) and augmented reality (AR). MR would be less expensive and potentially more effective pedagogically than on-site training that occurs either through human instruction or on-site simulation environments. How does MR compare with alternative training methods? They conducted research on AR for aerospace manufacturing and pilot training in aviation, and VR for astronaut training. The chapter discusses the advantages and possible limitations of MR in these case studies and the potential for adaptive intelligent systems to improve the training.

The chapter by *Biddle* focuses on commercial pilot training with Boeing. An off-site, adaptive instructional system with immersive flight simulation has the potential to have added pedagogical value, reduce training time, and reduce costs compared with on-site training. An ITS that considers individual demographic variables, flight experience, English language proficiency, and culture is expected to contribute to the adaptivity, interactivity, and other digital affordances. A sufficient assessment methodology is needed to make comparisons among the alternative training interventions.

Swartout, Nye, and *Rizzo* describe a Personal Assistant for Lifelong Learning (PAL3) that has an adaptive conversational agent to guide Navy personnel on their career paths. For those Sailors in technical areas, there is a need for fundamental skills in domains such as algebra, physics, electronics, and computer science. However, all Sailors also need training on how to cope with life skills. In this chapter the focus is on

training for suicide prevention, which raises questions about privacy and important nuances on how the conversational agent communicates with the Sailor.

Graesser and *Hu* raise the question of what populations, tasks, and subject matters are likely to benefit from conversational agents in ITSs. They report that struggling adult readers both like and learn from an AutoTutor ITS that interacts with them in natural language with two agents (a tutor and peer). However, the effectiveness of conversational agents on liking and learning is more complex in a Navy project with high ability trainees learning about electronics; some groups of Sailors like and learn from them, but others do not. The chapter proposes that researchers need to sort out the conditions in which adaptive conversational agents have added value.

The chapter by *Lajoie* and *Li* focuses on medical education. Medical students learn through BioWorld, an ITS that allows students to practice their diagnostic reasoning skills with virtual patient cases. BioWorld adopts a cognitive apprenticeship framework that situates medical students in a virtual hospital setting in which they review and diagnose patient cases by collecting patient data. Students have access to a medical library and a consultant. The researchers conducted empirical studies tracking the students' cognition, motivation, emotions, and metacognition in an effort to explore how these psychological components interact during learning.

Babin and Robinson review how tacit knowledge is ideally acquired through experiential education and training environments in military professional development. They propose that the communicative exchange between an expert and novice is much more important for tacit knowledge transfer than simply having the expert lecture to the novice. Tacit knowledge is best transferred through discourse when there is a common vocabulary, concrete concepts, and a common operating picture while the novice actively performs tasks. It is also important to have mutually respectful and trusting expert-novice relationships, multiple opportunities for the expert to model correct performance, and reflection over time.

Cheng, Prihar, Baral, Gurung, Botelho, Haim, C. Heffernan, Patikorn, Sales and *N. Heffernan* discuss the value of crowd sourcing in developing the content of an ITS, as an alternative to the traditional approach of learning scientists and software developers creating the content with authoring tools. The crowd includes teachers and students in addition to researchers. This team has successfully implemented the crowd-sourcing approach in their ASSISTments system. Teachers have an authoring tool to create and modify questions, hints, explanations, comments, feedback, and other content after examining a large number of open-ended student responses. This is a promising advance for the professional development of teachers who want to use adaptive instructional technologies.

Reference

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

CHAPTER 8 - UTILIZING MIXED REALITY TO SUPPORT ADAPTIVE WORKER TRAINING

Michael C. Dorneich[1], Peggy Wu[2], Stephen B. Gilbert[1], and Eliot Winer[1]
Iowa State University[1]; Raytheon[2]

Introduction and Background

Worker training is adapting to meet the rapid pace of technological change. Suggested approaches include improving access to on-demand training, developing workplace competency-based training, verifying and certifying worker skills, and building a robust apprenticeship system that leverages learning by doing in context (Lerman et al., 2020).

However, workforce development and workplace training face many challenges. Training can be conducted offsite in dedicated facilities or on-site. Offsite training may require dedicated facilities that need to be maintained, travel, and time away from the job (Carruth, 2017). Furthermore, dedicated training facilities may require expensive mockups to recreate the work environment in the fidelity needed to be effective if the context can be faithfully recreated. On-site training has the advantage of training the tasks in the context of work. Often referred to as On-the-Job Training (OTJ), OTJ typically uses an apprenticeship model where the trainee observes an expert interweaved with some hands-on experience. OTJ tends to be low cost as there is little to no preparation of materials ahead of time. However, it depends on the availability of experts who are qualified as trainers. It may also slow productivity as workers train on the line, where the same level of product quality must be maintained, and the trainees have a limited scope of errors while learning. Safety of employees and damage to equipment can also be a concern (Carruth, 2017). If dedicated training equipment is used on-site, their availability may be limited. For example, in the aerospace industry, highly complex and expensive equipment such as jet engines may not have an equivalent training prototype for learning purposes. This means OTJ can only occur when work is done on those parts. If those parts or procedures are rarely performed in certain facilities, trainees may need to travel to other facilities for either initial or refresher training and assessment.

The nature of the tasks being trained also plays a role in the type of feasible hands-on worker training. For dangerous, hazardous jobs, it may not be feasible to train on-site, and offsite training facilities may be limited and hard to maintain. Ideally, training should be supervised in a controlled setting that recreates realistic situations as closely as possible (Mossel et al., 2015). However, it may be challenging to recreate situations with all the context needed for effective training. Offsite training that does not include important contextual elements of the task environment may not adequately prepare workers for performance in the operational environment (Keinan & Friedland, 1996). Likewise, training for emergency procedures is difficult, and emergency response strategies are often trained through classroom instruction and simulation (Uhlig et al., 2016). However, this is inadequate given the potential low frequency of occurrence, the safety implications of a failure to respond appropriately, and the limited opportunities to practice (Carruth, 2017).

Worker training often relies on dedicated trainers or other workers in an apprenticeship-style approach. Engaging learners actively while under expert supervision enables the learner to gradually move from the periphery to full participation. Intelligent Tutoring Systems (ITSs) often use the apprenticeship learning model to train new skills and tie theory to practice (Dorneich & Jones, 2001; Katz et al., 2020). Adaptive ITSs can adaptively tailor their behavior to better support the learner in the moment, given their current state. Adaptive systems have four general categories of modification: who does what (the system or the

learner), when content is scheduled (i.e., task order, task types), how the system interacts with the learner, and what is being taught (i.e., content) (Feigh et al., 2012).

Mixed Reality (MR) includes virtual and real-world elements in the three-dimensional scene. Virtual reality (VR) composes the scene entirely from computer-generated graphics. Augmented Reality (AR) incorporates virtual graphical objects in the real-world scene (Pan et al., 2006). MR provides a training modality that can address many workplace training challenges. For instance, VR-based training has been shown to be a safe, controlled, and cost-effective way to train workers in authentic scenarios without exposing them to real-world hazards (Adami et al., 2021; Finseth et al., 2022). Three-dimensional AR content has been shown to improve learning and motivation (Akçayır & Akçayır, 2017; Meister, Wang et al., 2022a).

Furthermore, MR equipment is now very accessible, including VR headsets at consumer prices and the ability to display AR on tablets and smartphones. MR-based adaptive training can also be used with reduced supervision, enabling trainees the flexibility to train at times that do not interfere with the performance of their duties (Carruth, 2017). MR environments provide immersive worlds that provide users with a strong sense of being present in the virtual/augmented world (Dede, 1995; Dede et al., 2017). This can be used to create realistic environments to replicate real-world training.

Goals and Scope

This chapter explores the use of MR to develop training systems for worker skill development. Three use cases are discussed to demonstrate the potential of MR-based worker training approaches. In the first use case, an AR-based virtual engine model was evaluated to quantify new technicians' training effectiveness and trainee preferences for Manufacturing, Repair, and Overhaul (MRO). An AR app was developed to present a 3D model of a jet engine's High-Pressure Compressor (HPC) component, where parts can be viewed separately or in concert in an AR app. The app can be used in a classroom setting where there is no access to physical mockups. The second case describes an AR approach that was used to provide an exploratory learning environment for General aviation pilots to enhance their knowledge of weather. The AR was embedded in scenarios to train students to make safe, timely, and appropriate weather-related decisions. This presents a learning opportunity where novice pilot errors can be elicited in a safe AR environment with feedback provided. The final use case describes a VR-based adaptive stress training system to enhance resilience to stress during emergency procedure training of astronauts. The closed-loop, adaptive training system automatically adapted the environmental stressors to an appropriate level given the individual's current stress tolerance while gradually increasing stressors over time. Combined with a graduated stress exposure pedagogy, this training system did not require the supervision of a trainer to personalize the training. In addition, utilizing VR to create a realistic training environment is the first step toward a mobile training system for use on the ground, during spaceflight, or on the Martian surface. Finally, the chapter will discuss how MR can augment ITSs in an apprenticeship learning model to train new skills and tie theory to practice.

State of the Field and Supporting Research

Use Case 1: AR for manufacturing

Training Challenge. Within the aerospace industry, new technicians for MRO may come from a variety of experiences and backgrounds. Some new employees may be fresh graduates from a technical school, while others may have decades of experience in an adjacent industry, such as automotive repair, or deep expertise as retired military equipment maintainers. For this reason, developing training content that is engaging for

all is a challenge. Further, in highly hands-on fields such as engine repair, access to the right equipment is critical for instructors to convey knowledge. However, due to the size and cost of the engines and specialized parts, the actual equipment, or their equivalent training mockups, may not always be available. Instructors often need to use photos, diagrams, and/or a partial set of physical components based on availability. For training on large objects, the instructor's ability to share a common point of reference is important for building shared situation awareness. However, it is also important to enable trainees to control their own views to have sufficient time to see what they need to do, or for additional exploration. When a sufficient supply of physical artifacts and space is available, trainees can build shared situation awareness by attending to the instructor's reference and exploring at their own station. In the case of jet engines, trainees simply cannot have their own physical model of engines to explore.

AR-based approach. Using virtual engine models instead of physical parts or photos is one solution for providing instructors and trainees with access to highly accurate representations of engines that are otherwise unavailable. However, a full virtual engine-based training solution is time-consuming and costly to develop. A proof-of-concept prototype was created to evaluate the differences in training effectiveness and trainee preferences and to quantify the justification for a virtual engine training approach. The HPC component of a jet engine was modeled in 3D, where parts can be viewed separately or working in concert in an AR app. The app can be used in the classroom setting where there are no physical mockups.

Forty-three new technicians without prior exposure to the HPC component were selected to participate in an informal study. Participants were randomly assigned to one of four groups:

A) Eight trainees were provided a 5-minute introduction by an instructor. They were then provided individual tablets with a pre-installed AR app and instructed to explore the engine parts in the app at their own pace for a maximum of 10 minutes. No additional materials were provided.

B) As with A), eight trainees were provided the same 5-minute introduction and app to explore at their own pace. They were also provided a set of physical parts of the engine to share among the group and examine during their allotted time of 10 minutes.

C) Eight trainees received the same 5-minute introduction, then watched a 10-minute video capture of a walk-through of the AR app on a shared screen. The video showed a 360 rotation of each part and the text describing each part. The content was drawn from the same app for groups A and B. The video did not contain any narration.

Control) Nineteen trainees in the control group received a standard 15-minute lecture using images. It should be noted that the instructor provided the same content for developing the AR app.

Participants in all the groups received a total of 15 minutes of training. After the training session, all participants continued with other training sessions and a facility tour. After four hours, they received a multiple-choice knowledge retention test about the material presented in the app, such as names and basic functions of components. After the test, participants were invited to try the experience of the other groups, rank their preferences for each, and provide feedback on the different content delivery methods.

Results and Discussion. Table 1 illustrates the multiple-choice scores by group. Participants' overall mean score on the multiple-choice test was 58%. The average for the control group was 41%, whereas the average for the combined groups using the AR app directly or video capture of the AR app was 72%. The scores indicate that the knowledge test was designed to be adequately challenging to avoid ceiling or floor effects.

Table 1 presents a statistical comparison of the performance on the multiple-choice test for each of the four training groups. Pair-wise two-tailed *t*-tests of the test scores showed no statistically significant differences

between AR with or without parts (p=0.527). Compared to the AR app in general (either with or without parts), the video capture approaches statistical significance (p=0.069). The control group was significantly different from all the other conditions (p<0.05).

Table 1. Comparison of the four training group conditions.

Group	Condition	Exam Score (SD)	p X vs. A	p X vs. B	p Control vs. AR (A&B)	P X vs. C
A	AR no parts	64% (24%)				
B	AR with parts	71% (22%)	0.527			
C	Video	81% (12%)	0.094	0.292	0.069	
Control	Lecture	41% (13%)	0.032*	0.006*	< .001*	< .001*

Objective performance scores were lowest in the control group, who received standard training via lecture. However, there were no significant differences in groups A, B, and C. In the individual debriefs, trainees overwhelmingly preferred the AR app over the video capture and over the standard lecture style. While subjects preferred access to physical parts over no access, there was a lack of significant findings between the conditions of AR with or without parts. Similarly, there was no significant difference between AR apps in general versus video. We believe this makes intuitive sense within this use case due to the nature of the evaluation. The multiple-choice test only evaluated the participant's recall of factual information, such as the name of parts and their function. It did not test spatial knowledge about relative physical dimensions or how parts fit together. Future work includes evaluating the impact of using parts on building spatial knowledge.

Use Case 2: AR student pilot training in General aviation

Training Challenge. General aviation students and novice pilots have few opportunities to experience weather-related situations that require them to exercise their knowledge to make weather-related decisions (Berendschot et al., 2018; Johnson et al., 2017). Weather training is often restricted to ground-based classroom training focused on knowledge acquisition. Documented training gaps include failure to retain the knowledge of the differences in weather patterns and their associated visual cues, poor decision-making due to the inability to interpret, correlate, and apply weather information during flight, and resulting poor situation awareness and design-making ability (Carney et al., 2015). Pilots report that they do not feel prepared to make weather-related decisions during flight (AOPA Air Safety Institute, 2018; Major et al., 2017). Scenario-based weather training with flight simulators is immersive and visually realistic but is not always accessible and is not designed explicitly for visual weather training (Berendschot et al., 2018; FAA, 2009).

AR-based approach. Training general aviation pilots could be improved by providing enhanced opportunities to interact with accessible, dynamic, and visually realistic representations of weather phenomena. In this second use case, 3D AR scenario-based training learning experiences (see Figure 1) were integrated into existing text-based materials. This smartphone and tablet-based, AR-enhanced training, called interactive print, provided a low-cost, immersive, and visually realistic depiction of thunderstorms. The AR was embedded in scenarios to train students to make safe, timely, and appropriate weather-related decisions.

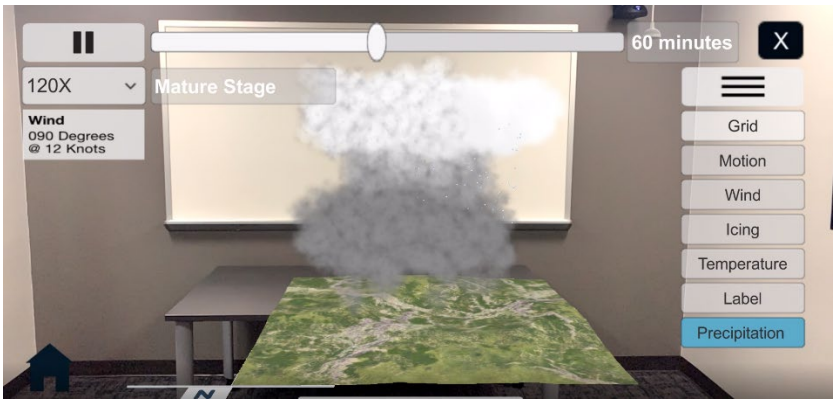


Figure 1. AR depiction of a developing thunderstorm cell while reading text.

Results/Benefits. A series of studies were conducted to assess the utility and benefit of AR scenario-based training. Evaluations found that the students significantly improved their factual knowledge and visual knowledge, with high levels of motivation (Meister, Miller et al., 2022; Meister, Wang et al., 2022b). When comparing traditional print to AR-enhanced interactive print (for full details, see Meister, Wang et al., 2022a), it was found that weather-related decision-making performance decreased in the AR condition. While initially counter-intuitive, discussions with participants anecdotally revealed that they were making the same mistakes novice pilots make in the air when trying to apply text-based book knowledge to in-flight weather-related decision-making. Participants likely relied on visual cues rather than their weather-related knowledge and guidance. This presents a learning opportunity where novice pilot errors can be elicited in a safe AR environment and feedback provided before encountering the situation for the first time in flight. Finally, participants preferred the 3D AR learning content when asked, stating that they believed it would help them comprehend and visualize the weather better.

Use Case 3: VR astronaut training

Training Challenge. Training for workplace emergencies is essential to job preparedness to mitigate safety risks in complex, high-criticality domains. Emergency response training usually focuses on repetitive skill training (Thompson & Mccreary, 2006). Astronaut emergency procedures training, for instance, relies on the repetition of procedures with increasing complexity and usually requires considerable resources in facilities and supervision (Balmain & Fleming, 2009). However, even highly trained operators' performance can be negatively affected by the existential threat inherent in emergency situations (Orasanu & Backer, 2020).

Stress training presents several challenges that limit the ability to provide training. Stress training requires careful application and practice under conditions that approximate the operational environment (Driskell et al., 2008). Failure to introduce stressors into the training may result in the trainee developing a poor mental model that does not account for stress factors that impact performance. Introducing too much stress too early to an inexperienced trainee may result in learned helplessness (Keinan & Friedland, 1996). Thus, a high-fidelity training environment is needed that also allows for the level of stressors introduced to the task to be manipulated based on the current competency level of the trainee. Current stress training is typically supervised by a trainer/psychologist who relies on their experience to determine the appropriate stressor levels (Robson & Manacapilli, 2014). Training typically requires considerable physical and instructional resources.

VR-based approach. Finseth et al. (2018) developed a VR-based adaptive stress training system to enhance resilience to stress during emergency procedure training of astronauts. The VR simulation of the International Space Station (ISS) was developed to simulate a spaceflight emergency fire. Figure 2 illustrates a closed-loop, adaptive training system that automatically adapts to environmental stressors (e.g., smoke, alarms, flashing lights) while gradually increasing stressors over time (Finseth et al., 2021). However, to keep individuals in the optimal zone of stress dictated by the system, changes to the environmental stressors were personalized by adapting them based on a real-time stress prediction generated by machine learning algorithms that input psychophysiological responses.

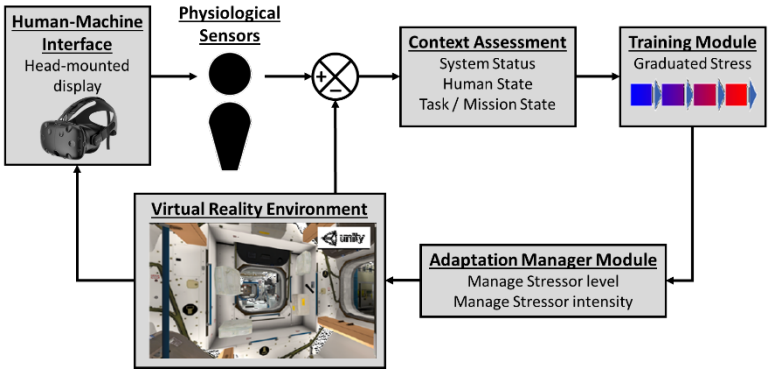


Figure 2. Conceptual diagram of the Adaptive Stress Training System.

Results/Benefits. Multiple studies were conducted to assess:(1) the effectiveness of the combination of graduated stress exposure in an interactive 3D VR environment to inoculate people against stress, (2) the ability of a procedure simulated in VR to be manipulated to evoke multiple levels of stress, and (3) the ability of a real-time physiology-driven VR adaptive system to enhance resilience to stress without degrading performance. In study 1, prior exposure to a high stress scenario enhanced relaxation behavior when confronted with a subsequent stressful condition in VR compared to a control group that did not have prior exposure to stressors when conducting training trials (Finseth et al., 2018). These results support the prior studies that graduated stress exposure enhances coping ability to acute stress. In study 2, varied combinations of stressors in a VR were able to induce differing levels of stress in participants (Finseth et al., 2022). The ability to use VR to elicit a multi-level subjective stress response in a predictable and controllable manner is a key requirement for graduated stress training. Finally, study 3 (Finseth et al., 2021) compared three groups of trainees: those exposed to training with no added stressors (skill-only), those exposed to a fixed schedule of increased stressors over time (graduated), and those who experienced stressor levels adapted to their current stress levels (adaptive). Stress was measured through subjective responses (stress, task engagement, worry, anxiety, and workload) and physiological responses (heart rate, heart rate variability, blood pressure, and electrodermal activity). Results suggest that all training conditions lowered stress, but the preponderance of trial effects for the adaptive condition suggests it is the most successful in decreasing stress over multiple trials.

The adaptive training system automatically adjusted the stressor levels based on a real-time measure of user stress. Combined with a graduated stress exposure pedagogy, this training system did not require the supervision of a trainer. In addition, utilizing VR to create a realistic training environment is the first step toward a mobile training system for use on the ground, during spaceflight, or on the Martian surface. Further work is needed to take the next step toward an adaptive ITS by adding personalized feedback and skill proficiency tracking.

Discussion

Workforce development can often be resource intensive, requiring dedicated trainers, facilities, and considerable time away from work. Utilizing MR can address challenges by bringing the operational context to the worker and training them in a virtual environment of appropriate fidelity. AR, furthermore, affords the possibility of training on the actual equipment. Previous work has shown that embedded VR-based Soldier training increased Soldier confidence because they could practice in environments that mirrored their preference for live training (Magee et al., 2011).

The three use cases described in the paper demonstrate the utility of both AR and VR in supporting knowledge acquisition and improving decision-making. The military has had several projects focusing on training Soldiers in virtual environments (Laviola et al., 2015). But the use of MR alone does not address the need for trainers to guide the learning process. Incorporating adaptive intelligent tutoring into MR learning approaches is a powerful way to lower the cost of workforce training. This may make it feasible to encourage more frequent training and practice to better support workforce development.

The first use cases compared engine parts recall training, using an AR-based app with parts, without parts, a video capture, and a standard lecture. Recall was found to be worst in the lecture case, but otherwise there were no statistically significant differences between the AR app and video capture. Suppose there are no differences in objective performance between a self-paced training app and a video. In that case, organizations can achieve substantial savings since videos tend to be more cost-effective to produce. However, this result might be a function of the types of knowledge being tested because no spatial or functional action knowledge was evaluated.

Further, anecdotally, the instructor who provided the training for all participants noted that subjects engaged with each other when interacting with the AR app, either with or without parts. The ability to control the 3D parts on their individual tablets and navigate to different perspectives allowed them to share different views at the same time by looking at each other's tablets. This was seen as an advantage over having one shared view where subjects had no agency in controlling their view. This may be especially useful in cases where employees with different roles receive training together. This is because each trainee can individually focus on the perspectives most relevant to their own tasks but also share that perspective with others to receive a more holistic view of the process.

The second use case demonstrated the utility of AR-based training of complex phenomena to support decision-making. Three-dimensional AR content has been associated with learning gains and motivation (Akçayır & Akçayır, 2017; Meister, Wang et al., 2022b). The application was a combination of user exploration and guided scenario-based learning activities. An ITS could proactively tailor feedback and differentiate instruction to address specific learner needs (Sottilare et al., 2017). The series of scenario-based learning activities chunked the learning experience into manageable amounts of information and scaffolded instruction. Scaffolded learning experiences begin with less complex cognitive activities and build up to more complex cognitive activities (Anderson & Krathwohl, 2001).

The third VR-based use case presented a VRE (virtual reality environment) that adapted automatically based on a real-time measure of worker stress. An ITS that is aware of the physical environment and objects can more effectively manage instructions (Laviola et al., 2015). Tracking the environment in VR is much easier than in a real environment and arguably affords a richer understanding of the learner's training context. Furthermore, the ability to use biosensors to derive the human cognitive and emotional state can strengthen the ITS model of the learner (Mathan & Dorneich, 2005). The Generalized Intelligent Framework for Tutoring (GIFT) has been configured to integrate with commercial off-the-shelf hardware and can utilize both VR technologies and external sensors (Heylmun et al., 2019)

Some limitations exist, such as the onset of cybersickness in VR. Cybersickness is the physical discomfort that can arise from VR experiences. It presents a barrier to the wide acceptance of the technology (Stanney et al., 2020). Our work has found a positive correlation between mental workload and cybersickness (Meusel, 2014) and a positive correlation between increased task complexity and cybersickness (Sepich et al., 2022). Performance was shown to be negatively correlated with cybersickness (Sepich et al., 2022), so it is important to consider how to mitigate cybersickness since training aims to develop high task performance.

Interactive learning in (safe) XR environments can provide a training environment that surpasses classroom based training toward more hands-on, scenario-based learning. Adding ITS components to these types of XR environments could automate some of the training and provide personalized feedback. XR also affords a collaborative experience with multiple users. Extending individual XR training to the multi-user training afforded by XR would allow multiple users to collaborate in a social setting. Users may provide different perspectives on the same object, which can provide opportunities for deeper exploration and collaborative learning.

Recommendations for GIFT and Intelligent Tutoring Systems

Use Case 1, in which trainees used AR to train on a jet engine, could have been handled by the current GIFT codebase. Unfortunately, no data was collected on how trainees used the AR, and they were assessed only by the multiple-choice knowledge retention test on factual knowledge. GIFT could have provided that assessment data. Moreover, not tracking trainees' engagement with the learning material (learning analytics) was a missed pedagogical opportunity. Previous researchers have successfully used patterns from learners' clickstreams within games or learning management systems to predict student learning and grades (Jayasekaran et al., 2022; Pal et al., 2022; Shute et al., 2021). If the AR environment logged user behaviors and sent them to GIFT, a domain module enhanced with "exploration pattern recognition" could detect exploration patterns authored by a trainer that indicate the trainee has not yet explored the complete engine, for example. Or, if the fragility of certain parts were emphasized in training, GIFT could detect a pattern of carelessness during exploration and remind the trainee to be more careful with parts.

Use Case 2 illustrates the importance of aiding learners in weighing the importance of different cues and sources of knowledge they have learned. The learner must be able to apply the theoretical knowledge they have learned to diagnose the situation they see in front of them. This challenge often arises in the applied domain of professional development rather than in traditional school-based learning. A person may have been trained on 12 different reasons an engine may not start, and there is an engine in front of the person that will not start. Will the learner be able to integrate the cues from previous knowledge and the cues from this particular engine? Analogously, a worker may have been trained on workplace ethics in an online course, and now a colleague is unintentionally asking the worker for confidential information. Has the worker learned the material enough to recognize the cues from the situation and apply relevant knowledge?

GIFT could be beneficial in ensuring that learners use all the knowledge cues available to them appropriately and reflect on which cues they are using and why. Several decades ago, a tool called the Diagnostic Pathfinder was created by Holly Bender and colleagues (Danielson et al., 2008), which demonstrated years of success at teaching medical problem-solving by enabling learners to cite particular observations as evidence for a diagnosis and then compare their diagnostic path with an expert's. This tool, like symbolic AI tools before it, such as EMYCIN (Bennett & Engelmore, 1984; Van Melle et al., 1984), offered the learner explicit feedback about the discrepancies between their actions and an expert's model. Some more recent ITSs only make that comparison implicitly within the domain module and offer direct feedback, but without explicitly documenting the expert model to the learner. The design decision of how explicitly to describe the expert model to the learner depends on the pedagogical context, just as a human

tutor might sometimes indicate only, "That's not quite right; try again," and other times offer a detailed explanation. Making two changes could enable GIFT to offer more detailed explanations and help trainees learn to balance their prioritization of knowledge cues. The first change would be to enable paths or patterns of action sequences (clickstreams) to be a unit of comparison (this change would also help with Use Case 1). The second change would be to visualize the comparison between the trainee's answer and the expert's answer rather than simply making that comparison behind the scenes.

Use Case 3 points out the importance of the learner's context when both learning and performing (in that use case, under high stress). As shown in several previous papers (e.g., Kim et al., 2018; Murphy et al., 2015), GIFT can integrate physiological signals into its Learner Module to help measure the learner's state and affective context. But this example allows us to consider what a GIFT-based measure of broader context might look like. When asking the question, "Why did a learner respond the way they did?" it could be helpful to look beyond the specific domain intricacies of the task at hand and examine the larger social context of the learner. What work-related pressures might the learner be experiencing in a workplace training setting? What social pressures might have influenced the learner's response in a school setting? Do gender- or race-based factors play a role, such as stereotype threat (Spencer et al., 2016)? GIFT could be enhanced to consider factors such as these. The discipline of human-computer interaction evolved from Wave 1, focusing on just the individual and what is in the head, to Waves 2 and 3, which consider the impact of broader social and cultural forces on people (Bødker, 2015). Could GIFT evolve similarly to take into account the organizational or cultural values of the learner? If a company highly valued evidence-based decision-making, that value could be translated to a learner skill of "Ability to justify decisions with evidence," perhaps clustered under a series of work performance skills that adhere to the organizational values. Or suppose the learner is known to hold personal cultural values (Blut et al., 2022) that focus more on their own needs than those of the entire work team. In that case, the pedagogical module within GIFT could perhaps have a conditional flag for that characteristic and offer feedback appropriately. Enabling GIFT to consider learners' organizational and cultural values would give it a significant advantage in offering motivating, personalized learning experiences across a broad range of professional development.

Conclusions

MR provides safe training environments for apprenticeship-style learning, whether the instructor is human or an intelligent agent, as in the case of ITSs. It allows trainees to be pre-exposed to different scenarios, working with expensive equipment rehearsing rarely-occurring processes to increase readiness. It affords maximum flexibility for either self-paced or structured lessons. For professional development, learning often occurs in group settings, where individuals may have different roles and learning objectives around the same use case. The spatial nature of MR not only allows individuals to take different perspectives and agency to control their views, but MR also offers the advantage of allowing trainees to take on the perspectives of others and even view a process or equipment in ways that are otherwise physically impossible. This may lead to a deeper understanding of the material and the different roles of their co-workers to improve overall teamwork. Further, MR can serve as a low-cost, fully instrumented environment, monitoring trainee performance and progress to provide more data for an ITS to calculate appropriate times for intervention. By analyzing learner behaviors, physiological state, and performance in the training environment, curricula can be adapted in real-time, increasing engagement, and reducing training time.

References

Adami, P., Rodrigues, P. B., Woods, P. J., Becerik-Gerber, B., Soibelman, L., Copur-Gencturk, Y., & Lucas, G. (2021). Effectiveness of VR-based training on improving construction workers' knowledge, skills, and safety

behavior in robotic teleoperation. Advanced Engineering Informatics, 50, 101431. https://doi.org/10.1016/J.AEI.2021.101431

Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. Educational Research Review, 20, 1–11. https://doi.org/10.1016/J.EDUREV.2016.11.002

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Pearson.

AOPA Air Safety Institute. (2018). 27th Joseph T. Nall Report: General Aviation Accidents in 2015. https://www.aopa.org/-/media/files/aopa/home/training-and-safety/nall-report/27thnallreport2018.pdf

Balmain, C., & Fleming, M. (2009). A methodology for training international space station crews to respond to on-orbit emergencies. International Conference On Environmental Systems. https://doi.org/10.4271/2009-01-2446

Bennett, J. S., & Engelmore, R. S. (1984). Experience Using EMYCIN. In B. G. Buchanan & E. H. Shortliffe (Eds.), Rule-Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project (pp. 314–328). Addison-Wesley Pub. Co.

Berendschot, Q., Ortiz, Y., Blickensderfer, B., Simonson, R., & DeFilippis, N. (2018). How to Improve General Aviation Weather Training: Challenges and Recommendations for Designing Computer-Based Simulation Weather Training Scenarios. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 3, 1792–1795. https://doi.org/10.1177/1541931218621406

Blut, M., Beatty, S. E., & Northington, W. M. (2022). Cultural personal values and switching costs perceptions: Beyond Hofstede. Journal of Business Research, 150, 339–353. https://doi.org/10.1016/J.JBUSRES.2022.06.005

Bødker, S. (2015). Third-Wave HCI, 10 Years Later-Participation and Sharing. Interactions, 24–31. https://doi.org/10.1145/2804405

Carney, T., Brown, L., Duncan, J., Whitehurst, G., Rantz, W., Seiler, B., & Mayes, P. (2015). Weather Technology in the Cockpit (WTIC) Project B: Unexpected Transition from VFR to IMC.

Carruth, D. W. (2017). Virtual Reality for Education and Workforce Training. 15th International Conference on Emerging ELearning Technologies and Applications, 1–6. https://doi.org/10.1109/ICETA.2017.8102472

Danielson, J. A., Mills, E.M., Vermeer, P.J., Bender, H.S. (2008). The Diagnostic Pathfinder: Ten Years Of Using Technology To Teach Diagnostic Problem Solving. In Leading Edge Educational Technology, Editors: Scott, T. B., Livingston, J. I., 71-103.

Dede, C. (1995). The evolution of constructivist learning environments: Immersion in distributed, virtual worlds. Educational Technology, 35(5), 46–52. https://doi.org/citeulike-article-id:1028204

Dede, C., Grotzer, T. A., Kamarainen, A., & Metcalf, S. (2017). EcoXPT: Designing for deeper learning through experimentation in an immersive virtual ecosystem. Journal of Educational Technology & Society, 20(4), 166-178.

Dorneich, M. C., & Jones, P. M. (2001). The UIUC Virtual Spectrometer: A Java-Based Collaborative Learning Environment. Journal of Engineering Education, 90(4), 713–720. https://doi.org/10.1002/j.2168-9830.2001.tb00663.x

Driskell, J. E., Salas, E., Johnston, J. H., & Wollert, T. N. (2008). Stress Exposure Training: An Event-Based Approach. In P. A. Hancock & J. L. Szalma (Eds.), Performance Under Stress (pp. 287–302). CRC Press. https://doi.org/10.1201/9781315599946-19

FAA. (2009). FITS Generic ADS-B, TIS-B and FIS-B Syllabus Version 1.0. https://www.faa.gov/training_testing/training/fits/training/media/generic/FIS_TIS_ADS.pdf

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. Human Factors, 54(6), 1008–1024. https://doi.org/10.1177/0018720812443983/ASSET/IMAGES/LARGE/10.1177_0018720812443983-FIG2.JPEG

Finseth, T., Dorneich, M. C., Keren, N., Franke, W. D., & Vardeman, S. B. (2022). Manipulating Stress Responses during Spaceflight Training with Virtual Stressors. Applied Sciences, 12(5), 2289. https://doi.org/10.3390/APP12052289

Finseth, T., Dorneich, M. C., Keren, N., Franke, W., Vardeman, S., Segal, J., Deick, A., Cavanah, E., & Thompson, K. (2021). The effectiveness of adaptive training for stress inoculation in a simu-lated astronaut task. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1541–1545. https://doi.org/10.1177/1071181321651241

Finseth, T., Keren, N., Dorneich, M. C., Franke, W. D., Anderson, C. C., & Shelley, M. C. (2018). Evaluating the Effectiveness of Graduated Stress Exposure in Virtual Spaceflight Hazard Training. Journal of Cognitive Engineering and Decision Making, 12(4), 248–268. https://doi.org/10.1177/1555343418775561/ASSET/IMAGES/LARGE/10.1177_1555343418775561-FIG2.JPEG

Heylmun, Z., Kalaf, M., Meyer, C., Padilla, C., & Woodman, L. (2019). Integrating Gift, Competencies, Virtual Reality, And Biometrics To Present Training Perspectives On Gauging Current Squad Capability. Proceedings of the 7th Annual GIFT Users Symposium, 157–163.

Jayasekaran, S. R., Anwar, S., Cho, K., & Ali, S. F. (2022). Relationship of Students' Engagement with Learning Management System and their Performance-An Undergraduate Programming Course Perspective. 2022 ASEE Annual Conference & Exposition.

Johnson, I., Whitehurst, G., Risukhin, V. N., Brown, L. J., Rantz, W., Ferris, T. K., Roady, T., Rodriguez-Paras, C., Tippey, K., & Futrell, M. J. (2017). PEGASAS: Weather Technology in the Cockpit. International Symposium on Aviation Psychology, 323–328.

Katz, S., Lesgold, A., Hughes, E., Peters, D., Eggan, G., Gordin, M., & Greenberg, L. (2020). Sherlock 2: An Intelligent Tutoring System Built on the LRDC Tutor Framework. In Facilitating the Development and Use of Interactive Learning Environments (pp. 227–258). CRC Press. https://doi.org/10.1201/9780367813512-13

Keinan, G., & Friedland, N. (1996). Training Effective Performance Under Stress: Queries, Dilemmas, and Possible Solutions: In Stress and Human Performance (pp. 266–286). Psychology Press. https://doi.org/10.4324/9780203772904-16

Kim, J. W., Sottilare, R. A., Brawner, K., & Flowers, T. (2018). Integrating Sensors and Exploiting Sensor Data with GIFT for Improved Learning Analytics. Sixth Annual GIFT Users Symposium.

Laviola, J. J., Williamson, B. M., Brooks, C., Veazanchin, S., Sottilare, R., & Garrity, P. (2015). Using Augmented Reality to Tutor Military Tasks in the Wild. Proceedings of the Interservice/Industry Training Simulation & Education Conference.

Lerman, R. I., Loprest, P. J., & Kuehn, D. (2020). Training for Jobs of the Future: Improving Access, Certifying Skills, and Expanding Apprenticeship (No. 166; IZA Policy Paper). http://hdl.handle.net/10419/243452

Magee, L., Sottilare, R., & Roessingh, J. J. (2011). Human Interaction in Embedded Virtual Simulations. Proceedings of the Interservice/Industry Training Simulation & Education Conference. https://doi.org/10.13140/2.1.1848.3201

Major, W. L., Carney, T., Keller, J., Xie, A., Price, M., Duncan, J., Brown, L., Whitehurst, G. R., Rantz, W. G., Nicolai, D., & Beaudin-Seiler, B. M. (2017). VFR-into-IMC Accident Trends: Perceptions of Deficiencies in Training. Journal of Aviation Technology and Engineering, 7(1), 4. https://doi.org/10.7771/2159-6670.1153

Mathan, S., & Dorneich, M. C. (2005). Augmented Tutoring: Enhancing Simulation Based Training through Model Tracing and Real-Time Neurophysiological Sensing. In D. D. Schmorrow (Ed.), Foundations of Augmented Cognition (pp. 964–973). Lawrence Erlbaum Associates, Inc. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.725.4633&rep=rep1&type=pdf

Meister, P., Miller, J., Wang, K., Dorneich, M. C., Winer, E., Brown, L. J., & Whitehurst, G. (2022). Designing Three-Dimensional Augmented Reality Weather Visualizations to Enhance General Aviation Weather Education. IEEE Transactions on Professional Communication, 65(2), 321–336. https://doi.org/10.1109/TPC.2022.3155920

Meister, P., Wang, K., Dorneich, M. C., Winer, E., Brown, L., & Whitehurst, G. (2022a). Evaluation of the Effectiveness of Augmented Reality Enhanced Aviation Weather Training. AIAA AVIATION 2022 FORUM, 3780. https://doi.org/10.2514/6.2022-3780

Meister, P., Wang, K., Dorneich, M. C., Winer, E., Brown, L., & Whitehurst, G. (2022b). Augmented Reality Enhanced Thunderstorm Learning Experiences for General Aviation. Journal of Air Transportation, 1–11. https://doi.org/10.2514/1.D0308

Meusel, C. R. (2014). Exploring mental effort and nausea via electrodermal activity within scenario-based tasks (Doctoral dissertation, Iowa State University).

Mossel, A., Peer, A., Goellner, J., & Kaufmann, H. (2015). Requirements Analysis on a Virtual Reality Training System for CBRN Crisis Preparedness. Proceedings of the 59th Annual Meeting of the ISSS - 2015 Berlin, Germany, 1(1). https://journals.isss.org/index.php/proceedings59th/article/view/2486

Murphy, J. S., Carroll, M. B., Champney, R. K., & Padron, C. K. (2015). Investigating the Role of Physiological Measurement in Intelligent Tutoring. Proceedings of the Second Annual GIFT Users Symposium, 105–114.

https://books.google.com/books?hl=en&lr=&id=M62MBgAAQBAJ&oi=fnd&pg=PA105&dq=Investigating+the+Role+of+Physio-logical+Measurement+in+Intelligent+Tutoring.+&ots=JRJPrFw_t1&sig=on9rQd0wP1hlanujNf_wSk-cKto#v=onepage&q=Investigating the Role of Physio-logical

Orasanu, J. M., & Backer, P. (2020). Stress and Military Performance. In Stress and Human Performance (pp. 100–136). Psychology Press. https://doi.org/10.4324/9780203772904-10/STRESS-MILITARY-PERFORMANCE-JUDITH-ORASANU-PATRICIA-BACKER-JAMES-DRISKELL-EDUARDO-SALAS

Pal, S., Ngampornchai, A., & Moskal, P. (2022). Examining the Impact of Online Lecture Viewing Behavior on Student Performance in a Flipped Classroom Blended Course. 2022 ASEE Annual Conference & Exposition.

Pan, Z., Cheok, A. D., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. Computers & Graphics, 30(1), 20–28. https://doi.org/10.1016/J.CAG.2005.10.004

Robson, S., & Manacapilli, T. (2014). Enhancing Performance Under Stress: Stress Inoculation Training for Battlefield Airmen. https://apps.dtic.mil/sti/pdfs/ADA605157.pdf

Sepich, N. C., Jasper, A., Fieffer, S., Gilbert, S. B., Dorneich, M. C., & Kelly, J. W. (2022). The impact of task workload on cybersickness. Frontiers in Virtual Reality, 3, 943409.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. In Journal of Computer Assisted Learning (Vol. 37, Issue 1, pp. 127–141). John Wiley & Sons, Ltd. https://doi.org/10.1111/JCAL.12473

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). www.GIFTtutoring.org

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. Annual Review of Psychology, 67(1), 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Stanney, K., Lawson, B. D., Rokers, B., Dennison, M., Fidopiastis, C., Stoffregen, T., Weech, S., & Fulvio, J. M. (2020). Identifying Causes of and Solutions for Cybersickness in Immersive Technology: Reformulation of a Research and Development Agenda. International Journal of Human–Computer Interaction, 36(19), 1783–1803. https://doi.org/10.1080/10447318.2020.1828535

Thompson, M. M., & Mccreary, D. R. (2006). Enhancing Mental Readiness in Military Personnel. https://apps.dtic.mil/sti/pdfs/ADA472674.pdf

Uhlig, T., Roshani, F. C., Amodio, C., Rovera, A., Zekusic, N., Helmholz, H., & Fairchild, M. (2016). ISS emergency scenarios and a virtual training simulator for Flight Controllers. *Acta Astronautica, 128*, 513-520. https://doi.org/10.1016/j.actaastro.2016.08.001

Van Melle, W., Shortliffe, E. H., & Buchanan, B. G. (1984). EMYCIN: A Knowledge Engineer's Tool for Constructing Rule-Based Expert Systems. In B. G. Buchanan & E. H. Shortliffe (Eds.), Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project (pp. 302–313). Addison-Wesley Pub. Co.

CHAPTER 9 - ADAPTIVE LEARNING CONSIDERATIONS FOR COMMERCIAL PILOT PIPELINE

Elizabeth Biddle
The Boeing Company

Introduction

Pilots typically enter the commercial transport aviation industry (e.g., passenger airlines or cargo transport carrier) after gaining initial instruction and experience through a military pilot career or regional commercial airline pathway. Consequently, commercial airline or freighter pilot training primarily focuses on type rating and recurrent training. Type rating training is a longer course, typically around 30 days, to orient and introduce a pilot to the systems and performance characteristics of an aircraft platform type required for a pilot to operate an aircraft type (e.g., Boeing 767). There are shorter curriculums that a pilot can complete to obtain a type rating certification depending on the similarity between aircraft types and a pilot's previous type rating qualification and experience. Recurrent training is an annual 2-3 day curriculum, which concentrates on non-normal events and manual flight skills with an operational emphasis that reflects safety concerns identified in the company or industry. The majority of commercial pilot training is performed at a flight training facility using different levels of flight simulators. Off-campus training is provided through online, computer-based training (CBT), sometimes called distance learning. CBT is used to teach the fundamental knowledge of systems and procedures specific to the aircraft type prior to the start of the simulator training events. While the number of flight hours required to enroll in a type rating and to fly commercial transport aircraft varies by regulator, this training assumes the pilot knows how to fly a transport category jet.

The commercial aviation industry expects 620,000 new pilots to enter the pilot workforce over the next 20 years (Boeing Pilot and Technical Outlook 2022-2041, 2022). Given the numbers of new First Officers entering the workforce versus the number of First Officers ready to transition to Captain, acceleration of the transition from First Officer to Captain is needed. Passenger and cargo carriers are building up early career programs focused on accelerating the development of pilot competencies and entry into the carrier's pilot operations. In the United States, many carriers have created entry-level pilot training programs to attract potential pilots and expedite the pilot training and experience pathway to airline operations. While these programs are helping to address the pilot pipeline issue, additional capabilities are needed to increase the training throughput and expedite time to mastery.

During the recent COVID-19 pandemic, numerous pilots were furloughed or left the airline industry due to downsizing and/or personal reasons. As the aviation travel industry rebounds from the pandemic to greater than pre-pandemic numbers of passengers flying domestically, resource availability to support training for airline pilots is further stretched. Given the limited number of costly, resource intensive flight simulators and a similar shortage of instructor pilots, the commercial aviation industry is seeking training solutions to provide instruction outside of the traditional brick and mortar flight schools. Finally, the pilot pipeline issue coupled with the current commercial pilot training demand has resulted in a variability in the global pilot population demographics – in terms of flight experience, English language proficiency, and culture – and is necessitating flexibility in training curriculums to address a range of learning needs.

Goals and Scope

Traditional brick and mortar flight schools are throughput limited in terms of number of simulators and instructor pilots available to support training. The pilot training bottleneck needs to be mitigated through

the implementation of training solutions accessed away from a training campus and without real-time oversight of an instructor pilot. Immersive learning and adaptive instructional systems (AISs) have the potential to expand the availability and reach of flight training and provide tailored instruction to address the diverse learning needs of the global pilot population. AISs, including intelligent tutoring systems (ITSs), are well-suited to distributed learning environments and provide real-time training performance feedback and learning recommendations. Results of AIS learning events can be accessed by the instructor pilot to prepare for and optimize the flight training events on campus. Commercial pilot training trends, the variability in the global pilot population and implications to training and recommendations on the application of AISs and ITSs for commercial pilot training will be discussed.

State of the Field and Supporting Research

Commercial pilot training is largely focused on grading specific pilot tasks and maneuvers. The training duration and flight simulator fidelity requirements are strictly regulated. The continued introduction of new technologies in the flight deck and air traffic management system, coupled with the pilot pipeline issue, has impacted flight operations and pilot training needs. A trend in the commercial pilot training industry is the implementation of evidence-based training (EBT; International Air Transport Association [IATA], 2021) that uses flight operations data to identify training needs for certain types of tasks, procedures and/or maneuvers. Instead of training the same tasks and maneuvers year after year, EBT recommends that the specific tasks and maneuvers performed during recurrent training be updated based on the trends in the flight operations data. EBT is a competency-based training and assessment (CBTA) program. CBTA is also expanding training to focus on non-technical competencies such as leadership and teamwork. EBT/CBTA is required for recurrent pilot training in airspace governed by the European Union Aviation Safety Agency (EASA) with additional regulators moving towards similar requirements.

Commercial Pilot CBTA

CBTA methodology for commercial pilot training was initially specified by the International Civil Aviation Organization (ICAO; 2013) intended for use in annual recurrent training. ICAO's CBTA methodology (2013) identifies 8 competencies. EASA added a ninth competency (*noted in the below bulleted list). The 9 competencies are defined with 7-11 observable behaviors that describe overt actions that an instructor pilot may witness during a training session to infer proficiency of a specific competency. The most current nomenclature for the commercial pilot competencies (IATA, 2021) is below:

- Application of Procedures and Compliance with Regulations
- Airplane Flight Path Management, Automation
- Airplane Flight Path Management, Manual Control
- Communication
- Situation Awareness and Management of Information
- Leadership & Teamwork
- Workload Management
- Problem Solving & Decision Making
- Application of Knowledge (*EASA addition to ICAO competencies)

EASA is the first regulator to require EBT/CBTA training program implementation. The evolution to EBT/CBTA implementation takes time, investment and change in standard practices. The initial transition to EBT/CBTA includes the incorporation of CBTA principles and methodologies into the pilot and instructor training and assessment programs. Industry data collection and analysis of flight operation is used during this phase to identify focus areas for recurrent training. The initial ICAO guidance recommended

procedures and maneuvers based on analysis of data collected from available flight operations data from a number of carriers. The recommended focus areas were recently updated based on analysis of more recent data (IATA, 2021). As the CBTA methods mature, the flight organization implements their internal operational data collection and analysis process to understand the organization's flight operation issues that can and should be addressed by recurrent pilot training.

The EBT/CBTA recurrent training high-level footprint involves an evaluation phase in which all 9 competencies are assessed to identify areas to emphasize in the remaining phases involving maneuvers, specialized procedures and scenario-based training sessions. Instructor pilots are taught to use the observable behaviors demonstrated during training as evidence to grade competencies, and in some cases, the instructor pilots may grade observable behaviors (IATA, 2021). The current CBTA guidance (IATA, 2021) still affords instructor pilots to use any of the observable behaviors they witness during the training session as evidence to substantiate the grades they assign at the end of a training session. The recommended grading approach is the implementation of a 5-point scale using the following attributes to create "word pictures" to assist the instructors with assigning grades. The "word pictures" include the following four elements (IATA, 2013):

1. Level of proficiency observed (e.g., the pilot did.., the pilot did not…)
2. How often the competency was demonstrated (e.g., very often, rarely)
3. The number of observable behaviors demonstrated
4. The outcome of the demonstrated behavior (e.g., safe landing)

The event based approach to training (EBAT; Fowlkes et al., 1998) used extensively in military training supports the design of scenario-based lessons to elicit performance of behaviors to demonstrate the competencies targeted by the lesson. This structured approach to scenario design with targeted performance assessment has been applied to competency-based training (Johnston et al., 2022). The EBAT methodology provides a means of consistency in the assessment of CBTA training by having the same events assessed across students. This structured assessment methodology reduces instructor workload by directing the instructor's attention to specific events rather than having to be on the constant lookout for potentially relevant behaviors.

The commercial aviation training community can leverage insights and development from other industries. The EBAT methodology discussed previously can help provide standardization of training and assessment methods. Given pilots will move between flight schools and potentially airlines during their flight career, standardized assessment methods will enable tailored training throughout their career. Further, the prescriptive nature of the EBAT assessment methodology is well-suited for implementation with ITSs and AISs.

Discussion

There is interest in the commercial aviation training community to integrate AISs with desktop and immersive flight simulations to reduce time required in large, fixed devices at a flight training campus and alleviate the training pipeline demands. ICAO is developing recommendations to update regulations regarding the use of digital learning solutions for portions of recurrent and type rating curriculums. ITS and AIS capabilities can expand digital learning capabilities with tailored learning experiences to support some of the known pilot demographic variables that impact training needs – specifically flight experience, English language proficiency, nation culture, and organizational culture.

Flight Experience and Learning Needs

Commercial pilots have a range of prior flight experience in terms of flight hours ranging from 200 to 10,000+ flight hours, yet the training curriculum is the same – regardless of experience levels. Recurrent and type rating training are designed with the assumption that the pilot has demonstrated proficiency in flying transport category aircraft, resulting in the training that challenges pilots with lower flight hours in a transport category aircraft. The less experienced pilots may require supplemental training on automated and manual flights to successfully complete the lesson objectives. Type rating, which has a demanding pace, often induces high workload and fatigue with less experienced pilots. AISs and ITSs are well-suited for optimizing the pace of learning for each pilot. Blended learning concepts in which the typical upfront loaded systems and procedure content, which can be 40 hours or more for a type rating course, can be implemented seamlessly with ITSs and AISs. Blended learning concepts such as chunking initial systems and procedural knowledge into short learning modules with opportunities to apply the knowledge learned with practical exercises will help promote learning and retention of distance learning content. Lesson spacing optimized to the pilot's workload will help reduce overload and fatigue.

Pilot English Language Proficiency

Roughly 2/3 of the people in the world who speak English are non-native English speakers, and consequently, English language proficiency remains a safety consideration in commercial aviation. Due to the number of aviation accidents attributed to English language communication related issues, ICAO released recommendations for English language proficiency training and assessment standards (ICAO, 2010) that were initially published in 2003 with a requirement for licensing compliance within 5 years (Fowler et al., 2021). The ICAO English language proficiency requirements identify 6 levels of English language proficiency, with a level 4 proficiency the minimum requirement for flight operations and related training. The level 4 proficiency is for verbal radiotelephony communications between aircraft and air traffic control (ATC). The ICAO standards are based on strict adherence to standard phraseology, pronunciation of numbers, and certain aviation terms per specification in the ICAO (2010) standard and assumes a rate of no more than 100 words per minute.

The ICAO language proficiency training and assessment standards are not regulated worldwide, and there is growing skepticism on the standard improving communications and increasing safety (Clark & Williams, 2020). Assessments are required for the non-native English speakers only, even though the phraseology, pronunciation and rate of speech required apply to native English speakers. Considering implications for a non-native English speaking pilot attempting a type or recurrent training session, the typical English language capabilities required for reading instructional materials, completing distance learning and classroom lessons and interactions with a native English speaking instructor pilot are not supported (Fowler et al., 2021). Most pilot training materials necessitating English language proficiency in reading, writing and comprehension far exceed the ICAO English language proficiency standards, leading to the potential for lack of understanding of the training content. Even when the non-native English speaking pilot's English is of an adequate level of proficiency for reading, writing and comprehension to understand the training materials, there is still an intensive workload demand (Farris et al., 2008). Adaptive learning strategies that can help assist the non-native English speaker through translations, simple phraseology and reduced rates of speech could be implemented with ITSs and AISs.

Cultural Considerations for Pilot Training

Culture – regionally, nationally and within organizations have impacts to performance in operations and needs for training. Nation culture influences attitudes related to flight crew interactions (Engle, 2000; Helmreich, 1984), which pose safety issues. The dimensions of power distance, uncertainty avoidance, and individualism contribute to the effectiveness of flight crew interactions. Cultures with higher power distance tend to place more emphasis on the deference of the First Officer to the Captain, while cultures high on individualism tend to feel less inclined to include their crew in decision-making. Cultural influences in flight deck operations and training have been attributed to the organization (airline) culture (Dahlstrom & Heestra, 2009). Airline culture on attitudes towards safety and use of automation effect pilots' attitude toward their airline and flight performance (Owen, 2013; Sexton et al., 2001). AISs and ITSs may be useful for identifying learning needs due to cultural issues and tailoring feedback recommendations accordingly.

Recommendations for GIFT and Intelligent Tutoring Systems

There are opportunities to leverage the Generalized Intelligent Framework for Tutoring (GIFT) to conduct the fundamental research regarding methods of implementing ITSs to address the pilot demographic concerns discussed in the prior section. There is little research on specific instructional methods in the context of commercial pilot training tailored to the pilot's prior flight experience, English language proficiency, or cultural needs. GIFT's data collection capabilities coupled with the ability to rapidly prototype ITS concepts provide a valuable experimental resource for such research. GIFT is also well-suited for the further development of fine-grained, quantifiable performance assessment of competencies to assist in the diagnosis of pilot learning needs.

Conclusions

The growth in the commercial aviation industry coupled with the diverging pilot demographic is creating a tremendous stress to current training resources and methods. AISs and ITSs integrated with simulation-based and immersive learning environments can increase training throughput and provide a means of addressing the pilot's learning needs outside of a traditional training campus. Improvements to off-campus learning will better prepare the pilots and increase the value of in-person training.

Finally, the development of quantifiable competency assessment methods will benefit from the adoption of learning data standards. The challenge in developing these methods will be the ability to collect learning effectiveness data from representative pilot populations.

Acknowledgements

Guidance regarding commercial aviation training regulations provided by Richard Caldwell, The Boeing Company.

References

Clark, L., & Williams, G. (2020). English Language Proficiency in Radiotelephony: A survey about its effect on the safety and efficiency of aviation. *The Especialist*. 41. 10.23925/2318-7115.2020v41i4a9.

Dahlstrom, N., & Heemstra, L. (2009). Beyond multi-culture: When increasing diversity dissolves differences. In (Editors: Strohschneider S. & Heimann R.), *Kultur and sicheres handeln*, pp.79-95. Verlag fur Polizeiwissenschaft.

Engle, M. (2000). Culture in the cockpit – CRM in a multicultural world. *Journal of Air Transportation World Wide,* Vol. 5, No. 1, 107-114.

Farris, C., Trofimovich, P., Segalowitz, N., & Gatbonton, E. (2008). Air Traffic Communication in a Second Language: Implications of Cognitive Factors for Training and Assessment. https://doi.org/10.1002/j.1545-7249.2008.tb00138.x

Fowler, R., Matthews, E., Lynch, J., & Roberts, J., (2021). Aviation English Assessment and Training. Collegiate Aviation Review International, 39(2), 26-42. http://ojs.library.okstate.edu/osu/index.php/CARI/article/view/8216/7644

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). The *International Journal of Aviation Psychology*, 8(3), 209–221. https://doi.org/10.1207/s15327108ijap0803_3

Helmreich, R. L. (1984). Cockpit management attitudes. *Human Factors*, 26(5), 583-589.

IATA (2013). *Evidence-Based Training Implementation Guide: 1st Edition*.

IATA. (2021). *Data Report for Evidence-Based Training.*

ICAO. (2010). *Manual on the Implementation of ICAO Language Proficiency Requirements.* Document 9835.

ICAO. (2013). *Manual of Evidence-Based Training*. Document 9995.

Johnston, J. H., Sottilare, R. A., Kalaf, M., & Goodwin, G. (2022). Chapter 8 – Training for Team Effectiveness Under Stress. In, Sinatra, A. M., Graesser, A. C., Hu, X., Goldberg, B., Hampton, A. J., and Johnston, J. H. (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 9 - Competency-Based Scenario Design*. Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center. https://gifttutoring.org/documents/

Owen, E. L. (2013). Assessing the status of airline safety culture and its relationship to key employee attitudes (Master's Thesis). Middle Tennessee University.

Sexton, J. B., Wilhelm, J.A., Helmreich, R.L., Merritt, A.C., & Klinect, J.R. (2001). *Flight Management Attitudes & Safety Survey (FMASS): A Short Version* (Technical Report: 01-01). University of Texas at Austin Human Factors Research Project.

The Boeing Company (2022). Boeing Pilot and Technical Outlook 2022-2041. https://www.boeing.com/commercial/market/pilot-technician-outlook/

CHAPTER 10 – CONSIDERATIONS IN CONSTRUCTING AN INTELLIGENT TUTORING SYSTEM FOR SENSITIVE TOPICS: ADAPTING THE PAL3 FRAMEWORK FOR SUICIDE PREVENTION TRAINING

William Swartout, Benjamin Nye, and Albert (Skip) Rizzo
USC Institute for Creative Technologies

Introduction

Constructing an intelligent system for training an academic topic such as physics or algebra is formidable but building a training system for sensitive topics such as suicide prevention, where users may be emotionally involved, is even more challenging. Some of the issues include privacy — users may not want to share sensitive information if they think it may be shared with superiors or others; adaptation — people may be motivated to get suicide prevention training for a variety of reasons, such as helping themselves or helping someone else, which means that the training will be most effective if it is tuned to the needs of a particular user; interaction tone — a matter-of-fact interaction style may be very appropriate for academic content, but a more sensitive, and non-stigmatizing tone for interaction may be needed for sensitive topics; and finally, availability — users need easy access to content so that it may be used if a crisis occurs. We sought to use the Personal Assistant for Lifelong Learning (PAL3) framework to build a training system for suicide prevention. PAL3 already had some of the desired capabilities, such as availability, since it runs on iOS and Android smartphones, but a number of additional enhancements were required. In this chapter we briefly discuss the suicide problem in the military, outline the PAL3 framework, and discuss enhancements we made to the PAL3 framework as we developed a system for suicide prevention training.

Background

Based on recent Centers for Disease Control statistics, the occurrence of suicide in the United States has become a serious public health crisis (Centers for Disease Control and Prevention, 2021). Within the general civilian population, many more Americans die by suicide than homicide. While homicide is the 16th leading cause of death, suicide ranks as 10th, with one American dying by suicide every 11-minutes (Drapeau & McIntosh, 2020). Moreover, suicides have been rising nationally in the United States since 1999 with half of the states seeing more than a 30% increase in suicide rates from 1999 to 2016 (National Center for Injury Prevention and Control, 2018). Suicide is also one of the leading causes of death among young people. In 15–24 year-olds it is the third leading cause of death and ranks 2nd in the 25-34 year-old cohort.

These numbers are particularly concerning when considering young service members in the military. Despite unprecedented suicide prevention efforts undertaken in the United States Department of Defense (DoD), suicide rates among military service members remain elevated relative to the pre-9/11 era. Suicide is the 2nd leading cause of death in the military (Armed Forces Health Surveillance Center (AFHSC), 2012). The most recently reported suicide rate for active-duty military was 25.9 deaths per 100,000 population (Tucker et al., n.d.). There has been a per-year increase in the suicide mortality rate ratio (RR) since 2011 among active-duty service members (per-year RR=1.04; CI=1.02-1.05). Despite advantages in access to health care, mental health care, employment, and exercise compared to the general population, service members experienced equivalent increases in suicide rates compared to the US population

(Tucker et al., n.d.). Furthermore, while the most recent suicide rate for active-duty service members is similar to the rate in the US general population, military rates observed in recent years differ dramatically from decades of historic trends where military suicide rates were consistently much lower than the general population (Eaton et al., 2006). For example, from 1990 to 2000, the US military suicide rates were 11.82 to 12.98 per 100,000 population, 25-to-33% lower than the US civilian population (Eaton et al., 2006). This is like many decades of prior military research (Eaton et al., 2006; Rothberg & Jones, 1987). In the Department of Veterans Affairs , suicide prevention is also a top clinical priority. U.S. Veteran suicide rates have also been rising in recent years, and the Veteran suicide rate is currently 1.5 times the rate of the non-Veteran US population (U.S. Department of Veterans Affairs, 2020). Thus, the need for improved suicide prevention practices in the military has become an issue of critical concern. While there are many pressing medical and mental health matters to address among Service Members and Veterans, suicide prevention is a top priority for the DoD.

State of the Field and Supporting Research

To address this priority, the DoD and Veterans Administration (VA) have implemented a variety of classroom/web-based programs that have primarily focused on training leaders and clinical care providers in strategies for better recognizing the signs of suicide risk and in the provision of interventions to their at-risk subordinates or patients. For example, the Ask, Care, Escort Suicide Intervention (ACE-SI) has been the gatekeeper component of the Army's suicide intervention strategy (U.S. Department of the Army, 2015). The primary goal of this program has been to train Army leaders E6 and above to identify peers at risk for suicide and safely accompany them to a helping resource. ACE-SI aims to challenge leaders to engage using Motivational Interviewing skills (Ask), offer support and assistance through common factors strategies (Care), and safely implement supportive action by accompanying them or directing them to the appropriate helping resource (Escort).

The Navy Leader's Guide for Managing Sailors in Distress (Navy Medicine, 2021) provides Navy leaders with psychoeducational materials that address mental health and wellness and includes a module on suicide prevention. The VA's Safety Planning Intervention is designed as a brief clinical intervention that healthcare providers can implement with Veterans at risk for suicide. At risk patients are identified as those who may have made a suicide attempt or engaged in other types of suicidal behavior, reported suicidal ideation, have psychiatric disorders that increase suicide risk, or who are otherwise determined to be at risk for suicide (Stanley & Brown, 2012). This approach teaches clinicians how to conduct a structured interview that aims to help patients identify their emotional warning signs or triggers and to formalize a plan of action (or behavioral contract) for reducing their subsequent suicide risk (i.e., identify internal coping strategies, specify social, family, friend, and professional contacts, and in the encouragement of harm reductions strategies). These programs represent a strong effort to teach leaders and healthcare key principles for recognizing and supporting those at risk for suicide. However, complimentary strategies are needed to provide service members and veterans similar psychoeducational knowledge, self-awareness, and suicide prevention tactics directly.

In the past, suicide prevention training was delivered mainly as a group lecture in a classroom. There are several problems with this approach that our work seeks to overcome. The lecture setting necessitates a one-size-fits-all approach to content. The group setting makes it very difficult to adapt training in response to individual learner needs or motivations. Classroom training is delivered periodically, which means it may not be available outside of class or when it is most needed. Finally, the classroom setting may discourage students from asking questions about sensitive topics or revealing their concerns.

PAL3 Framework

The PAL3 framework (Swartout et al., 2016) was designed to provide learners with an adaptive, always available learning environment to promote learning outside of the classroom. The design of PAL3 follows four core principles:

- *Useful Learning*: Recommend learning content that is relevant to the learner's goals and needs.
- *Personalized Learning*: By analyzing learning pathways, recommend topics and lessons that maximize learning rates and mitigate skill decay.
- *Engaged Learning*: Leverage techniques from the learning sciences, games, and social media to create engagement and learning over time, even when between traditional classes and training.
- *On-Demand Learning*: By leveraging mobile learning (e.g., smartphones), content is always with a learner, whenever and wherever they are, including making content available when offline.

An overview of the PAL3 framework is shown in Figure 1. The Learning Record, built on the Veracity Learner Record Store (LRS) framework (lrs.io), stores learners' past training experiences and how they did, their mastery of relevant topics, and their goals. The Resource Library holds a variety of different types of learning resources. These can include HTML websites, videos, models and simulations, interactive computer tutors, and even other apps. PAL3 can make use of a broad array of existing resources. In most cases it is not necessary to create special content for PAL3. To add content to the Resource Library and make it usable by PAL3, usually all that is required is to add metatags to the content indicating how much active exploration the resource involves (further described in the next section) and what knowledge components the resource can help a user learn. These metatags are used by the recommender, described below. Because PAL3 may need to be used in situations where online connectivity is not available, PAL3 can download and cache resources for offline use, including local versions of resources (e.g., videos, static web pages, quizzes, tutoring dialogs).

Two of PAL3's core capabilities are the Recommender and its Engagement mechanisms.

Recommender for Lessons

The Recommender uses the information in the Learning Record to adaptively recommend learning exercises to the user. Recommendations are based on three factors, which each require an increasing amount of information about resources to apply.

1. *Novelty:* The recommender prefers resources that the learner has not already seen, which is done by calculating a familiarity estimate based on the number of exposures to the resource so-far. The novelty factor requires no metadata about a lesson, enabling limited adaptivity even with arbitrary resources.
2. *Exploration:* The second factor is how much active learning and degrees of freedom the learner needs to benefit from the resource. The exploration factor is a single number, representing a continuum meant to represent distinctions such as Passive / Active / Interactive / Constructive (Chi & Wylie, 2014). Passive resources such as videos or simple web articles are assigned the lowest exploration level. More open-ended have high exploration levels, such as interactive simulations or model construction, where the user has a large space of options or complexity to manage. If a learner's mastery of a topic is low, the recommender prefers passive resources with more

knowledge components (overviews) and low-exploration active resources with fewer knowledge components, while more active resources will be recommended for those with greater mastery.

3. *Deficits:* Learning resources are tagged with the knowledge components (KC) (Aleven & Koedinger, 2013) they can address, while the Learning Record expresses mastery in terms of KCs. Learning resources that can address specific learner deficits are preferred.

The learner is presented with the recommended resources and is free to either follow the recommendation or navigate to some other learning resource. Users engage with their selected resources, resulting scores are recorded in the learning record and the whole process iterates.

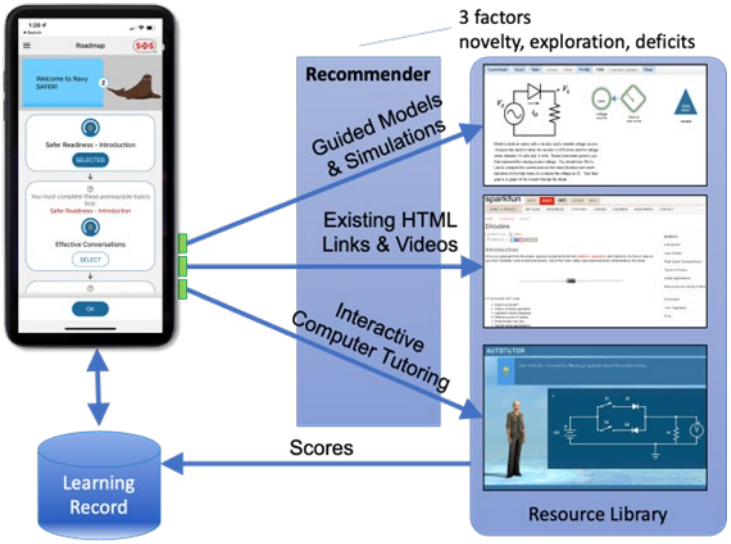


Figure 1. PAL3

Engagement Mechanisms

The engagement mechanisms for the framework follow three primary types: mastery learning, effort self-regulation, and social learning. These align directly to the three main panels: Goals, Study Pace, and Teams. The Goals panel provides an open learner model showing the learner's progress toward mastering the current topic and their larger goal. Open learner models help with metacognition about learning and skill levels, which have been shown to produce engagement and learning gains (Long & Aleven, 2017). The Study Pace panel enables learners to set a target for daily study time, to help them regulate their study pace over the week. This is inspired by fitness tracking apps, such as Fitbit step tracking. The Teams area allows learners to opt-in to a team, which competes against other teams for king-of-the-hill for each topic. This enables a collaborate/compete dynamic, where members within a team are incentivized to support each other's learning while competing against other teams. This structure is central to many social games and social media programs, which produce sustained engagement (Pirker et al., 2018; Shonfeld & Resta, n.d.).

PAL3 Evaluations

We have conducted two controlled evaluations of the PAL3 framework. The first study, which used Microsoft Surfaces rather than smartphones, showed that PAL3 significantly reduced knowledge decay among Sailors about electronics (Hampton et al., 2018). The second controlled study showed significant learning gains in leadership knowledge with junior Navy officers (16% gain from pre to post; *N*=24;

$p<0.001$; effect size 0.76). Topics studied included communication and counseling, leadership, making adjustments for moves and family life, and initial content for suicide bystander training.

Approach

Individuals at risk of suicide will not always be identified if we depend exclusively on traditional in-person mental health clinic visits. This is underscored by findings in a recent 2021 review (Tang et al., 2021) that found that "…the majority of people who die by suicide have never seen a mental health professional or been diagnosed with a mental illness." The authors go on to suggest that online interventions, including mobile apps and online psychotherapy appointments, having shown preliminary success, may be a useful option for reducing suicidal ideation and for breaking down barriers to treatment such as physical distance and stigma. Moreover, suicidal actions often involve stressors and losses that add to long-building stress (Ho et al., 2018). As a result, interventions to strengthen protective factors and reduce vulnerabilities to high stress can reduce risk of suicide and other destructive behaviors.

To address this issue, the PAL3 framework was adapted to develop SAFER, the Safety Assistant for Excellence and Resilience. SAFER was designed to align to US Navy suicide prevention efforts, which include broad-based general military training (GMT) to build skills and understand available resources. However, suicide prevention skills and goals vary widely between different people, due to the history of and current level of experienced stress, concerns about friends or family, or the persons need to act as a leader to build social supports. As a result, personalized learning is important for each individual to build the skills and mindset that is relevant to how they can recognize and reduce suicide risk.

Compared to earlier PAL3 training domains such as electronics or leadership skills, SAFER suicide prevention presents unique challenges for personalized learning that required significant changes to the PAL3 framework. The four challenges were: Relevance and risk estimation, Content for prevention, Plans on how to apply skills to a real situation, and Privacy of sensitive data. These represent additions to the PAL3 framework and also required modifying or disabling earlier capabilities not appropriate for SAFER's use cases.

Relevance and Risk - Adaptive Intake Survey and SOS Button

Fairly quickly when designing SAFER, we recognized a key concern: what if a Sailor comes to the app because they are currently at high risk? This is a non-trivial issue and ties in tightly with privacy issues, since Sailors would be less likely to be frank and open with a system that will report back on them. The decision was made to search for potential risk factors and, if identified, suggest ways to reach out for help. This was accomplished by an initial intake interview with the pedagogical agent, which asks about reasons for visiting the system and about different types of risk factors.

The first question in the suicide prevention interview determines if they are ready to complete the survey, if they have concerns about completing it, or if they came because they need immediate help (Figure 2a). If they indicate that they need help, we open the Safety Button, also called the SOS button (Figure 2b). The Safety Button opens a content tree which can be navigated by clicking through the tree options or by searching for content. The Safety Button area can also be opened to directly display a specific piece of content, as is shown below. The resource gives clickable phone numbers for suicide hotlines, crisis chat links, and suggestions about how to increase safety against self-harm. Content in the Safety button is unique in that it is nearly always available via the upper right-hand button, even when offline (all associated content is downloaded). While the content in the Safety Button is currently limited to seeking help and helping others, this could be expanded to be context-sensitive to the current training goal and

could be used to offer a fast way to search for just-in-time skills (e.g., reminders on how to perform CPR). If they indicate they are not comfortable completing the survey, we ask for their reason and in that process, they can also return to complete the survey. For each option they select in the survey, the system adds or subtracts counters from a set of attributes. These attributes are:

- Self-At-Risk: Risk factors for harm to self (overall)
- Others At-Risk: Risk for others (e.g., concerned for a friend)
- Prevention: Interest in prevention in general (e.g., a leader)
- Disengaged: Response pattern shows lack of attention
- Negative Feelings: Feeling depressed, hopeless, anger, etc.
- Stress: Indicates high levels of stress and stress-related issues
- Sleep Issues: Poor sleep quality and fatigue
- Exercise: Lack of physical activity
- Social Support: Feeling a lack of social network or help
- Unsecured Guns: They have unsafely stored firearms
- Suicidal Ideation: Indicates thoughts or consideration of suicide (can trigger SOS Button)

The survey is adaptive, where questions are displayed or hidden based upon the current levels of attributes. For example, if the user shows high Self At-Risk early in the survey, we open additional questions to ask about Suicidal Ideation and suggest ways to seek help. However, if they show low risk and we have not directly asked if they had Suicidal Ideation, we ask that near the end of the survey, just to be sure we do not miss asking this critical question. This enables the survey to emulate an interview by asking additional questions about areas of concern, while keeping the questions for each learner brief (e.g., about 5 minutes).

The questions in this survey are primarily based on established clinical surveys for assessing risk factors such as mood disorders, sleep problems, and suicide risk. These short and well-validated screening measures include: 1. The Generalized Anxiety Disorder scale (Spitzer et al., 2006); 2. The State-Trait Anxiety Inventory (Spielberger, 1989); 3. The Patient Health Questionnaire-9 (Depression) (Kroenke & Spitzer, 2002); 4. The Beck Hopelessness Scale (Beck et al., 1988); 5. The Insomnia Severity Index (Morin et al., 2011); and 6. The PTSD Checklist-M (Blevins et al., 2015; Weathers et al., 2013).

New questions were also added to directly align to the learning topics available, such as questions about prior experience and confidence in applying skills from certain topics. Additionally, questions for social support and reasons for using the system were created ad-hoc, due to these being tailored toward Navy Sailors.

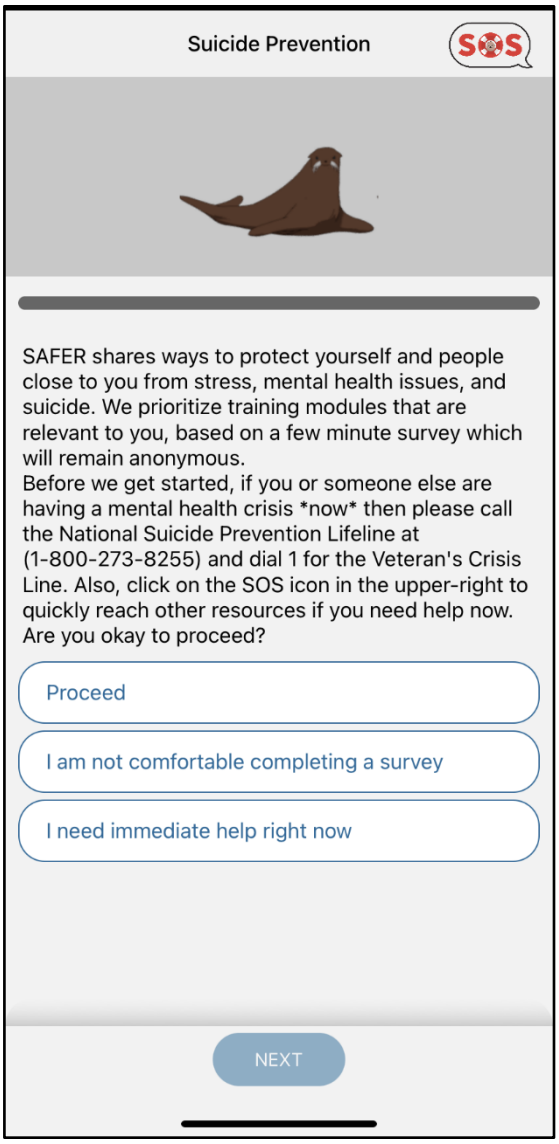


Figure 2a: Intake Survey (First Question)

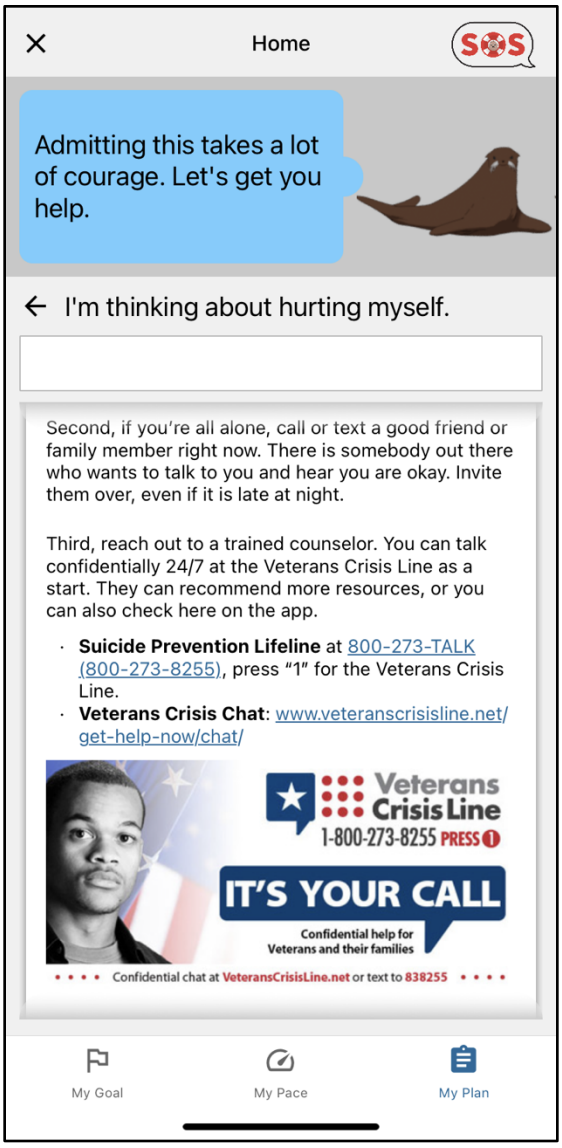


Figure 2b: Safety Button/SOS Suggesting help

When the survey is complete, their responses to questions generate a personalized roadmap for learning topics based on their interests and risks that are relevant to them (Figure 3). This roadmap considers three factors: relevance based on attributes, if the topic was mastered already (if any prior resources), and prerequisites for topics (which topics should be mastered before others). As a result, the roadmap updates to reflect prior learning and show the current priorities. For a new learner, the attributes determine the initial roadmap. Each topic can have weights associated with attributes, which may be positive (more relevant) or negative (less relevant). These enable calculating a weighted sum for the relevance of a topic to the learner, based on their attribute profile from the survey. This is expected to increase engagement and usefulness of the content, by providing the most relevant topics first.

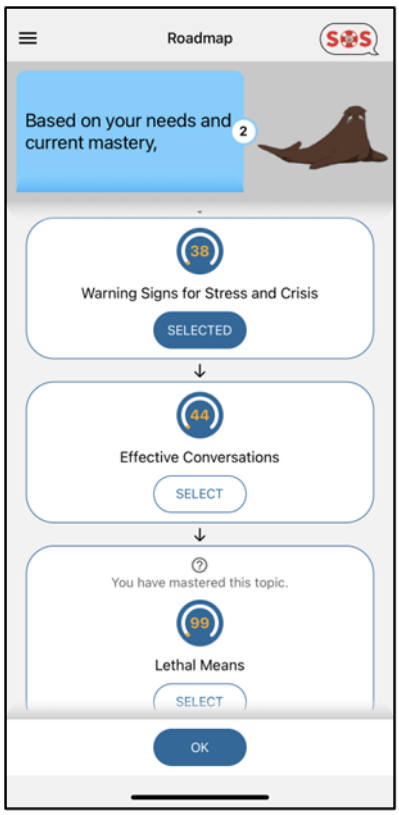


Figure 3. Personal Roadmap

Content for Prevention - Planning Ahead for Safety

The core content for SAFER is aligned to the Navy's General Military Training (GMT) content for suicide prevention, which is managed by Navy N17, the 21st Century Sailor program. However, further structure was required to organize the training into topics with a prevention focus. The unifying theme was "Planning ahead for safety" where Sailors build skills, mindsets, and behaviors that increase resilience and reduce stress so that if a major stressor or loss occurs, they have greater readiness. Inspiration for learning content was informed and adapted from a variety of well-vetted web-delivered sources including: National Institute for Mental Health (National Institute for Mental Health, 2022), Navy Leaders Guide for Managing Sailors in Distress (Navy Medicine, 2021), National Center for PTSD (U.S. Department of Veterans Affairs, 2021) and Beyond Blue (Beyond Blue, 2022).

As part of this structure, Sailors are introduced to the concept of the suicide Safety Plan, a tool designed for clinicians to help a person at-risk to understand and plan ahead for the people, resources, and strategies that they can use when they notice their own warning signs (Stanley et al., 2008). In SAFER, as a clinician is not available, the Safety Plan is treated as a practice opportunity: familiarity with the tool should assist them in working with a clinician if needed and even for Sailors who may never need a safety plan, it offers a concise outline of key prevention skills. The Veterans Administration Safety Plan template (U.S. Department of Veterans Affairs, n.d.) has seven sections which can be summarized as: Personal warning signs, Internal coping strategies, People and places for healthy distractions, People to contact for help, Professional help for a crisis, Making a safer environment/home, and Reasons for Living.

The topic areas in SAFER mirror these sections, consisting of:

- Introduction to Safety Planning: Summarizing the reasons for planning ahead, the role of pro-activity reducing risk factors and stress, and the concept of a safety plan.
- Warning Signs for Stress and Crisis: Identifying physical, behavior, and emotional warning signs in yourself and in others.
- Quick Coping Strategies: Physical tools (e.g., breathing) and cognitive tools (e.g., disrupting cognitive distortions and dysfunction patterns of thought) to reduce risk.
- Stronger Support Networks: Recognizing different types of social support, understanding your support network, and building stronger support networks.
- Reaching Out (to professional help): Overcoming barriers to help-seeking and understanding the professional resources available in the Navy.
- Effective Conversations (talking to others at-risk): Practicing how to reach out to others and talk with them supportively and productively.
- Lethal Means: Understanding why securing methods of suicide can reduce risk long-term and understanding the best options to secure the guns in their household.

Each topic contains multiple types of resources, typically starting with a non-stigmatizing and motivating rationale (e.g., a video or infographic) followed by a review which may include a multiple-choice quiz or an OpenTutor (Nye et al., 2021) conversational tutoring lesson. After skills are introduced, practice activities are presented until the learner shows mastery of the topic. In the current topics, to keep learners moving through the material smoothly, the current set of lessons are calibrated to help learners reach mastery without frustration (e.g., simpler assessments, low repetition). By comparison, some prior PAL3 subject areas included more challenge problems or simulation-based practice. These more challenging practice opportunities may be appropriate for future content, which depends on mastery of foundational topics. For example, topics that were considered but which were not integrated were Emergency Response (recognizing an acute crisis and helping connect them to care), Command Climate (leadership strategies to improve social support and help-seeking), and Postvention (leadership steps to prepare for and respond to a death by suicide).

In addition to adding content, some systems of PAL3 were modified to support SAFER. As part of a synergistic research effort, the COPE Tutor was developed to support use-cases such as suicide prevention content. The COPE Tutor is a substantial expansion of the OpenTutor framework, an open-source project which delivers authorable and incrementally improved open-response tutoring dialogs. COPE was developed to address needs observed when shifting PAL3 tutoring dialogs from the electronics domain to areas such as leadership and peer pressure. During this shift, it was noted that direct feedback was often inappropriate for sensitive topics (e.g., a learner says, "It would be hard because I would be depressed." and then the tutor says "No. That's not right."). Systematic changes were made to the OpenTutor dialog system, such that dialogs could be specified as "Sensitive" vs. "Traditional". This also involved changes to symbols and color schemes, to avoid "red for wrong" but instead using more neutral tones for corrections. Compared to the Traditional dialog policy, Sensitive dialogs avoid strong negative feedback, tone down positive feedback (e.g., avoid "Great!"), provide more encouraging prompts, and optionally provide a "survey says" board to focus attention more on the correct answers rather than on the feedback. While a separate evaluation is determining the impact of these changes, initial testing indicates that they enable meaningful dialogs on more sensitive topics that might be too callous using a traditional more direct tutoring approach.

Planning Ahead - Incrementally Building a Safety Plan

In addition to using the Safety Plan to help structure content, the training also helps the learner develop their own personal safety plan. While this plan would not be near the level of a plan developed with a well-trained clinician, this safety plan helps them think about how they would leverage protective factors and strategies in their own life. It also can be exported and shared as a PDF, in the case that they might need it in the future as a starting point for a professionally-aided safety plan or to share with a bystander during an unexpected crisis.

Building a safety plan required three additions to develop SAFER. First, a new main area in the PAL3 framework was developed ("My Plans", as shown in Figure 4a). This area allows any PAL3 goal to be associated with one or more "plans" which can be accessed as fillable forms. Plans may be as simple as a single field (e.g., a "Notes" form) or can involve multiple sections which accept text, phone numbers, locations, and other fields. The Suicide Prevention goal has one plan: the Safety Plan.

Within each topic, one or more special "Planning" lessons can be added. These resemble conversational tutoring lessons, but rather than assessing responses and giving tutoring feedback, they ask for information to help complete a section of a Plan (Figure 4b). Each answer can be associated with custom validation, to help users to improve their response if it is likely to be unsuitable (e.g., too short, invalid phone number). Each Planning lesson in SAFER is based on question prompts adapted from the VA Safety Planning guide for clinicians and associated short-form guides with additional question prompts (Stanley et al., 2008). It must be emphasized that this is not close to emulating a professional. While an actual therapist would know the client's history, help them think about scenarios where they felt warning signs, and ask about barriers to certain strategies, the SAFER planning dialogs are comparatively shallow and meant to encourage reflection and an initial draft of planning. After the Planning dialog is complete, the learner is asked if they want to update their Safety Plan based on their responses. Each dialog completed fills out a section of the plan, as shown in Figure 4c.

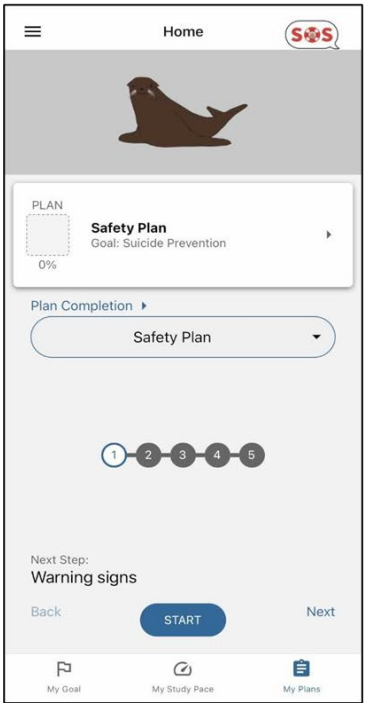


Figure 4a: Safety Planning.

Figure 4b: Gathering Info

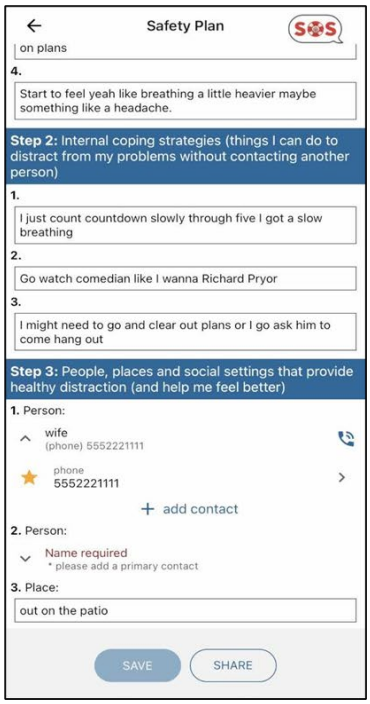


Figure 4c: Safety Plan

Privacy of Sensitive Data - Local Storage and Controls

During the design of these features and content, privacy and sensitivity of data was a key issue. Privacy considerations are particularly sensitive in a military setting, because health issues (including sometimes mental health issues) can impact a Sailor's readiness to serve in certain jobs, roles, or even remain in military service. While this is uncommon in practice, since the military invests heavily in each member and tries to return them to service, concerns about career repercussions can delay Sailors from seeking mental health services or make them careful to choose services with clear policies about disclosure (Ho et al., 2018). As a result, it was decided that a high degree of privacy and personal control over information would be the default for the system.

A three-tiered privacy model was chosen for the current version of SAFER:

1. User ID-Linked: When starting the app, learners can make an arbitrary user ID based on any available email, to enable discretion over how easily that can be tied to a specific person. This ID is associated with data such as which resources are completed and for certain persistent settings (e.g., study pace).
2. Local-Only: Other data is only stored locally on-device. If the user creates an account on a new device, they must manually re-enter it (though in the future, a method to manually transfer it is planned). This data includes the attributes calculated after completing an intake survey and the contents of the Safety Plan.
3. Non-Persistent: Finally, some data is not stored after entering. This includes the responses given to the intake survey and any responses in a planning dialog where the user does not update their local safety plan.

Thus, maximally sensitive data is not stored long-term (such as specific survey responses). Sensitive but less-specific data is stored only on-device, with no server storage or synchronization. The remaining data is relatively non-specific, such as which resources the learner has completed. This is not particularly identifying, in that all resources will eventually be recommended, and learners have the option to start with any topic they choose, regardless of the recommender system order. Additionally, to retain greater privacy, the PAL3 Teams area was disabled for SAFER so that learners train individually. While there could be cases where cohorts were appropriate, it was decided that testing with users to better understand their privacy considerations and preferences would be required before enabling or adapting this feature.

By prioritizing privacy, we expand access: more users should be able to trust that they can use SAFER, and they will be able to use the app more authentically. In addition, SAFER offers quick access to crisis hotlines and other tools to help both bystanders and individuals who are at-risk. However, there are downsides to this level of privacy. Even if a user answers with high-risk responses, we are not able to automatically notify a human to contact them. Moreover, even if we wanted to, the data that a user provides would be insufficient to know their phone number or location to reach them. As a result, we can only trigger the SAFER coach and Safety Button to suggest seeking help from friends or hotlines. We expect that this should be appropriate for the level of risk that users have when actively in the system (e.g., that an acutely suicidal person would not be likely to spend their time in a learning/training app). However, it is still a limitation of the way the app is currently designed.

Future versions of the app will hopefully be able to further tune the privacy and data management. One added level of security would be to further de-identify User ID-Linked data in the app, so that only one system component is possible to make that connection. This would facilitate an organization (e.g., the Navy) to analyze patterns of system usage without any meaningful user ID's (i.e., only arbitrary ID mappings), for an additional layer of security and privacy. A second addition would be to combine the current Safety Button triggers, which occur when high risk is detected, with a notification system that allows a user to accept or reject directly calling a crisis hotline or similar human connection. By further reducing the friction to reach a trained human helper, this should increase the likelihood that a person with acute risk can get help.

Discussion

Most intelligent tutoring systems have been developed to run on laptops or desktops. The instructional content is typically drawn from some academic discipline such as algebra, electronics or physics. The tutoring systems are often used as an augmentation of a classroom experience. Students may be challenged, bored, or engaged by the content, but the content itself is not a risk for students. As we developed SAFER, we found that we were moving into a very different space that required us to make changes from the usual approach to intelligent tutoring systems.

First, we recognized that people using the system could have very different motivations for seeking suicide prevention training. They might be concerned about helping a friend or loved one with suicidal thoughts, or they might be having suicidal thoughts themselves, and there are various risk factors that they or others might have. This wide variability meant that a single path through the content was not likely to be effective and led us to create the initial survey to elicit users' motivations and risks. Second, SAFER is designed to run on mobile devices, which means it is always available. While increased availability is good, it also means that people may use it when they are away from instructors or classmates. That could be problematic if the person were feeling suicidal or became suicidal while using SAFER. That is why the initial interview that we added to the PAL3 framework in creating SAFER not only assesses a user's motivations for using the system, but it also assesses their risk and if the risk is high enough suggests that they need to seek counseling with a real human and makes it easy to contact help.

Thus, the initial interview is not only finding out more about the learner so that the learning experience can be customized, but also suggesting that they stop using the app and seek counseling if they seem at high risk. This is not something that typically happens in intelligent tutoring systems but is necessary for this domain.

Third, we were concerned that if users felt their data might be shared with others, particularly supervisors, people would be reluctant to use the system. To allay those concerns we designed the systems so that all personal data stays on the user's device. Elements such as the safety plan are only shared with others if the user decides to share them, and the default is not to share.

Fourth, the fact that some users might be confronting high stress and other risk factors meant that the system responses to user inputs had to be couched to reflect potential user sensitivities. A simple "right" or "wrong" response that might be acceptable in a conventional training system needed to be modified to be more supportive and nuanced.

Recommendations for GIFT and Intelligent Tutoring Systems

Several key aspects of the PAL3 framework are relevant to tutoring systems such as the Generalized Intelligent Framework for Tutoring (GIFT), which can deliver both web-based training for desktops and connect with team simulations. First, SAFER targets a use-case where adaptive mobile training is used as personalized training that is intended to complement in-person team training (e.g., an on-site GMT training session). While SAFER uses this pedagogy for suicide prevention training, training for squad level simulations could use the same general design: personalized competency-building → team training exercise → goal-setting for team and individuals → additional personalized training. Depending on the training goals, this pattern could be used for multiple different team training designs such as: a) Collaborative Learning: having each member of a team learn different things to share in-person, b) Common Ground: establishing a baseline of prerequisites prior to team training, c) Role-Based Training: practicing skills that are relevant to only one role within a larger group, and d) Goal-Setting: following team training, set personalized practice goals that will improve each member's contribution to the next team training.

Second, persistent mobile adaptive learning such as SAFER can provide continuity across many different team training environments. This is important, because a common concern among military learners is that they train extensively but often lack easy ways to re-visit and review material later when it is needed.

Finally, as intelligent tutoring systems expand into new domains, such as mental health, we believe that it will be important to support capabilities from PAL3 such as an initial survey that enables up-front, persistent personalization and also data-sensitivity settings that determine which data is shared versus kept private/local for the user.

Conclusions

In this chapter, we have discussed the changes that we found necessary to make to the PAL3 framework as we moved from domains that intelligent tutoring systems typically cover such as electronics or leadership to suicide prevention training. The resulting SAFER system currently exists as an advanced prototype ready for evaluation.

It is important to note that the SAFER approach is not designed to replace the care of a live provider. Rather, it can fill a gap where a live provider is not available or where the user is hesitant to speak with

one. This gets at the core of what is needed most to engage troubled service members who are resistant to or disengaged from mental healthcare services (but who are in the most need and perhaps at the highest risk of suicide). The ability to leverage a mobile application that can help enhance service member resiliency via systematic access to critical self-awareness building and psychoeducation content, offers such a safety net. Although SAFER is not a substitute for live professional care when needed, it can provide a complement to that care--and, unlike real human clinicians, mobile apps are always available, never tire, have ready access to an extensive library of relevant learning resources, and maintain a steady and consistent presence. Moreover, with additional software enhancements, the PAL3 system could assemble a knowledge base of a users' issues through repeated interactions with users that could be used to guide further development of suicide prevention content. Thus, whether to fill a gap in absent care, or serve as its complement, the potential for PAL3 to reach all service members in support of their mental health needs offers a pragmatic and pro-social example of the potential benefits of intelligent tutoring systems.

Acknowledgements

The project of effort depicted here was or is sponsored by the U.S. Army under contract W911NF-14-D-0005 and through ONR (Office of Naval Research) contract N00014-16-C-3027 with support from MOM RP, Navy N1 and N17 and the Office of Naval Research. The content of this paper does not necessarily reflect the position or policy of the US Government and no official endorsement should be inferred.

References

Aleven, V., & Koedinger, K. R. (2013). Knowledge component (KC) approaches to learner modeling. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design Recommendations for Intelligent Tutoring Systems - Volume 1: Learner Modeling* (pp. 165–182).

Armed Forces Health Surveillance Center (AFHSC). (2012). Deaths while on active duty in the U.S. Armed Forces, 1990-2011. *MSMR*, *19*(5), 2–5.

Beck, A. T., Steer, R. A., & Pompili, M. (1988). *BHS, Beck hopelessness scale: manual*. Psychological corporation San Antonio, TX.

Beyond Blue. (2022). *Welcome to Beyond Blue*. Beyond Blue. https://www.beyondblue.org.au/home

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, *28*(6), 489–498.

Centers for Disease Control and Prevention. (2021). *WISQARS Leading Causes of Death Reports, 1981–2019*. https://wisqars.cdc.gov/data/lcd/home

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243.

Drapeau, C. W., & McIntosh, J. L. (2020). *U.S.A. Suicide: 2019 Official Final Data*. American Association of Suicidology. https://suicidology.org/wp-content/uploads/2021/01/2019datapgsv2b.pdf

Eaton, K. M., Messer, S. C., Garvey Wilson, A. L., & Hoge, C. W. (2006). Strengthening the validity of population-based suicide rate comparisons: an illustration using U.S. military and civilian data. *Suicide & Life-Threatening Behavior*, *36*(2), 182–191.

Hampton, A. J., Nye, B. D., Pavlik, P. I., Swartout, W. R., Graesser, A. C., & Gunderson, J. (2018). *Mitigating Knowledge Decay from Instruction with Voluntary Use of an Adaptive Learning System*. Springer International Publishing.

Ho, T. E., Hesse, C. M., Osborn, M. M., Schneider, K. G., Smischney, T. M., Carlisle, B. L., Beneda, J. G., Schwerin, M. J., & Shechter, O. G. (2018). *Mental Health and Help-Seeking in the US Military: Survey and Focus Group Findings*. Defense Personnel and Security Research Center. https://apps.dtic.mil/sti/citations/AD1059321

Kroenke Kurt, & Spitzer Robert L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515.

Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. *User Modeling and User-Adapted Interaction*, *27*(1), 55–88.
Morin, C. M., Belleville, G., Bélanger, L., & Ivers, H. (2011). The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. *Sleep*, *34*(5), 601–608.
National Center for Injury Prevention and Control. (2018). *Suicide rising across the US*. Centers for Disease Control and Prevention. https://www.cdc.gov/vitalsigns/suicide/index.html
National Institute for Mental Health. (2022). *Transforming the understanding and treatment of mental illnesses: Anxiety Disorders*. https://www.nimh.nih.gov/health/topics/anxiety-disorders/
Navy Medicine. (2021). *Navy Leaders Guide for Managing Sailors in Distress*. https://www.med.navy.mil/Navy-Marine-Corps-Public-Health-Center/Population-Health/Health-Promotion-and-Wellness/navy-leaders-guide-for-managing-sailors-in-distress/
Nye, B. D., Sanghrajka, R., Bodhwani, V., Acob, M., Budziwojski, D., Carr, K., Kirshner, L., & Swartout, W. R. (2021). OpenTutor: Designing a Rapid-Authored Tutor that Learns as you Grade. *The International FLAIRS Conference Proceedings*, *34*(1). https://doi.org/10.32473/flairs.v34i1.128576
Pirker, J., Rattinger, A., Drachen, A., & Sifa, R. (2018). Analyzing player networks in Destiny. *Entertainment Computing*, *25*, 71–83.
Rothberg, J. M., & Jones, F. D. (1987). Suicide in the U.S. Army: epidemiological and periodic aspects. *Suicide & Life-Threatening Behavior*, *17*(2), 119–132.
Shonfeld, & Resta. (n.d.). Competitive game effect on collaborative learning in a virtual world. *Collaborative Learning in a Global World*. https://books.google.com/books?hl=en&lr=&id=ySKADwAAQBAJ&oi=fnd&pg=PA91&dq=shonfeld+resta&ots=GMLw0Xp_s8&sig=117_a_1D790yIheEkZBGKfSG6lk
Spielberger, C. D. (1989). *State-trait Anxiety Inventory: A Comprehensive Bibliography*. Consulting Psychologists Press.
Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.
Stanley, B., & Brown, G. K. (2012). Safety Planning Intervention: A Brief Intervention to Mitigate Suicide Risk. *Cognitive and Behavioral Practice*, *19*(2), 256–264.
Stanley, B., Brown, G. K., Karlin, B., Kemp, J. E., & VonBergen, H. A. (2008). *Safety plan treatment manual to reduce suicide risk: Veteran version*. https://sefbhn.org/assets/zero-suicide-recommended-evaluation-tools/safety-plans/stanley-brown-safety-manual.pdf
Swartout, W., Nye, B. D., Hartholt, A., Reilly, A., Graesser, A. C., VanLehn, K., Wetzel, J., Liewer, M., Morbini, F., Morgan, B., Wang, L., Benn, G., & Rosenberg, M. (2016). *Designing a Personal Assistant for Life-Long Learning (PAL3)*. FLAIRS-29, Key Largo, FL.
Tang, S., Reily, N. M., Arena, A. F., Batterham, P. J., Calear, A. L., Carter, G. L., Mackinnon, A. J., & Christensen, H. (2021). People Who Die by Suicide Without Receiving Mental Health Services: A Systematic Review. *Frontiers in Public Health*, *9*, 736948.
Tucker, Smolenski, & Kennedy. (n.d.). Department of Defense Suicide Event Report (DoDSER): Calendar year 2018 annual report. *Psychological Health Center of Excellence*.
U.S. Department of the Army. (2015). *Army Health Promotion (AR 600-63)*.
U.S. Department of Veterans Affairs. (n.d.). *Safety Plan Worksheet*. https://www.healthquality.va.gov/guidelines/MH/srb/PHCoEPatientSafetyPlanSelfPrint3302020508.pdf
U.S. Department of Veterans Affairs. (2020). 2020 National veteran suicide prevention annual report. *Office of Mental Health and Suicide Prevention*.
U.S. Department of Veterans Affairs. (2021). *PTSD: National Center for PTSD*. https://www.ptsd.va.gov
Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The ptsd checklist for dsm-5 (pcl-5). *Scale Available from the National Center for PTSD at Www. Ptsd. va. Gov*, *10*(4), 206.

CHAPTER 11 - AGENT-BASED INTELLIGENT TUTORING SYSTEMS FOR PROFESSIONAL DEVELOPMENT

Arthur C. Graesser and Xiangen Hu
Institute for Intelligent Systems, University of Memphis

Introduction

Conversational agents have been integrated with adaptive learning environments for at least three decades (for reviews, see Graesser & Li, 2022; Wang et al., 2022). The conversational agents vary in the extent to which they have realistic human voices, facial persona, emotional expressions, and body movements. Agents in minimalist chat systems communicate in printed messages with communicators depicted by a static facial icon. Agents in most systems developed in K12, college, and other instructional contexts are animated talking heads. Agents in augmented reality, virtual reality, and the metaverse are avatars in 3d worlds. Conversational agents can take on different roles, such as tutor, coach, mentor, peer, companion, adversary, and so on.

Conversational agents can perform different tasks and pedagogical functions. They can guide the learner on how to navigate a complex human-computer interface or to make progress in scenario-based tasks by nudging them (via hints, questions, or requests) to generate physical and verbal contributions. Agents can define or explain something when asked, pop in and explain something when the learner is stuck, give feedback on learner actions with explanations/justifications, and hold conversations in natural language (Nye et al., 2014; Wang et al., 2022). Pairs or groups of conversational agents can model social interaction, prompt a debate or disagreement to stimulate deeper learning, and stage scenarios for team training (Graesser et al., 2017). The agents can be designed to handle emotions in addition to cognitive skills, knowledge, and abilities (Arroyo et al., 2014; D'Mello & Graesser, 2012, 2023; Taub et al., 2020). Indeed, these conversational agents have been designed to perform essentially any task or function that humans perform (and more).

At this point in history, conversational agents have rarely been integrated in learning environments for adult professional development, at least compared with the dozens, if not hundreds, of systems developed for K12 and college. This raises two fundamental questions. First, what are adults' impressions of these conversational agents? We know from our previous projects with AutoTutor (Graesser, 2016; Nye et al., 2014) that some adults have negative attitudes toward the agents for a variety of reasons that are discussed later. If the majority of adults have a negative impression of the agents, then that does not bode well for adoption. Hence, data need to be collected on adult impressions in different populations, subject matters, and tasks. Second, is there added value of conversational agents in improving learning? If there is no added value, then there is no reason to use them. An answer to the second question is no doubt more nuanced. In essence, we need to know the populations, subject matters, and tasks in which conversational agents have added value in learning gains. It is important to acknowledge that answers to the first question are very different from the answers to the second question because there tends to be a zero or negative correlation between liking and learning of difficult material (Graesser & D'Mello, 2012).

Goals and Scope

This chapter explores the two fundamental questions (i.e., addressing liking and learning) in two adult populations and contexts. The first is an *ElectronixTutor* system to help Sailors learn about the

fundamentals of electronic circuits so they can pass tests to progress in their ranks as electronics technicians. The second is an AutoTutor system to help struggling adult readers at literacy centers improve their reading comprehension skills so they can land a decent job. These two cases represent two ends of a continuum of learning, skills, and abilities of adults. The Sailors have extremely high knowledge, skills, and abilities (as manifested on the Armed Services Vocational Aptitude Battery) and are thereby selected for training in the Navy Nuclear Power Training Center to be electronics technicians. In contrast, adults who use *AutoTutor-ARC* (Adult Reading Comprehension) have significant challenges in knowledge, skills, and ability for many reasons (e.g., immigration, poverty, cognitive limitations) so they need help from reading literacy centers. Both ElectronixTutor and AutoTutor-ARC have a design with two agents in a *trialogue* (Graesser et al., 2017) with a tutor agent and a peer student agent. The two systems are comparable in this sense whereas they differ in subject matter and many other features.

Liking and learning data have been collected for ElectronixTutor and the AutoTutor-ARC. The architecture of ElectronixTutor has been articulated in previous publications (Graesser et al., 2018; Hampton & Graesser, 2019) whereas the data are reported in a recent technical report (Nye et al., 2022). The architecture and data on AutoTutor-ARC has been reported in recent publications (Chen et al., 2022; Fang et al., 2022; Graesser et al., 2019). This chapter summarizes the highlights of the results with respect to liking and learning.

ElectronixTutor for Sailors in the Navy

Sailors at the Navy Nuclear Power Training Center receive training on electronics fundamentals in order to complete the Nuclear Field A School (NFAS) and receive an Electronics Technician Nuclear (ENT) rating. These Sailors have high ability according to their scores in the Armed Services Vocational Aptitude Battery (ASVAB) and the Basic Electricity and Electronics (BEE) test. There are different job roles that are relevant to their career paths, such as Equipment Operator (EO) and Reactor Operator (RO).

ElectronixTutor was used in a pilot study in the first grading period of the NFAS Electronics Fundamentals curriculum (Nye et al., 2022). The subject matter in this period covered semiconductor fundamentals, biased positive-negative junction, solid-state rectifiers, direct current power supply filters, and solid-state power supplies. The Sailors assigned to the pilot study (N=390) were compared with those who did not receive training with ElectronixTutor (N=721). The course had human instructors in a classroom over several days (28 hours). The extra training with ElectronixTutor was typically 1.5-3 hours that were spread over the days in a lab with 20 computers. The Sailors could freely choose the days and times to use ElectronixTutor, which typically was found to be in the afternoon and early evening.

ElectronixTutor was originally designed to have AutoTutor conversational agents interact with the Sailor in conjunction with other intelligent tutoring systems (ITSs) and digital facilities to improve learning (Graesser et al., 2018). The original prototype included AutoTutor conversational agents, three other ITSs developed at other universities, a point and query system to have immediate answers to Sailor questions about a circuit, Navy technical documents, and topic summaries. There was a recommender system that could recommend different learning resources and topics to pursue, given the performance of the learner that was stored in the Learning Record Store. However, the version of ElectronixTutor in this pilot study had a more focused set of learning resources. It included AutoTutor trialogues, YouTube videos of relevant electronics topics, and digital reading materials designed by instructors.

Figure 1 shows a screenshot of the AutoTutor component of ElectronixTutor, which was the primary component that guided the interaction with the Sailor. There is a tutor agent in the upper left, a peer student agent in the upper right, a main question for the Sailor to verbally answer (“Considering atomic interactions, how is N-type material made?”), and a diagram. The bottom is a space for the Sailor to type

in an answer. The bottom left is a pallet of alternative resources the Sailor can click on, such as access to a YouTube video, digital reading materials, and the conversational history.

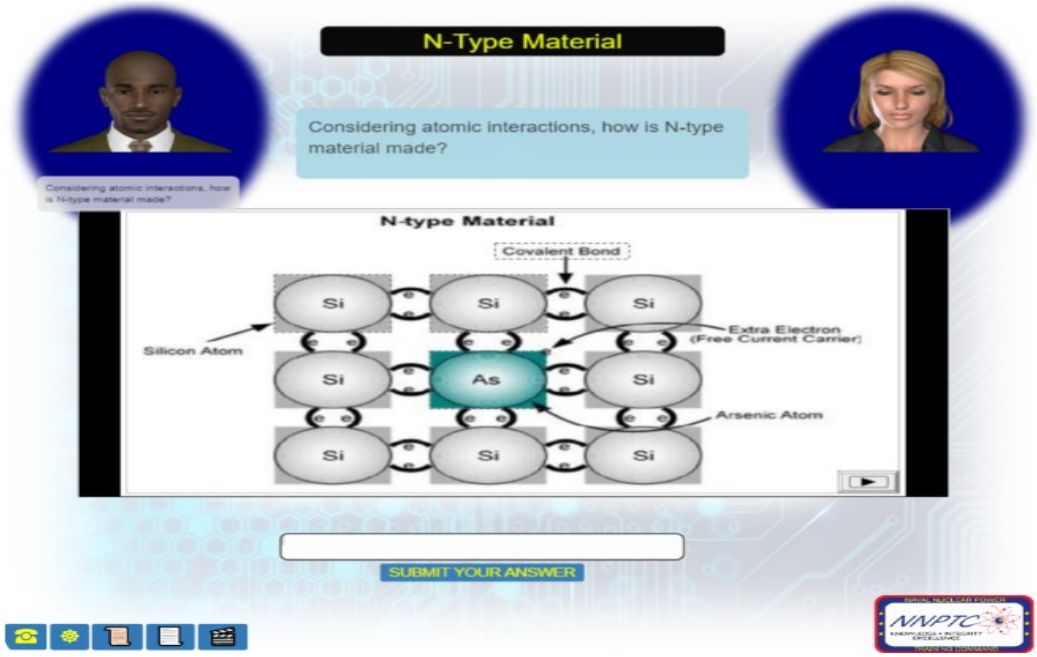


Figure 1. Screenshot of an AutoTutor component of ElectronixTutor.

When the Sailors type in their verbal responses to the main question, the system evaluates the answer using advances in natural language processing in computational linguistics. It is beyond the scope of this chapter to specify how this is accomplished, but some relevant highlights should convey the essence. AutoTutor uses an "Expectation & Misconception Tailored" (EMT) dialogue approach that compares the Sailor's verbal response to a set of expected good answers (i.e., sentence-like expressions) and anticipated misconceptions (Graesser, 2016; Nye et al., 2014). Automated computational linguistics algorithms compare the semantic overlap (between 0 and 1) between the Sailor answer and each of the expectations and anticipated misconceptions to the main question. When the Sailor gives an initial answer that has high semantic overlap scores and covers all expectations, they get positive feedback and then a computer summary that is well articulated and complete. When the Sailor misses particular expectations, AutoTutor gives hints and directed prompts to encourage the Sailor to fill in missing ideas and words; this can continue with multiple turns and exchanges until all of the expectations are covered. When the Sailor expresses information that matches a misconception, AutoTutor gives negative feedback and corrections. Consequently, it often takes many turns to fill in all of the correct information and remediate misconceptions.

Nye et al. (2022) have reported the results of the pilot study with respect to their liking and learning of ElectronixTutor. With respect to liking, there were many questions on a Likert scale that inquired about their subjective response to ElectronixTutor, the YouTube videos, and other learning resources. Overall, 78% agreed or strongly agreed that they had a positive response to AutoTutor. That is of course good news for advocates of conversational agents. That being said, there were differences between the different subsamples of Sailors. Sailors with EO ratings had a 91% positive response rate whereas the mean rating was only 63% for Sailors with RO ratings. Clearly, there are differences between populations of adults. The RO Sailors tended to have higher scores and be higher in ability than the EO Sailors, which supports

the hypothesis that agents may not be a good fit for individuals with higher knowledge, skills, and abilities.

Regarding learning, there was a course exam on the content covered in the first period of the course. This test has Sailors give verbal responses to electronics problems, which are scored by instructors with respect to being accurate and complete. There were significant learning gains for the pilot Sailors compared with the Sailors in the control group. These gains were consistent after controlling for both a prior psychometrically validated test of BEE (Basic Electricity and Electronics) average test scores, Sailor rating (RO vs. EO), and adjusting for test difficulty. Interestingly, the EO Sailors were the primary beneficiaries; the RO Sailors did not show a significant benefit from the pilot intervention. Failure rate on the major course test for EO Sailors was 8.98% for the comparison group and 4.76% for the ElectronixTutor group. This is positive news for the added value of conversational agents in a population of higher ability adults who are learning difficult content like electronics. Minimizing failure in a course saves considerable costs in both money and nonoptimal decisions on career trajectories of individual Sailors. However, this trend does not generalize to all subgroups of these adults. There is a need to identify which categories of adults will benefit from conversational agents.

AutoTutor-ARC for Struggling Adult Readers

Approximately one in five adults aged 16 or older have literacy skills at a low level of proficiency (OECD, 2016), which is a major barrier in their advancing in their careers. Adults with low literacy skills are heterogeneous in demographic and psychological characteristics, such as age, race/ethnicity, country of origin, educational level, literacy skills, interests, and goals. Therefore, an ITS that personalizes their training would be highly desired. Many of the struggling adult readers have problems at the level of text comprehension as opposed to word decoding skills and other basic language components. AutoTutor-ARC was developed to help adults acquire the comprehension skills (Graesser et al., 2019). The long-term goal is to have this system available in reading literacy centers and vocational development centers.

AutoTutor-ARC has lessons that cover many comprehension skills, which are summarized in Figure 2. As in ElectronixTutor, there is a tutor agent in the top left, a peer agent in the top right. There is a list of lessons, each of which takes 20-60 minutes to complete as a struggling adult reader. The lessons cover many levels of comprehension, types of texts, and media. Struggling adult readers are helped at literacy centers with professional and volunteer instructors who discuss their life and career goals, their challenges, and approaches to improving their reading skills. These literacy professionals also tutor individuals and teach small groups on tasks and texts to improve their reading. The hope is that AutoTutor-ARC will be a useful resource for instructors and adults. A Center for the Study of Adult Literacy (CSAL, https://sites.gsu.edu/csal/) was funded by the Institute of Education Sciences of the US Department of Education to advance these goals.

Struggling adult readers have significant writing problems so it would not make sense to expect them to have a natural language conversation with a conversational agent via typing as opposed to speech. Consequently, there was very little typed natural language input by the adult, other than short words and phrases. Most of their actions were responding to questions and requests by the conversational agents by clicking, dropping, and dragging. In essence, multiple choice questions by a tutor or peer agent were woven into the conversational flow, with feedback on their decisions and hints for follow-up responses when their selections were incorrect. In order to improve motivation, many of the lessons had the peer agent ask for help (to boost the adult's self-esteem) and some lessons had a game competition between the adult and the peer agent, with the peer agent never winning (to boost the adult's self-concept). This is a low literacy population who has low self-efficacy, esteem, and self-regulation skills. All of the events and actions of adults were stored in a Learning Record Store in the cloud.

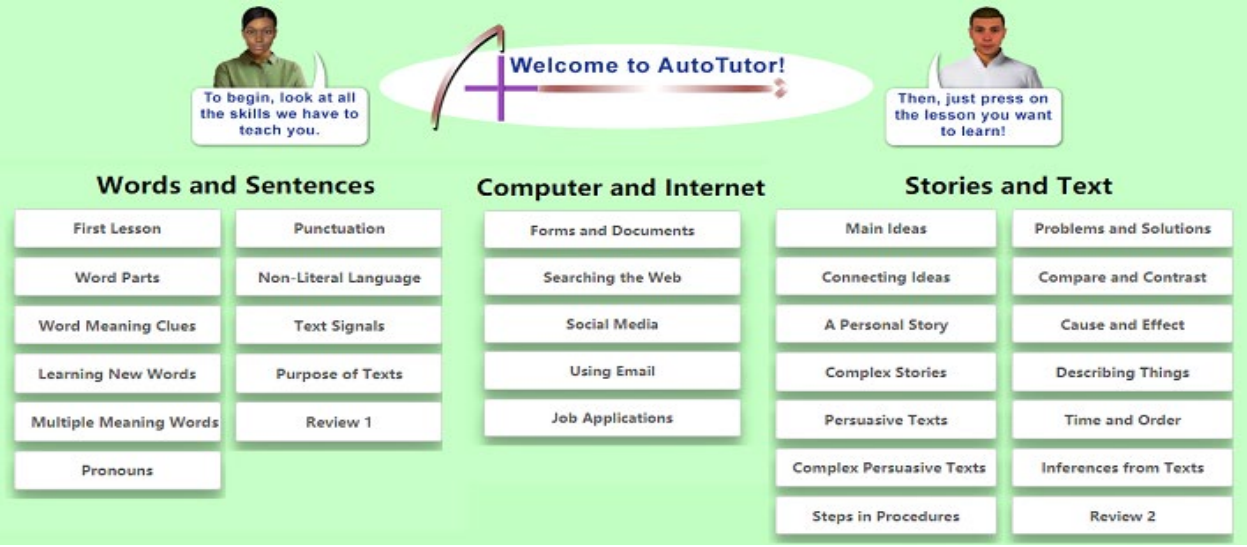


Figure 2: Scope of lessons in AutoTutor-ARC.

A pilot study was conducted to assess the effectiveness of a hybrid computer-instructor intervention with AutoTutor-ARC in literacy centers in Atlanta and Toronto (Fang et al., 2022). The adults started out with reading skills between the 3rd and nearly 8th grade level. A 4-month,100-hour intervention covered lessons reflected in Figure 2. Unlike the human instruction, AutoTutor-ARC was able to record all of the events and actions of the adults while using the computer system. This included both the time and accuracy of answering the conversation-based questions woven into AutoTutor-ARC. We also recorded the text reading times, text difficulty, and various events outside of the conversational flow (such as selecting a lesson summary video or an option to have a text read to them).

The pilot study (*N*=252) revealed that there were four clusters of adults who had distinctive profiles while interacting with AutoTutor-ARC. Clustering analyses revealed there were (1) relatively *proficient readers*, although below the 8th grade level starting out (answers to AutoTutor-ARC were accurate and comparatively fast), (2) *conscientious readers* (accurate but slow), (3) *underengaged readers* (relatively inaccurate but fast), and (4) *struggling readers* (inaccurate and slow). There were three psychometric tests to see how much the comprehension skills improved before and after the intervention, including a recent test developed at Educational Testing Service (Sabatini et al., 2019). The results revealed that clusters 1-3 all benefited from the intervention, primarily the conscientious readers; however the struggling readers in cluster 4 did not benefit at all from AutoTutor-ARC. Perhaps the struggling readers need a different intervention from AutoTutor-ARC at their zone of proximal development.

It is important to identify early during AutoTutor-ARC whether the adult is in cluster 4. We are currently trying to identify the point at which a reliable assessment can be made so the adult can shift to a different learning environment. We have developed algorithms (based on response time, accuracy, and question difficulty) that can assess the extent to which an adult is engaged in the learning activity, as opposed to quickly making decisions or mind-wandering (Chen et al., 2021). This algorithm can detect disengagement within a 1-2 minute time span. Perhaps this algorithm can signal early detection of struggling readers so their intervention can be sensibly changed.

Regarding liking, the struggling adult learners had a uniformly positive response to the AutoTutor agents. Their responses to self-report questions essentially had an overwhelming ceiling effect. This is compatible

with the conclusion that conversational agents are particularly suited to populations with lower knowledge, skills, and abilities.

Discussion and Recommendations for GIFT and Intelligent Tutoring Systems

The results of these two case studies on ElectronixTutor for Navy Sailors and on AutoTutor-ARC for struggling adult readers clearly indicate that the conversational agents benefit some populations of learners in particular contexts but not others. Most of the adults liked and learned from the agents, but others did not. For example, some of the high ability Navy Sailors did not have a positive impression of the agents and did not have learning benefits. Regarding the literacy study, there were struggling readers who liked the agents but did not improve their comprehension skills, presumably because AutoTutor-ARC was beyond their zone of proximal development. This disconnect between liking and learning needs to be seriously considered in the Generalized Intelligent Framework for Tutoring (GIFT) more generally. It is insufficient to judge the value of a learning environment based on impressions of the learners because deep learning takes effort and effortful learning is not particularly fun (Graesser & D'Mello, 2012). Some populations of learners enjoy a good challenge whereas others exit as soon as significant difficulties are apparent. But more specifically to this chapter, the data we have presented suggest we need more research that investigates the populations of learners and associated subject matter and tasks when conversational agents have added value. GIFT would benefit from tracking relevant information in the Learning Record Store and having production rules that adapt to the learner.

One simple generalization is that conversational agents are suited to adults with lower knowledge, skills, and abilities. This generalization is too simple, however, because the most struggling adult readers had no improvements in comprehension skills even though they liked the intervention. A more nuanced program of research is needed in order to decide when particular agent technologies (including chat, Virtual Reality, Augmented Reality, and the Metaverse) have added value and to make recommendations on when they should be launched.

The design of recommender systems in GIFT would benefit from a research base. We believe that fuzzy production rules are generally appropriate to capture the research generalizations.

IF <Population P, Learner Profile L, Subject Matter M, & Difficulty D> THEN <Digital Facility D>

The rules should be fuzzy so that the analytical systems can track what happens when the rule does not apply the ideal values in addition to when it does apply them. A fuzzy production system covers all values of a variable but has most of the observations at the expected value, with gradient decreases as values deviate from the expected value. Such variation is required for successful testing to see what happens when predictions are not followed. Indeed, we believe that all production rules should be fuzzy rather than brittle as a general policy in GIFT.

GIFT will need to find ways to handle incompatibilities between liking and learning in the arena of agents, as well as adaptive learning environments more generally. Available evidence suggests that adults will not learn much if they have a negative impression of a learning environment early on. Steps will need to be taken to contextualize and persuade the learner why it is important to have agents, or alternatively to remove the agents altogether and replace them with something different. For those learners who like the agents, GIFT systems need to optimize learning in a way that blends both deep learning and engagement. One way is to plant problems that put the learners in the state of cognitive disequilibrium, such as agents who disagree, which is hopefully resolved at some point (D'Mello & Graesser, 2012). Fortunately, there are algorithms that detect disengagement (Chen et al., 2021) so the learning record store can track both engagement and learning, both of which predict liking.

In summary, this chapter has articulated the value of conversational agents in facilitating learning as well as liking of the learning environments. Some populations who are attempting to learn particular subject matter indeed like the agents and learn from them whereas others do not, so there is a need for GIFT to sort out the conditions when agents have added value. The complex relationship between liking and learning underscores the need for GIFT to track engagement and learning throughout the learning sessions so that the system can optimize both dimensions. Moreover, conversational agents have a foundation for communicating and contextualizing what and how the adults will be learning. The conversations make learning visible.

References

Arroyo, I., Muldner, K., Burleson, W., & Woolf, B. P. (2014). Adaptive interventions to address students' negative activating and deactivating emotions during learning activities. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), Design recommendations for intelligent tutoring systems: Instructional management, Vol. 2 (pp. 79-91). Orlando (FL): Army Research Laboratory (US); 2014.

Chen, S, Fang, Y., Shi, G., Sabatini, J., Greenberg, D., Frijters, J., & Graesser, A.C. (2021). Automated disengagement tracking within an intelligent tutoring system. Frontiers in Artificial Intelligence, 3, 1-16.

D'Mello, S. K. & Graesser, A. C. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems, 2(4), 23: 1-38.

D'Mello, S.K., & Graesser, A.C. (2023). Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In P. Shutz and K.R. Muis (Eds), Handbook of Educational Psychology, vol. 4. Washington, D.C.: American Psychological Association.

Fang, Y., Lippert, A., Cai, Z., Chen, S., Frijters, J., Greenberg, D., & Graesser, A. (2022). Patterns of adults with low literacy skills interacting with an intelligent tutoring system. International Journal of Artificial Intelligence in Education, 32, 297-322.

Graesser, A.C. (2016). Conversations with AutoTutor help students learn. International Journal of Artificial Intelligence in Education, 26,124-132.

Graesser, A.C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In. B. Ross (Eds.), The Psychology of Learning and Motivation, vol. 57 (183-225). Elsevier.

Graesser, A.C., Forsyth, C., & Lehman, B. (2017). Two heads are better than one: Learning from agents in conversational trialogues. Teachers College Record, 119, 1-20.

Graesser, A.C., Greenberg, D., Olney, A.M., & Lovett, M.W. (2019). Educational technologies that support reading comprehension for adults who have low literacy skills. In D. Perin (Ed). Wiley adult literacy handbook (pp. 471-493). New York: Wiley.

Graesser, A.C., Hu, X., Nye, B.D., VanLehn, K., Kumar, R., Heffernan, C., Heffernan, N., Woolf, B., Olney, A.M., Rus, V., Andraskik, F., Pavlik, P., Cai, Z., Wetzel, J., Morgan, B., Hampton, A.J., Lippert, A.M., Wang, L., Cheng, Q., Vinsen, J.E., Kelly, C.N., McGlown, C., Majmudar, C.A., Morshed, B., & Baer, W. (2018). ElectronixTutor: An intelligent tutoring system with multiple learning resources. International Journal of STEM Education, 5:15, 1-21.

Graesser, A.C., & Li, H. (2022). Intelligent tutoring systems and conversational agents. In R. Tierney, F. Rizvi, K. Ercikan, and G. Smith (Eds.), International Encyclopedia of Education, edition 4. Elsevier.

Hampton, A. J., & Graesser, A. C. (2019). Foundational principles and design of a hybrid tutor. In R. A. Sottilare & J. Schwarz (Eds.) Proceedings of the First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference (pp. 96–107), Orlando, FL, USA, July 26–31, 2019.

Nye, B.D., Graesser, A.C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education, 24, 427–469.

Nye, B.D., Core, M., Swartout, B., Hu, X., Morgan, B., & Graesser, A. (2022). ElectronixTutor content and system testing to support adaptive learning for nuclear field electronics. Final report.

OECD. (2016). Skills matter: Further results from the survey of adult skills. OECD Skills Studies. Paris, France: OECD Publishing.

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinberg, J., & Chao, S.-F. (2019). SARA reading components tests, RISE forms: Technical adequacy and test design, 3rd edition (No. ETS RR–19-36). Princeton, NJ: Educational Testing Service.

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. Computers & Education, 147, 103781.

Wang, F., Li, W., & Zhao, T. (2022). Multimedia learning with animated conversational agents. In R.E. Mayer and L. Fiorella (Eds), The Cambridge Handbook of Multimedia Learning, Edition 3 (pp. 450-460). Cambridge, UK: Cambridge University Press.

CHAPTER 12 – CONSIDERATIONS FOR INTELLIGENT TUTORING SYSTEMS FOR MEDICAL EDUCATION

Susanne P. Lajoie[1] and Shan Li[2]
McGill University[1], Lehigh University[2]

Introduction

Intelligent tutoring systems (ITSs) are created based on a model of the learner (student model), the domain (expert model), and how to teach the topics in question (pedagogical or tutoring models). There is variation in how such ITSs are designed and such variation is often based on the learning theories that guide their design. When creating an ITS in medicine, the military, or other professions, it is essential that one realizes the high-stakes nature of assessing and fostering proficiency since improper assessments can lead to errors and learning outcomes that have great consequences (Lajoie, 2009). For several years, our research has been dedicated to fostering medical student learning in the context of BioWorld, an ITS that provides a safe practice environment for students to deliberately practice their diagnostic reasoning skills with virtual patient cases (Lajoie, 2009, 2021). Diagnostic reasoning in this context refers to the dynamic thinking process that leads to the identification of a diagnosis that best explains the clinical evidence (Szaflarski, 1997). Rigorous studies have been conducted to study the relationships between learners' cognition, emotions, motivation, and metacognition in the context of BioWorld. In this chapter, we discuss the interdisciplinary multimodal methodologies (computer interactions, verbal reports, facial expressions of emotions, and electro-dermal responses) used to provide evidence of the complex interplay of cognition and affect on learning in medicine. We will discuss our findings and provide more general recommendations for the future of ITSs.

Goals and Scope

It is often the case that school-learning does not readily transfer to real world applications. For example, one may be an excellent student but flounder when asked to apply schooled learning in practice. Lesgold et al. (1988) found that airmen who had top grades in circuit tracing courses, did not do well when troubleshooting an electronic fault on an aircraft. In medical school, the same situation applies, where students take basic science courses independent of opportunities to use this knowledge with real patients until they have passed their coursework. A common educational problem is that students do not have opportunities to apply what they learn in meaningful contexts where such skills should be applied (Greeno, 1998).

Many professions have a form of apprenticeship where trainees learn from more skilled others who model the skills that are needed to perform a task. However, traditional apprenticeships have their downfalls. In medicine for example, medical students learn from expert physicians, but they only see a limited number of patients and diseases, based on the specific weeks they are in rotation. Furthermore, not all experts are good teachers, and they may have difficulty articulating their knowledge in a manner that is easy for a trainee to follow. In the section below, we outline a cognitive apprenticeship framework that guides the design of the BioWorld ITS.

A Cognitive Apprenticeship Design for an ITS: BioWorld

The goal of a cognitive apprenticeship is similar to a traditional one in that trainees' apprentice within specific domains of study where more skilled others help them participate in real-world activities. A cognitive apprenticeship differs from a traditional one in that it has a structured framework to support learners in a situated learning environment. In particular, a cognitive apprenticeship consists of articulating the domain content knowledge and strategies needed to solve problems along with methods for effective teaching, appropriate sequencing of instruction and considerations for the sociology of the domain (Collins & Kapur, 2014). This framework is useful for ITS designers since the domain content knowledge and sequence of problems are clearly articulated along with rules for teaching and tutoring that are based on student profiles. These profiles determine the amount of expert modelling needed at a particular point in time, and are used to tutor, coach, scaffold, and fade support when learners reveal they can do things independently. Learners are given opportunities to reflect on and articulate their understanding.

BioWorld (Lajoie, 2021) was designed using the cognitive apprenticeship framework. BioWorld situates medical students in a virtual hospital setting where they review and diagnose patient cases by collecting appropriate patient data, such as patient symptoms and history, and conduct diagnostic tests to rule in (or out) their diagnoses. Trainees can consult a medical library and ask for a medical consult as they would at the hospital. Figure 1 presents the interface.

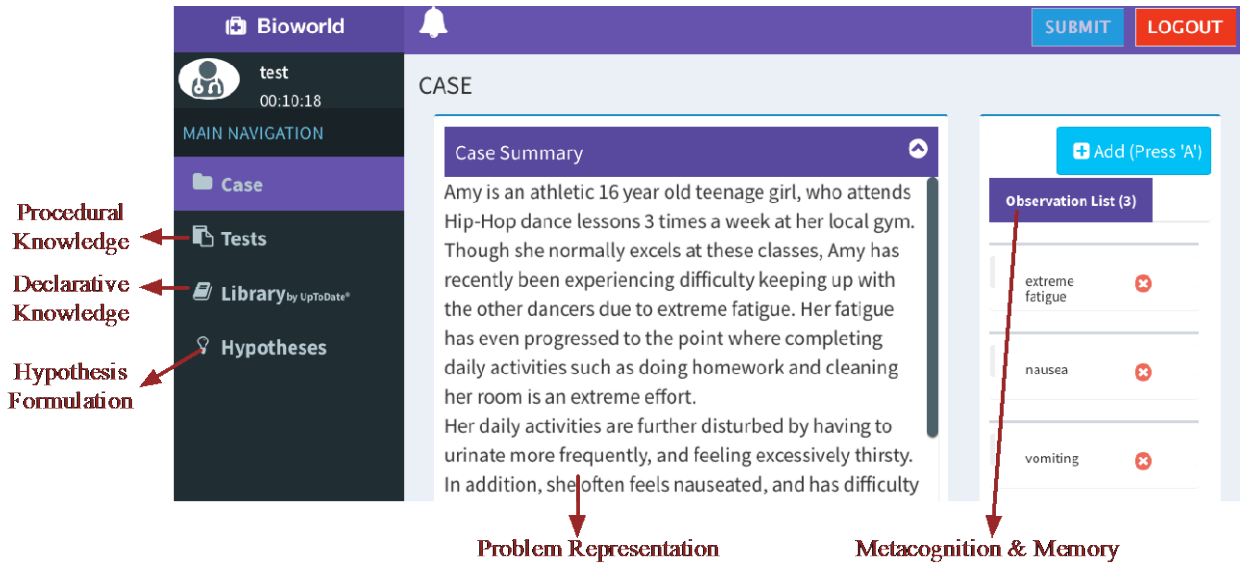


Figure 1. The Main Interface of BioWorld

The domain content knowledge and strategies were designed with subject matter experts in medicine, to ensure the created virtual patient cases had all of the necessary content embedded in the system. Rules for tutoring were based on modeling and assessment of student profiles that were measured against a series of best paths for solving problems that were created by expert physicians. An expert overlay was used to assess student models to determine levels of assistance and student proficiency levels. Students articulated their knowledge through their interactions with BioWorld. Log files were analyzed to determine the antecedents and consequences of their actions. Student beliefs and confidence levels were collected after

each diagnosis students made while solving a case. After they submitted a final diagnosis, they would rank their evidence in terms of significance to solving the case, and they would write a final case summary that could be handed to the next physician who would see the patient. A final student assessment was presented that indicated whether their final diagnosis was correct and the percentage of overlap their actions had with experts (level of expert proficiency).

BioWorld presents opportunities for deliberate practice by embedding expert models of diagnostic reasoning during and after problem solving. Students are encouraged to reflect on their knowledge during problem solving by explicitly posting their evidence in the evidence palette as they formulate their diagnoses. They also reflect on their beliefs and confidence in their diagnosis by selecting a percentage belief on the belief meter after each diagnosis. As mentioned above, after they submit their final diagnosis, they are presented with visual comparisons of their problem-solving process with that of experts. Together, this set of visualizations (evidence/observation palette, belief meter, and comparisons with expert problem-solving processes) present opportunities for reflection on their learning and helps to promote self-regulation during and after diagnosing patient cases (Lajoie et al., 2021a; Lajoie et al., 2021b). Students are tutored individually but the sociology of the medical environment is represented by the features that simulate the sociology of the hospital environment.

State of the Field and Supporting Research

Learning analytics has revolutionized the way we study ITSs in medical education. In the context of BioWorld, we explicitly look at the development of proficiency by exploring the relationships between behavioral, cognitive, metacognitive, and affective dimensions of diagnostic reasoning. We utilize multimodal data to provide such evidence and we use learning analytics to interpret the patterns and trajectories towards expertise.

Multichannel data, including self-reports, digital trace (log files), facial expression, eye movement, electrodermal activity (EDA), and concurrent think-aloud, are collected as students solve BioWorld cases. The collection of multimodal data allows us to capture the complexity of diagnostic reasoning and develop a holistic understanding of expert-novice differences in the dimensions of behavior, cognition, metacognition, and affect. Moreover, we aim at a precise understanding of the development of proficiency in diagnostic reasoning, and multimodal data helps to gain fine-grained insights into this process. Additionally, we learn that multimodal data is more accurate than a single data source, thus having the potential to address the mixed results in the literature and enhance the generalizability of our findings.

Apart from collecting multimodal data about learners and learning context, we use Artificial Intelligence (AI) techniques to advance our understanding of the relationship between learning to diagnostically reason, and the emotions that influence such learning. An important part of our work is to investigate medical students' behavioral patterns in solving patient cases, leveraging educational data mining and learning analytics techniques. As an example, we used recurrence quantification analysis (RQA) to examine the temporal structures of students' self-regulated learning (SRL) behaviors in diagnostic reasoning (Li et al., 2022c). We found that low performers had more single, isolated recurrent behaviors in problem-solving, whereas the recurrent behaviors of high performers were more likely to be part of a behavioral sequence. In addition to examining the temporal structures of SRL behaviors, we also examined the sequential patterns of students' SRL behaviors using sequential mining techniques (Li et al., 2022c; Zheng et al., 2021b). We found that the behavioral patterns of less efficient students were more disorganized compared to efficient students (Zheng et al., 2021b). Students in the less efficient group collected significantly more irrelevant evidence, ordered more lab tests, and proposed more incorrect hypotheses than efficient students. Additionally, our findings indicated that high performers were more

likely to demonstrate behavioral patterns that were cyclically sustained across the three SRL phases, i.e., forethought, performance, and self-reflection (Li et al., 2022c). Examining the expert-novice differences in behavioral trajectories lays a solid foundation for the design of feedback systems in ITSs, especially for those based on cognitive apprenticeship models. Providing feedback on the behavioral patterns and trajectories to the novice in real-time enables them to appreciate the usually unobservable differences in problem-solving patterns with experts.

AI techniques are also used to detect cognitive and metacognitive processes in learning with ITSs. In a recent study, we built machine learning models to predict students' cognitive engagement using their facial behaviors in diagnosing virtual patients with BioWorld (Li et al., 2021a). Specifically, we trained five types of supervised machine learning algorithms (i.e., Naïve Bayes, k-NN, decision tree, random forest, and support vector machine) on three categories of facial behaviors: eye-gaze, head pose, and facial action units. We found that the support vector machine model could accurately predict whether students were cognitively engaged in problem-solving or not. Moreover, we used a text mining technique to infer two types of metacognitive judgements, i.e., Feeling-of-Knowing (FOK) and Judgement of Learning (JOL), from students' think-aloud protocols (Lajoie et al., 2021a). We further examined the joint role of metacognitive judgement and achievement emotions in predicting diagnostic efficiency. The results suggested that FOK judgements positively predicted diagnostic efficiency, whereas JOL and the achievement emotion of anger negatively predicted diagnostic efficiency.

It becomes evident that we cannot claim a complete understanding of learning with ITSs without addressing the affective aspect of students' learning. In the BioWorld context, we use advanced techniques to detect and study emotions. For instance, we combined facial expressions and electrodermal activities to understand emotion dynamics in SRL (Zheng et al., 2022c). We found that students with better performance demonstrated more stable emotions in the forethought phase, less stable emotions in the self-reflection phase, and a higher level of emotional arousal in the self-reflection phase. We used growth curve modeling to examine how discrete academic emotions unfold in different phases of SRL and how the changes of these emotions influence learning performance (Zheng et al., 2022a). The results showed that curiosity and confusion declined across the three phases of SRL, whereas boredom increased in the self-reflection phase of SRL. The initial levels of curiosity and enjoyment positively predicted students' performance. In addition, we are interested in emotion variability and how it relates to students' performance in clinical reasoning. Specifically, we examined the changes of emotion variability in SRL phases and the differences in emotion variability between high and low performers (Li et al., 2021c). We found that high performers demonstrated less fluctuations of emotional states than low performers across the three SRL phases (i.e., forethought, performance, and self-reflection), although the differences between the two groups were not statistically significant. Emotion variability in the forethought phase influenced students' performance the most, compared to that in the performance and self-reflection phases. In sum, findings from our studies on emotion variability highlighted the importance of maintaining stable emotions in the three SRL phases and particularly the forethought phase to gain high performance.

Learning is essentially a complex dynamical system (Li et al., 2022b) whereby multi-components (behavioral, cognitive, metacognitive, and affective) interact with each other over time to yield an outcome. Therefore, we examine the interplay between learning components as students interact with the BioWorld environment (Lajoie et al., 2021b; Li et al., 2022a). As a representative example, we examined the co-occurrences of emotions and SRL behaviors in clinical reasoning (Lajoie et al., 2021b). Our study revealed that high and low performers differed on the co-occurrences of, and sequential transitions between, emotions and SRL behaviors. For instance, we found that the top ranked co-occurrence of SRL behaviors and emotions for low performers were elaboration and surprise, whereas evaluation and anger co-occurred the most for high performers. In another study, we used epistemic network analysis (ENA) to examine the interplays between SRL activities and the use of different types of knowledge (Li et al.,

2022a). We found that domain knowledge and metacognitive knowledge co-occurred most frequently, followed by the co-occurrence of domain knowledge and planning, regardless of the levels of task difficulty. Moreover, we found that high performers made more connections than low performers between metacognitive knowledge and domain knowledge, as well as between metacognitive knowledge and self-reflection, when solving the easy task. In contrast, low performers showed stronger connections between task information and other elements such as domain knowledge, planning, and evaluation than high performers. In the difficult task, high performers tended to make stronger connections between self-reflection and all three types of knowledge (i.e., task information, domain knowledge, and metacognitive knowledge) than low performers. There is no doubt that the complex interplay between learning components will receive increasing attention from SRL researchers in the future. As such, research on learning with ITSs will need to study these relationships between affect and SRL as well.

Discussion

We designed the BioWorld system for medical students to deliberately practice clinical reasoning skills. We performed AI and learning analytics techniques on the collected multimodal data to understand the complexity of clinical problem-solving in different dimensions, i.e., behavior, cognition, metacognition, affection, and their interrelationships. Particularly, we explored the differences in behavioral patterns, cognitive engagement, metacognitive judgements, emotion dynamics, and the interplay of learning components, between high- and low-performers. Our research on BioWorld has theoretical, methodological, and practical significance to the field of ITSs in medical education. Meanwhile, we see many opportunities moving forward.

First, there is a clear need to study emotion dynamics in learning with ITSs. In our previous work, we examined the dynamic aspect of students' emotions in SRL phases and found that different patterns of emotions are linked with SRL processes and diagnostic performance (Lajoie et al., 2018, 2021b; Li et al., 2021b; Zheng et al., 2021b). For instance, we found that epistemic emotions occur most frequently at the *forethought* phase of SRL. The *performance* phase is the right time for instructors to provide emotional support for students in the negative-boredom group to help them become more curious-positive. Achievement emotions should be attended to in the *self-reflection* phase of SRL, where instructors can help ease students' negative emotions induced by inferior performance (Zheng et al., 2022a). For future research, we argue that emotion dynamics can be studied at a more concrete level since the current emotion detection methods allow researchers to capture longitudinal and time-series data of emotions. Examining emotion dynamics could provide researchers with new insights about the temporal changes of students' emotional responses that go beyond the static approaches of studying emotions such as the frequency and duration of emotions. We address this gap by first introducing a taxonomy of emotion dynamics features, i.e., emotional variability, emotional instability, emotional inertia, emotional cross-lags, and emotional patterns (Zheng et al., 2022b). Furthermore, we present some predominant analytical techniques that can quantify emotion dynamics from longitudinal and time-series data.

Future research is needed to investigate feedback systems in ITSs that use timely predictions generated by advanced AI techniques and multimodal data (di Mitri et al., 2018). As aforementioned, we used machine learning algorithms to predict students' cognitive engagement states (i.e., engaged or less engaged) in real-time based on their facial behaviors (Li et al., 2021a). Taking that study as an example, there is a possibility that we can integrate such a cognitive engagement detection system into the BioWorld platform, which could provide instructors and students with timely feedback on students' engagement levels in problem-solving.

Recommendations for GIFT and Intelligent Tutoring Systems

The Generalized Intelligent Framework for Tutoring (GIFT) provides an extensive framework for authoring ITSs and is quite comprehensive in its consideration of the multi-componential nature of learning. It already considers multimodal data in constructing learner profiles that include cognitions, behavior, affect, and metacognition. GIFT uses such data to inform pedagogical strategies to improve learning.

As we learn more about the complex interactions between these complex features used to model human learning and performance, fine tuning of scaffolding will become even more prevalent. These nuanced assessments tell us the antecedents and consequences of emotions on cognitions and self-regulated learning; however, decisions around when, what, and how to scaffold during learning are still central to effective pedagogy. Although advances are being made in integrating affective computing into ITSs (D'Mello, 2013; D'Mello, Kappas & Gratch, 2018), there is still work to be done in automatically detecting and responding to student affect in an effective manner (Graesser, 2020).

As we saw in our own research, learning analytics can be used to classify clusters of performance sequences and determinations can be made regarding which sequences lead to successful performances and why. However, how do we use the temporal dynamics of SRL, emotion variability, type and location of physiological arousal, and indicators of cognitive engagement to give the right type of assistance? Empirical research will need to be conducted on scaffolding of cognitive competencies, along with scaffolds that regulate emotions so that learners stay on task. Furthermore, emotions and engagement vary based on the phase of SRL and consequently, different types of scaffolds for planning, performance and reflection need to be considered to keep individuals engaged. One question for consideration is what should a scaffold look like? Should the scaffold be to increase positive emotions such as curiosity and joy?; should scaffolds include reappraisal prompts when failures or impasses are reached?; should scaffolds focus on making the task content more accessible by providing more assistance around the cognitive skills?; should scaffolding be on specific phases of SRL (i.e. scaffolding specific to forethought, performance and reflection phases) and concurrently look at emotions in each of these phases to address better learning and performance?

Learning is not all or none and ITSs need to build learner profiles that capture the longitudinal changes in students' knowledge of themselves, and knowledge of the tasks and learning strategies (Matcha et al., 2019). Providing students with their real-time learning trajectories and patterns, and possibly a comparison with experts, is a promising direction that can enhance student awareness and reflection of their learning progress.

There are still many things to consider in ITSs for professionals, such as physicians-in-training. For example, the task itself is often high stakes, urgent, and lives can be at stake. The timeliness and appropriateness of feedback in urgent situations is critical for trainees to learn how to act under duress. The importance of the sociology of the workplace will need greater attention. When we consider modelling teams and groups the student modelling issues and scaffolding needs grow exponentially. However, learning analytics may help this process through the use of visualizations that externalize the team performance as a way for team members to reflect on their own and others performance (Zheng et al., 2021a). These visualizations are a form of scaffolding in themselves. Decisions will need to be made as to how to scaffold a team. For example, socially shared regulation (Järvelä et al., 2018) of a task is essential in team performance, where team leaders and team members each have a responsibility to work together to complete a goal, such as saving a patient in the operating room (OR). If one member, be it the leader or other team member, fails in their role, everyone may fail and safety is compromised. Furthermore, the group dynamics can impose its own challenges where emotion regulation strategies are needed in addition to fostering cognitive components of the task at hand (Lajoie & Poitras, in press). It is

likely that scaffolding can occur at the individual and team level by having specific pedagogical agents designed for each team member and perhaps a meta-agent to oversee the team itself, like a wizard of oz. We are beginning to see this type of approach in Lester et al.'s research (Saleh et al., 2020) and we have seen specific pedagogical agents for different features of SRL in Azevedo's research (Azevedo et al., 2022). It is likely that a network of agents can work in tandem to facilitate teamwork using a multi-componential profile to determine what needs to be scaffolded.

Conclusions

It has long been recognized that improving learning and performance is determined by recognizing the many ways that individuals learn. Each of us is motivated to learn for different reasons, we experience success and failure differently, and we have our own specific aptitude strengths and weaknesses along with our own specific prior knowledge in specific domains. Furthermore, our ability to manage our emotions during learning may depend on our abilities, the task, the consequences, or the people around us who support or challenge us positively or negatively. For this reason, student modelling is key. We conclude our chapter with these thoughts about student modelling as a way to lead us forward in future ITS designs. At the heart of student modelling is the need to identify differences in proficiency levels so that appropriate levels of adaptive feedback can be provided based on evidence of what a student knows or does not know in a specific context (Pellegrino et al., 2001). Pellegrino et al. (2001) referred to an assessment triangle where valid inferences could be made regarding learner cognition based on specific observations. However, "student modelling is beginning to transition into more inclusive models of learning where emotions, metacognition, and self regulation are considered important aspects of the learning transition" (Lajoie, 2021, *p*. 472). We are finally moving closer to developing computers that can care (Self, 1999) by linking models of learning and affect by tightly coupling the use of multi-componential data to provide appropriate assistance.

References

Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., Cloude, E., Dever, D., Wiedbusch, M., Wortha, F., & Cerezo, R. (2022). Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning with an Intelligent Tutoring System. *Frontiers in Psychology*, *13*, 813632. https://doi.org/10.3389/fpsyg.2022.813632

Collins, A. & Kapur, M. (2014). Cognitive apprenticeship. In R. K. Sawyer (Ed.). *The Cambridge handbook of the learning sciences* (pp. 109-127). NY: Cambridge University Press. http://ebooks.cambridge.org/ebook.jsf?bid=CBO9781139519526

di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*(4), 338–349.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4). https://doi.org/10.1037/a0032674

D'Mello, S. K., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement, *Emotion Review, 10*(2), 174-183.

Graesser, A. (2020). Emotions are the experiential glue of learning environments in the 21st century. *Learning and Instruction.* (1-5). https://doi.org/10.1016/j.learninstruc.2019.05.009

Greeno, J. G. & MMAP (1998). The situativity of knowing, learning and research. *American Psychologist, 53*(1), 5-26.

Järvelä, S., Hadwin, A., Malmberg, J., & Miller, M. (2018). Contemporary perspectives of regulated learning in collaboration. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.) *International handbook of the learning sciences* (pp. 127-136). Routledge.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *Development of Professional Expertise: Toward*

Measurement of Expert Performance and Design of Optimal Learning Environments (pp. 61–83). Cambridge University Press.

Lajoie, S. P. (2021). Student modeling for individuals and groups: the BioWorld and HOWARD platforms. *International Journal of Artificial Intelligence in Education*, *31*, 460-475. https://doi.org/10.1007/s40593-020-00219-x

Lajoie, S. P., Li, S., & Zheng, J. (2021a). The functional roles of metacognitive judgement and emotion in predicting clinical reasoning performance with a computer simulated environment simulated environment. *Interactive Learning Environments*, 1–12. https://doi.org/10.1080/10494820.2021.1931347

Lajoie, S. P. & Poitras, E. (in press). Technology rich learning environments: Theories and methodologies for understanding solo and group learning. In K. Muis & P. Schutz (Eds.). *Handbook of educational psychology, 4th edition. NY, NY:* Routledge. (invited)

Lajoie, S. P., Zheng, J., & Li, S. (2018). Examining the role of self-regulation and emotion in clinical reasoning: implications for developing expertise. *Medical Teacher*, 40 (8), 842-844.

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., & Gube, M. (2021b). Examining the interplay of affect and self regulation in the context of clinical reasoning. *Learning and Instruction*, *72,* 101219. https://doi.org/10.1016/j.learninstruc.2019.101219

Lesgold, A., Lajoie, S. P., Bunzo, M., & Eggan, G. (1988). Sherlock: A coached practice environment for an electronics troubleshooting job. *Technology and Learning, 2,* 1-3.

Li, S., Huang, X., Wang, T., Pan, Z., & Lajoie, S. P. (2022a). Examining the interplay between self-regulated learning activities and types of knowledge within a computer-simulated environment. *Journal of Learning Analytics*.

Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021a). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education*, *163*, 104114. https://doi.org/10.1016/j.compedu.2020.104114

Li, S., Zheng, J., Huang, X., & Xie, C. (2022b). Self-regulated learning as a complex dynamical system: Examining students' STEM learning in a simulation environment. *Learning and Individual Differences*, *95*, 102144. https://doi.org/10.1016/j.lindif.2022.102144

Li, S., Zheng, J., & Lajoie, S. P. (2021b). The frequency of emotions and emotion variability in self-regulated learning: What matters to task performance? *Frontline Learning Research*, *9*(4), 76–91.

Li, S., Zheng, J., & Lajoie, S. P. (2022c). Temporal structures and sequential patterns of self-regulated learning behaviors in problem solving with an intelligent tutoring system. *Educational Technology & Society*, *25*(4), 1–14.

Li, S., Zheng, J., Lajoie, S. P., & Wiseman, J. (2021c). Examining the relationship between emotion variability, self-regulated learning, and task performance in an intelligent tutoring system. *Educational Technology Research and Development*, 1–20. https://doi.org/10.1007/s11423-021-09980-9

Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226–245.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Saleh, A., Chen, Y., Hmelo-Silver, C., E., Glazewski, K. D., Mott, B.W., & Lester, J. C. (2020). Coordinating scaffolds for collaborative inquiry in a game-based learning environment. *Journal of Research in Science Teaching, 57* (9), 1490-1518. https://doi.org/10.1002/tea.21656

Self, J. A. (1999). The distinctive characteristics of intelligent tutoring systems research: ITSs care, precisely. *International Journal of Artificial Intelligence in Education, 10*, 350–364.

Szaflarski N. L. (1997). Diagnostic reasoning in acute and critical care. *AACN clinical issues*, *8*(3), 291–302. https://doi.org/10.1097/00044067-199708000-00002

Zheng, J., Huang, L., Li, S., Lajoie, S. P., Chen, Y., & Hmelo-Silver, C. E. (2021a). Self-regulation and emotion matter: A case study of instructor interactions with a learning analytics dashboard. *Computers & Education*, *161*, 104061.

Zheng, J., Lajoie, S. P., Li, S., & Wu, H. (2022a). Temporal change of emotions: Identifying academic emotion trajectories and profiles in problem-solving. *Metacognition and Learning*.

Zheng, J., Li, S., & Lajoie, S. P. (2021b). Diagnosing virtual patients in a technology-rich learning environment: A sequential mining of students' efficiency and behavioral patterns. *Education and Information Technologies*. https://doi.org/10.1007/s10639-021-10772-0

Zheng, J., Li, S., & Lajoie, S. P. (2022b). A review of measurements and techniques to study emotion dynamics in learning. In V. Kovanovic, R. Azevedo, D. Gibson, & D. Ifenthaler (Eds.), *Unobtrusive Observations of Learning in Digital Environments*. Springer.

Zheng, J., Li, S., Lajoie, S. P., & Wang, T. (2022c). Using facial expressions and electrodermal activities to understand emotion dynamics in self-regulated learning. *Manuscript Submitted for Publication*.

CHAPTER 13 - CAN THE USE OF INTELLIGENT TUTORS IMPROVE TACIT KNOWLEDGE TRANSFER IN EXPERIENTIAL LEARNING ENVIRONMENTS?

LisaRe Brooks Babin and Rebecca L. Robinson
Army University, Institutional Research and Assessment Division

Introduction

The current research effort discussed in this chapter is focused on understanding tacit knowledge transfer and how intelligent tutoring tools, like the Generalized Intelligent Framework for Tutoring (GIFT), might better prepare students for experiential education and training environments. It is hypothesized that tacit knowledge is best transferred from expert to novice when: there is a common vocabulary, concrete concepts, and a common operating picture that both the expert and the novice share to facilitate better productive discourse necessary for the exchange of nuanced skills, only acquired by doing. It is the exchange between the expert and the novice that is critical to tacit knowledge transfer, not simply the expert conveying information to a passive novice. The focus of this research is to identify advisor knowledge, skills, and abilities (KSAs) and develop a tool that will test tacit knowledge transfer of that knowledge. It also provides recommendations for future research investigating different aspects of the learning environment to improve tacit knowledge transfer and how intelligent tutors, like GIFT, may facilitate the learning process.

The study of tacit knowledge has been very limited since it was first defined by Michael Polanyi (1966), who is considered the father of tacit knowledge. Polanyi contrasted learning that occurs from reading with learning from doing. A more modern perspective would be the benefits of on-the-job training where a new hire takes advantage of the knowledge of an expert who gives insights into how to do the job well through repeated observation and continued discourse. For the military, on-the-job training is often how new personnel are trained. Unfortunately, there is limited understanding of tacit knowledge transfer across the services (Babin & Garven, 2019). The majority of research on this topic has been conducted by the United States Army Research Institute for the Behavioral and Social Sciences (ARI). From 1994 to 2008, several studies were conducted to ascertain how the military might understand and leverage tacit knowledge transfer more effectively, especially for officers (Antonakis et al., 2002; Avolio & Yammarino, 2003; Boyce et al., 2005; Cianciolo et al., 2001; Hedlund et al., 1998; 1999; 1999a; 1999b; 1999c; 2000; Horvath et al., 1994; Taylor et al., 2008).

The primary focus of the ARI research was on leadership KSAs. Army doctrine (FM 6-22; HQDA 2022) emphasizes the importance of leadership KSAs by stating, "Leadership is fundamental to Army operations as an element of combat power…" (preface). Further, FM 6-22 states in the introduction, "Army leaders are the competitive advantage the Army possesses that neither technology nor advanced weaponry and platforms can replace" (p. ix). Understanding leadership competencies and how to maximize the development of those competencies is essential to the Army maintaining that competitive advantage. Taylor et al. (2015) state "the nature of effective leadership in the military involves direct and indirect command and control of individuals, as well as small and large teams in a complex, rapidly evolving environment, where identifying leader and leadership development tools that can shape and develop effective leaders is critical" (p. 1). It is the direct and indirect ways that leaders engage with their Soldiers to convey their expert knowledge that is of interest to the current research effort.

Horvath et. al. (1994) conducted a literature review of tacit knowledge in the military to begin 14 years of research for ARI. They concluded that while effective leadership is the key to successful military

operations, the actual knowledge that a leader has and how it is used to direct personnel is not understood at all. Subsequent research by ARI began to provide clarity regarding what it means to be an effective leader, how to identify and measure leader tacit knowledge, and how to encourage the transfer of tacit knowledge from experts to novices. (Antonakis et al., 2002; Avolio & Yammarino, 2003; Boyce et al., 2005; Cianciolo et al., 2001; Hedlund et al., 1998; 1999; 1999a; 1999b; 1999c; Hedlund et al., 2000; Horvath & Sternberg, 1996; Taylor et al., 2008).

Building on the findings of the ARI research, this article focuses on advisor KSAs as a similarly nuanced skillset that is important to future multidomain operations. Advisor KSAs, like leadership, require practical intelligence acquired through real-life experiences (Sternberg, 1988). Whether it is during Security Force Assistance (SFA) training with international partners or Gender Advisors (GENAD) informing commanders and staff on gender considerations, for both jobs, effective decision making is essential to successful military operations (ATP 3-96.1; HQDA, 2022). Effective decision making is the result of intense study (explicit knowledge) and experiential learning (tacit knowledge). It is hypothesized in this chapter that enhancing explicit knowledge before an experiential learning event improves tacit knowledge transfer from experts to novices thus improving military readiness.

Similar to ARI's tacit knowledge research on leadership, this project started with a literature review of relevant KSAs and how those attributes are developed, measured, and employed within the Army. Unlike leadership, advisory attributes are not well understood. The primary military document describing SFA advisor's KSAs is the Security Force Assistance Brigade ATP 3-96.1 (HQDA, 2020). It details roles and responsibilities across multiple areas of expertise, and lists needed competencies like leadership, teamwork, communication, adaptability, and dependability. Additionally, desired competencies also included individuals that are empathetic, proactive, disciplined, and demonstrate endurance. Unfortunately, no guidance was given regarding how these competencies are developed. Brown (2018) drilled down on five characteristics that he identified as the most important to being a successful advisor: humble, empathic, self-aware, diplomatic, and having vision to effectively problem solve. He stated that "Advising is a separate and distinct skill from all other military specialties with the arguable exception of special operations" (p. 3). The advisor skillset makes for an ideal study of tacit knowledge transfer because of the multiple levels of expertise that is gained from intense study and experience.

Similar to the SFA advisor KSAs, GENADs must be good communicators, function well within a diverse team, and have strong cognitive and metacognitive skills for effective decision making. GENADs work with military commanders and staffs on a daily basis to integrate gender considerations into military, planning, operations, evaluations, and reporting. They often have to problem solve on the spot as military activities are being conducted requiring them to be flexible in their understanding of the fluid operational environment. Unlike for SFA advisors though, there is no Army doctrine or processes for identifying the needed competencies to be an effective GENAD for successful military operations. Some clarity can be achieved by reviewing our international partners' doctrine and training materials, but research needs to be conducted to ascertain the GENAD's needed KSAs which are critical to understanding how the knowledge of expert advisors can best be transferred during experiential education and training events (Nordic Defense Cooperation, 2022).

Once the KSAs have been identified for the advisor skillset, it is also important to understand how that knowledge is passed from expert to novice. Vast research has been conducted on expertise, but there are still many gaps in how to create learning environments that maximize expert knowledge and leverage that in novice training. An intriguing study was conducted by Gill (2021) titled "The Reciprocal Nature of Pedagogical and Technical Knowledge and Skill Development between Experts and Novices". In this article, the author emphasized the complex nature of the expert and novice relationship that is important to the transfer of tacit knowledge in educational environments. Specifically, the author describes the importance of technical knowledge and practice for both the expert and the novice, as well as the

relationship building between the two learners. Gill (2021) describes a reciprocal nature of the learning and developed a Sociocultural Contextual Framework to describe the actions and key components of the learning environment to maximize the tacit knowledge transfer and mutual development between expert and novice. In conclusion, Gill (2021) identified key components to an ideal learning environment that include: 1) developing mutually respectful and trusting relationships, 2) ensuring strong content knowledge for clear communications, 3) providing multiple opportunities to model correct performance, and 4) facilitating reflection over time. These factors facilitate relationship building and positive practice essential to effective tacit knowledge transfer.

The State of the Field

The U.S. Army is focused on understanding processes to improve individual, team, and unit performance for competing in multidomain operations globally. Army personnel come from diverse backgrounds and bring variable levels of KSAs to their jobs. Most Army assessments of KSAs are given at accessions or during classroom and training events, but the outcomes result often is a simple "Go/No Go" evaluation in the moment instead of proficiency over time. It would be a great improvement to better understand where a Soldier or Army Civilian performs on a continuum of expertise, instead of marking them only as a "pass" or "fail". A more sophisticated approach would allow for enhanced selection to learning events, improved tailoring of the curriculum for increased rigor, and potentially faster tacit knowledge transfer for improved readiness.

The proposed research project discussed in this chapter follows the scientifically established methodology developed by ARI researchers to measure tacit knowledge, mentioned earlier. The first phase is focused on identifying advisor KSAs to better understand the needed practical knowledge that advisors should acquire through directed study, on the job training, and experience via experiential learning and training environments (Horvath et al., 1996). As previously discussed, GENAD KSAs are not formally documented so training materials and interviews from subject matter experts must be conducted to fully understand the needed KSAs that should be targeted for novice training.

Additionally for the research project, work-related situations will be identified as the basis of the measurement instrument. The instrument is a situational judgment test (SJT) that is commonly used for measuring tacit knowledge (Antonakis et al., 2002). During interviews with GENADs, real-life scenarios will be developed based on their day-to-day job activities. These scenarios need to include enough information for the participant to come to logical and realistic conclusions on a continuum of expertise from novice to expert (Hedlund et al., 1998). Typically, each scenario will have between 5-20 response items. The response items can then be rated by the participant based on the quality of each response to the scenario presented. Additionally, participants can generate their own solution to the problem (Hedlund et al., 2000).

The development of the response items is an iterative process. A range of participants, from novice to expert, will be asked to provide their open-ended responses to the scenarios. The responses will then be sorted and combined into themes that can be evaluated using a rubric from "good" to "poor" responses. A panel of seasoned GENADS will evaluate the response themes and refine the response items to ensure that discreet responses can effectively discriminate novice from expert performance.

Once the scenarios and response items have been approved by the subject matter experts, the validation phase of the research will begin. Students attending basic advisor courses, who volunteer to participate in the research, will be given the measurement and their responses will be evaluated based on the expert panel's ratings of good to not so good performance. Ideally, clear break points will be identified to reliably identify levels of advisor expertise. Continued refinement of the measure may be needed to ensure it reliably

identifies advisor ability along a continuum from novice to expert. Previous research by Antonakis et al. (2002), developed between 13 and 19 scenarios for their different leadership SJTs.

Once validated, the final assessment tool can be utilized in a number of different ways, whether it is for selection to a course, refinement of the curricula shaped to student performance as a premeasure, or a talent management tool for job placement. Further experimentation can be conducted to better understand influencing factors facilitating tacit knowledge transfer between experts and novices. Of particular interest to this project, is the manipulation of levels of concrete knowledge developed before an experiential learning or training environment.

Recommendations for GIFT and Intelligent Tutoring Systems

One way to enhance concrete knowledge acquisition is to use intelligent tutors like GIFT. GIFT is a free open-source intelligent tutoring system (ITS) framework developed by US Army DEVCOM Soldier Center that can be used to create computer-based adaptive training (Sottilare & Goodwin, 2017). GIFT includes authoring tools that allow instructors to incorporate their existing content for use and provides a straightforward way for them to author adaptive tutoring lessons in any topic area (Sinatra et al, 2022). There are many benefits to using GIFT from creating linear online lessons as a prerequisite to an experiential learning activity to developing adaptive modules that include remediation based on questions that a student misses for performance improvement before or during traditional classroom instruction.

The findings from the current research could inform further use of GIFT to target the development of needed KSAs for military students in general, and advisors, in particular. By developing SJTs for specific jobs, that information could then inform training tools and assessments in GIFT, expanding the applications of this intelligent tutoring tool. Developing SJTs can be quite time-consuming, but if automated and aligned to current GIFT training capabilities, it could be a more efficient way to identify needed skillsets for specific actions or overall jobs. Additionally, there may be overlap of needed skillsets across military jobs that could maximize training for larger groups of related learners. For instance, if the current research effort identifies that good communication skills are needed to be a good advisor, as is theorized above, then ensuring that tutoring tools enhance communication skills would be an important next step. This communication feedback could also be great for other military jobs that have been identified as needing good communication skills like Public Affairs Officers, Civil Affairs Soldiers, etc. While the jobs may be different, elements of the jobs may be the same and can be maximized by targeted tutoring modules aligned to that KSA. Additionally, knowing a student's strengths and weaknesses going into a training session might help improve the ITS' adaptive capabilities to increase the amount of learning during the event.

Future research might take an experimental approach using GIFT to manipulate concrete knowledge about any topic before an experiential learning event and measure levels of tacit knowledge that were transferred from experts to novices using the SJT. Students could be assigned to one of two groups: 1) the treatment-as-usual group (no explicit knowledge required) and 2) the enhanced treatment group (must pass a pre-test of explicit knowledge of specific doctrine, structure, and terminology before attending the different courses). Student evaluations could be collected from the instructors, video of the practical exercises could be analyzed, and peer evaluations could be collected to ascertain performance measures of each student during the course. Lastly, students could be given a tacit knowledge test to access the level of tacit knowledge they acquired during the course. The performance and test scores would be compared across the groups to analyze the impact, if any, of the treatment on students becoming more expert. Additionally, if feasible, longitudinal data could be collected on the students to compare how they performed in their actual jobs, depending on the group they were assigned.

If the development from novice to expert can be achieved more efficiently and rapidly using tools such as GIFT, military readiness for future operations, particularly for difficult to learn tasks, can be vastly improved– giving the U.S. military overmatch of peer and near-peer competitors.

References

Antonakis, J., Hedlund, J., Pretz, J. E., & Sternberg, R. J. (2002). Exploring the nature and acquisition of tacit knowledge for military leadership (ARI Research Note 2002-04). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA400486)

Avolio, B.J., & Yammarino, F.J. (2003). Development of officer leadership for the Army: Preliminary results. (ARI Contractor Report 2004-01). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Babin, L.B., & Garven, A.J. (April, 2019). Tacit knowledge cultivation as an essential component of developing experts. Journal of Military Learning. Ft. Leavenworth, KS: Army University Press. https://www.armyupress.army.mil/Portals/7/journal-of-military-learning/Archives/April-2019/Babin-Garven-Tacit-Knowledge.pdf

Boyce, L. A., Wisecarver, M. M., Zaccaro, S. J. (2005). Understanding, predicting, and supporting leader self-development (ARI Technical Report 1173). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA442647)

Brown, E.E. (December, 2018). The definition of Advisor: Comprehending the mission to advise Foreign Security Forces. Small Wars Journal. McLean, VA: Small Wars Foundation. https://smallwarsjournal.com/jrnl/art/definition-advisor-comprehending-mission-advise-foreign-security-forces

Cianciolo, A. T., Antonakis, J., & Sternberg, R. J. (2001). Developing effective military leaders: Facilitating the acquisition of experience-based, tacit knowledge (ARI Research Note 2001-11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA400614)

Gill, D.D. (2021). The reciprocal nature of pedagogical and technical knowledge and skill development between experts and novices. Design and Technology Education: An International Journal, 26, 46-65.

Hedlund, J., Horvath, J. A., Forsythe, G. B., Snook, S., Williams, W. M., Bullis, R.C., Dennis, M., and Sternberg, R. J., (1998). Tacit knowledge in military leadership: Evidence of construct validity (ARI Technical Report 1080). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedlund, J., Sternberg, R.J., Horvath, J.A., Forsythe, G.B., & Snook, S. (1999). Tacit Knowledge for military leaders: Lessons learned across organizational levels. (ARI Research Note 99-29). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedlund, J., Sternberg, R. J., & Psotka, J. (2000). Tacit knowledge for military leadership: Seeking insight into the acquisition and use of practical knowledge (ARI Technical Report 1105). Alexandria, VA: U.S. Army Research Institute.

Hedlund, J., Williams, W.M., Horvath, J.A., Forsythe, G.B., Snook, S., Wattendorf, J., McNally, J.A., Sweeney, P.J., Bullis, R.C., Dennis, M. & Sternberg, R.J. (1999a). Tacit knowledge for military leaders: Company commander questionnaire (Research Product 99-08). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedlund, J., Williams, W.M., Horvath, J.A., Forsythe, G.B., Snook, S., Wattendorf, J., McNally, J.A., Sweeney, P.J., Bullis, R.C., Dennis, M. & Sternberg, R.J. (1999b). Tacit knowledge for military leaders: Platoon leader questionnaire (Research Product 99-07). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedlund, J., Williams, W.M., Horvath, J.A., Forsythe, G.B., Snook, S., Wattendorf, J., McNally, J.A., Sweeney, P.J., Bullis, R.C., Dennis, M. & Sternberg, R.J. (1999c). Tacit knowledge for military leaders: Battalion commander questionnaire (Research Product 99-09). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Horvath, J.A. & Sternberg, R.J. (1996). Tacit Knowledge in Military Leadership: Supporting Instrument Development. (ARI Technical Report 1042), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Horvath, J.A., Williams, W.M., Forsythe, G.B., Sweeney, P.J., Sternberg, R.J., McNally, J.A., & Wattendorf, J. (1994). Tacit knowledge in military leadership: A review of the literature. (ARI Technical Report 1017), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

HQDA. (September 2020). Field Manual Army Techniques Publication 3-96.1, Security Force Assistance Brigade. Washington, DC: Headquarters, Department of the Army.

HQDA. (November 2022). Field Manual 6-22, Army leadership. Washington, DC: Headquarters, Department of the Army.

Nordic Defense Cooperation. (2022). Gender Advisor Course Curricula. https://www.forsvarsmakten.se/siteassets/english/swedint/engelska/course-curricula/genad.pdf

Polanyi, M. (1966). The tacit dimension. Chicago, IL: University of Chicago Press.

Sinatra, A.M., Robinson, R., Goldberg, B., & Goodwin, (July, 2022). Generalized Intelligent Framework for Tutoring (GIFT) Master Gunner Course Pilot. Poster for the 2022 ArmyU Learning Symposium at Fort Leavenworth, KS.

Sottilare, R. & Goodwin, G. (October, 2017). Adaptive instructional methods to accelerate learning and enhance learning capacity. Conference presentation at the International Defense and Homeland Security Simulation Workshop, Barcelona, Spain.

Sternberg, R.J. (1988). The triarchic mind: A new theory of human intelligence. New York: Viking.

Taylor, T.Z., Psotka, J., & Legree, P. (2015). Relationships among applications of tacit knowledge and transformational/transactional leader styles: N exploratory comparison of the MLQ and TKML. *Leadership & Organization Development Journal*, *36*(2), 120-136.

Taylor, T.Z., Higley, L., & Grabarczyk, D. (2008). A U.S. army reserve noncommissioned officer (NCO) tacit knowledge inventory: Flexible structure for squad-level leader self-development (ARI Research Product 2008-01). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

CHAPTER 14 - AUTHORING TOOLS FOR CROWDSOURCING FROM TEACHERS TO ENHANCE INTELLIGENT TUTORING SYSTEMS

Li Cheng[1], Ethan Prihar[1], Sami Baral[1], Ashish Gurung[1], Anthony T. Botelho[2], Aaron Haim[1], Cristina Heffernan[3], Thanaporn Patikorn[4], Adam Sales[1], and Neil T. Heffernan[1]
Worcester Polytechnic Institute[1], University of Florida[2], The ASSISTments Foundation[3], Rajamangala University of Technology Suvarnabhumi[4]

Introduction

Free and accessible authoring tools for crowdsourcing are a great resource for creating intelligent tutoring systems that can leverage the power and strengths of the crowd which includes students, teachers, and researchers. Currently, most authoring tools for Intelligent Tutoring Systems are created by educational technology companies for use by their employees. Committed to open science, we have taken a broader and more inclusive approach.

Several years ago, our article, *The Future of Adaptive Learning: Does the Crowd Hold the Key?* in the International Journal of Artificial Intelligence (see Heffernan et al., 2016) promulgated a vision for every teacher (and even every learner) to be given access to the authoring tools in order to create learning content. Furthermore, we believe that running randomized controlled trials on the crowdsourced content to allow newly created learning content to compete with prior content is equally important to optimize learning content and advance the learning sciences. In essence, we think an authoring tool is the combination of a functionality that allows authors to contribute learning content and an engine that operates randomized controlled trials or machine learning to examine which content is more effective and what content works for which students. We believe that this combination is crucial for making a complete authoring ecosystem for creating intelligent tutoring systems that provide personalized learning.

The increasing popularity of intelligent tutoring systems has called for more access to authoring tools for these learning environments so as to leverage the strengths of crowdsourcing. Since 2016, Dr. Neil Heffernan and his team have made more advancements in creating free authoring tools for crowdsourcing in ASSISTments, which is an online intelligent tutoring system that provides assistance for K-12 students and also provides learning assessments data which teachers can use to improve classroom instruction; and has been acknowledged as highly effective with Tier 1 rating from Evidence for ESSA (U.S. Department of Education, 2015).

This book chapter aims to demonstrate the authoring tools of crowdsourcing that have been designed and implemented in ASSISTments and how we help teachers use the tools, and provide an updated vision for crowdsourcing in intelligent tutoring systems. Specifically, we provide a brief background on crowdsourcing for intelligent tutoring systems. We then highlight our endeavors on crowdsourcing authoring tools in the ASSISTments platform through two infrastructures: TeacherASSIST, which harnesses the power of teachers by providing them an authoring tool to add student-supports such as hints and explanations; QUICK-Comments, which combines crowdsourcing and machine learning/artificial intelligence (AI) to assist teachers in providing feedback to students' open-ended responses. Following the discussion on our crowdsourcing authoring tools and how we help teachers use them, we conclude this book chapter by sharing our insights on leveraging crowdsourcing to improve intelligent tutoring systems and suggestions on helping teachers use crowdsourcing tools.

Crowdsourcing and Intelligent Tutoring Systems

The idea of crowdsourcing has developed since the early 2000s when Howe (2006) discussed using Web 2.0 tools to engage the crowds in performing tasks. Some common examples of crowdsourcing include Wikipedia, Stack Overflow, and Reddit. These platforms aggregate content from their communities and allow everyone to create and edit the content. With the rapid development of information communications and technologies (ICT), crowdsourcing has gained popularity in the education domain. Through a systematic review on crowdsourcing in education, Jiang et al. (2018) developed a definition of crowdsourcing in education as "a type of online activity in which an educator, or an educational organization proposes to a group of individuals via a flexible open call to directly help learning or teaching" (p. 3). We concur with this definition as it emphasizes that crowdsourcing is an online activity to gather input from a group of individuals to improve learning and/or teaching. While the definition does not specify who is in the group, we believe the individuals could range from experts to novices. The crowd does not necessarily have to be comprised of experts. Even a crowd of novices may serve as a helpful resource (Heffernan et al., 2016).

Crowdsourcing from novices such as learners not only provides a large amount of content that could be made useful for future learners but can also benefit the learners who contribute content (Heffernan et al., 2016). As an expert in learnersourcing, Juho Kim has done exemplary work in crowdsourcing from learners and found that learnersourcing is beneficial to enabling more interactive, collaborative, and data-driven learning (Kim, 2015). A study by Juho Kim and the team found that learner-generated reflective summary answers regarding information about a video can generate a video outline with subgoals that were comparable in quality to expert-generated subgoals, and the learners were more engaged and had a better understanding of the learning material (Weir et al., 2015). In another study that asked learners to generate, revise, and evaluate explanations when solving a problem, Williams and colleagues (2016) found the crowdsourced explanations were perceived as helpful by future learners, had the same quality as the ones generated by an experienced instructor, and enhanced learning of the learners who provided explanations when compared to solving problems and receiving answers without providing explanations.

A variety of research has been conducted to crowdsource with learners or more knowledgeable crowds. Research suggests many benefits of crowdsourcing in education, such as creating educational contents, providing practical experience, facilitating the exchange of complementary knowledge, and providing abundant feedback (Jiang et al., 2018). Specifically, in an intelligent tutoring system, crowdsourcing content contributions from users, not just designers, has a great potential to expand the breadth and diversity of learning materials and learning support (Heffernan et al., 2016), which further contributes to personalized learning in intelligent tutoring systems. Many researchers have studied crowdsourcing in education (see Alenezi & Faisal, 2020; Jiang et al., 2018) and some researchers used machine learning to analyze crowdsourced content (e.g., Kamath et al., 2016; Krause et al., 2017; Williams et al., 2016) which may inform personalized learning. However, there is a scarcity of studies on crowdsourcing directly implemented for creating intelligent tutoring systems. Floryan and Woolf (2013) used crowdsourcing to collect and analyze previous students' work within an intelligent tutor and used an intelligent algorithm to coalesce data to automatically construct expert knowledge bases to be used by future students. They compared human-created knowledge bases with the crowdsourced expert knowledge bases and found that crowdsourced expert knowledge bases had qualities similar to that of human-crafted knowledge bases and were generated in significantly less time. In another example, Khosravi et al. (2019) developed a platform called RiPPLE (Recommendation in Personalized Peer-Learning Environments) to recommend personalized learning activities to students based on their knowledge state from a pool of crowdsourced learning activities that were generated by educators and the students themselves. Evaluation of the platform showed students had measurable learning gains and perceived the platform as beneficial for supporting their learning.

Although crowdsourcing has gained attention from researchers, there is still very limited evidence on how crowdsourcing contributes directly to intelligent tutoring systems. Even less is known about how to make crowdsourcing accessible for the stakeholders by providing authoring tools. A systematic review on authoring tools for designing intelligent tutoring systems by Dermeval et al. (2018) suggests that a powerful intelligent tutoring system relies on the combination of artificial intelligence and human intelligence. Providing free and accessible authoring tools for crowdsourcing is a promising strategy to harness human intelligence of the crowd to improve intelligent learning systems. Over the past few years, we have designed and developed several infrastructures in ASSISTments to crowdsource teachers' and researchers' input by providing free and convenient authoring tools for teachers to create student-supports such as hints and explanations and for researchers to develop, deploy, and disseminate educational studies. In the next section, we provide an overview of two crowdsourcing authoring tools we have developed.

Authoring Tools for Crowdsourcing to Support Students and Teachers

The authoring tools we have developed for crowdsourcing support both students and teachers. We designed and developed an authoring tool to crowdsource from teachers to provide just-in-time support for students. The student support includes hints, explanations, and scaffolding questions for a given problem; feedback messages to common wrong answers for a given problem; and YouTube videos that explain a concept. We also use crowdsourcing to learn the common questions students ask and answers for the questions. In addition to supporting students, we use crowdsourced instructional recommendations to support teachers. The support for teachers for a given problem includes hints and explanations, responses to common wrong answers, and feedback on students' open responses. As some mathematics curricula do not have a sufficient quantity of problems for students, we also crowdsource from teachers to develop new and better problems. Furtherly, we use a model of student affect with sensor-free detectors to track students and crowdsource feedback that teachers can give to students (Botelho & Heffernan, 2019). Finally, we crowdsource ideas from researchers through an Ed Tech Research Infrastructure to Advance Learning Science (E-TRIALS) Platform, which scales up research to fundamentally improve educational research and stimulate theory- and evidence-based improvements on the ASSISTments platform. In the next subsections, we highlight two of our crowdsourcing projects with teachers.

Teachersourcing On-Demand Assistance for Students: TeacherASSIST

Beginning in 2018, ASSISTments created the TeacherASSIST program, which was inspired by Chris LeSiege, a teacher in Maine, who made a comment for every math problem in an entire textbook for his students. Dr. Heffernan created TeacherASSIST to support teachers to write hints or explanations for their students and also share with students of other teachers. TeacherASSIST crowdsourced content from teachers who had been creating their own tutoring messages for their students and redistributed their content to students outside their classes (Patikorn & Heffernan, 2020). This content came in the form of hints and explanations for middle school mathematics problems. Figure 1 shows an example of two tutoring messages written for the same problem by two different TeacherASSIST teachers (Prihar et al., 2021).

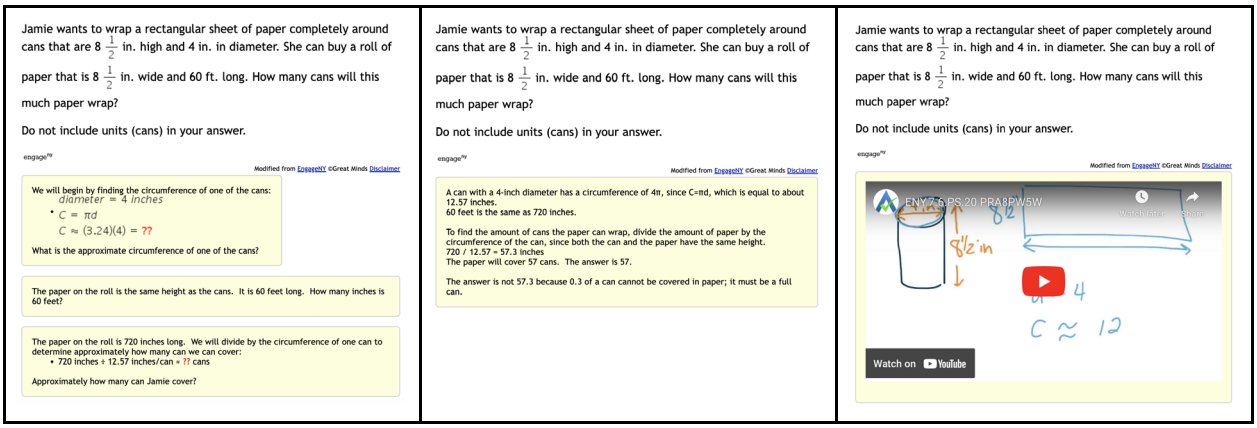


Figure 1. Two Tutoring Messages as Seen by A Student Using ASSISTments.
***Note*. The top is the problem body and the message with a yellow background is the tutoring message. The left tutoring message is a series of hints, the middle message is an explanation in text, and the right message is an explanation with video.**

Through TeacherASSIST, 40,000 new instances of tutoring for about 26,000 different problems were aggregated from teachers already creating content for ASSISTments. From 2018 through 2020, two large-scale randomized controlled experiments were conducted in which students were randomized on a per problem basis between receiving crowdsourced tutoring or just the answer when they were struggling. It was found that providing crowdsourced tutoring messages to students had a statistically significant positive impact on their learning compared to only providing students with the answer to the problem they were struggling with (Patikorn & Heffernan, 2020). Since the publication of these findings, ASSISTments has scaled up the distribution of crowdsourced content within the platform. The same experiment was repeated after scaling up using data from 2020 to 2021 and the findings of the original study were confirmed through this repeat analysis. (Prihar et al., 2021). Additionally, the repeat analysis was able to identify differences in the quality of different content creator's tutoring messages, which can be used to select from multiple tutoring messages with the intention of maximizing students' learning.

After identifying that some content had a greater benefit to students than other content, ASSISTments moved to using reinforcement learning to determine the most effective tutoring messages. Thompson sampling, a multi-armed bandit algorithm, was used to learn over time which tutoring message was most effective for each problem. While Thompson sampling was able to increase students' next problem correctness over a two month trial period, using Thompson sampling prevented the use of statistical techniques that rely on the independence of samples to compare the quality of tutoring. Moving forward, a balance must be found between optimizing student learning and statistically evaluating the quality of different tutoring messages.

Crowdsourcing and Machine Learning to Support Teachers: QUICK-Comments

Rapid growth and development in the domain of Natural Language Processing (NLP) and Machine Learning methods over the past few years have facilitated innovation in automated grading and feedback generation for responses to open-ended problems. One of the most prominent areas where we can realize the power of NLP is in the *'Google Smart Reply'* tool which reviews the content of an email and recommends a response. Various learning platforms have leveraged the power of NLP primarily in the assessment and feedback generation of long-form written responses (Allen et al., 2016; McNamara et al., 2013; Roscoe et al., 2019); however, responses to open-ended questions in mathematics tend to be short,

precise, and often contextually sparse as there is an expectation that the teacher can inherently infer the context. This presents a unique, albeit challenging, context for leveraging NLP in the automated assessment and feedback generation for student open-responses.

Our lab analyzes this unique challenge through the QUICK-Comments project by crowdsourcing teacher authored feedback messages and scores to open-ended problems. We utilize open-response assessment data from teachers (both numeric score and feedback messages to students' open-ended work) and implement various machine learning and NLP techniques to aid teachers in evaluating new open-ended works. For the development of this tool, we (Erickson et. al, 2020) explored NLP methods such as bag-of-words, term frequency-inverse document frequency (tf-idf) and GloVe combined with various machine learning approaches like Random Forest, Long Short Term Memory (LSTM) to predict scores to be given to open-ended responses and have shown strong model performance in this prediction task. The best overall model from the study (Erickson et al., 2020) was the Random Forest Model with the AUC score (multi-class ROC AUC) of 0.850, Root Mean Squared Error (RMSE) score (calculated over ordinal prediction and label) of 0.615 and multiclass Kappa of 0.430. To better understand students' textual responses and further improve the auto-assessment method for open-responses, we explored other NLP approaches based on sentence-level semantic representations in Baral et al. (2021). The method called SBERT-Canberra which utilizes a pre-trained model of SBERT and is based on contextual similarity of student responses, outperformed previously developed approaches for score prediction across all three evaluation metrics used in Erickson et al. (2020); with the AUC score of 0.856, RMSE score of 0.577 and Kappa of 0.476. The SBERT-Canberra method is further extended in the QUICK-Comments tool to predict feedback messages to give to students' open responses (Botelho et al., 2023).

Manually assessing open-ended problems introduces a fair amount of complexity to teachers as there can be multiple correct answers. The responses explain students' understanding of a particular topic that requires the teacher to infer their understanding. With the goal of easing this process by automating open response assessment for teachers, the QUICK-Comments tool assesses the student response and automatically suggests a numeric score and a set of feedback messages for the teacher to choose from. Figure 2 shows the QUICK-comments tool inside ASSISTments, where for each student response, the teacher receives a suggested score and a set of feedback messages to select from. The teacher can either select one of the suggested feedback messages, edit the selected feedback message, or write a new feedback message of their own. Starting April 2021, we began closely monitoring 47 teachers who are actively using the QUICK-Comments. Thus far, the teachers have graded ~46,000 student responses and provided about ~47,000 unique feedback messages while utilizing the QUICK-Comments tool. Currently, we are working on releasing QUICK-comments to a more significant cohort of teachers and iteratively work on improving the efficacy of QUICK-Comments. With the QUICK-Comments tool, we continue to crowdsource these assessment data from the teachers by collecting the scores and feedback messages even during the instances when they disagree with the recommendations from our models.

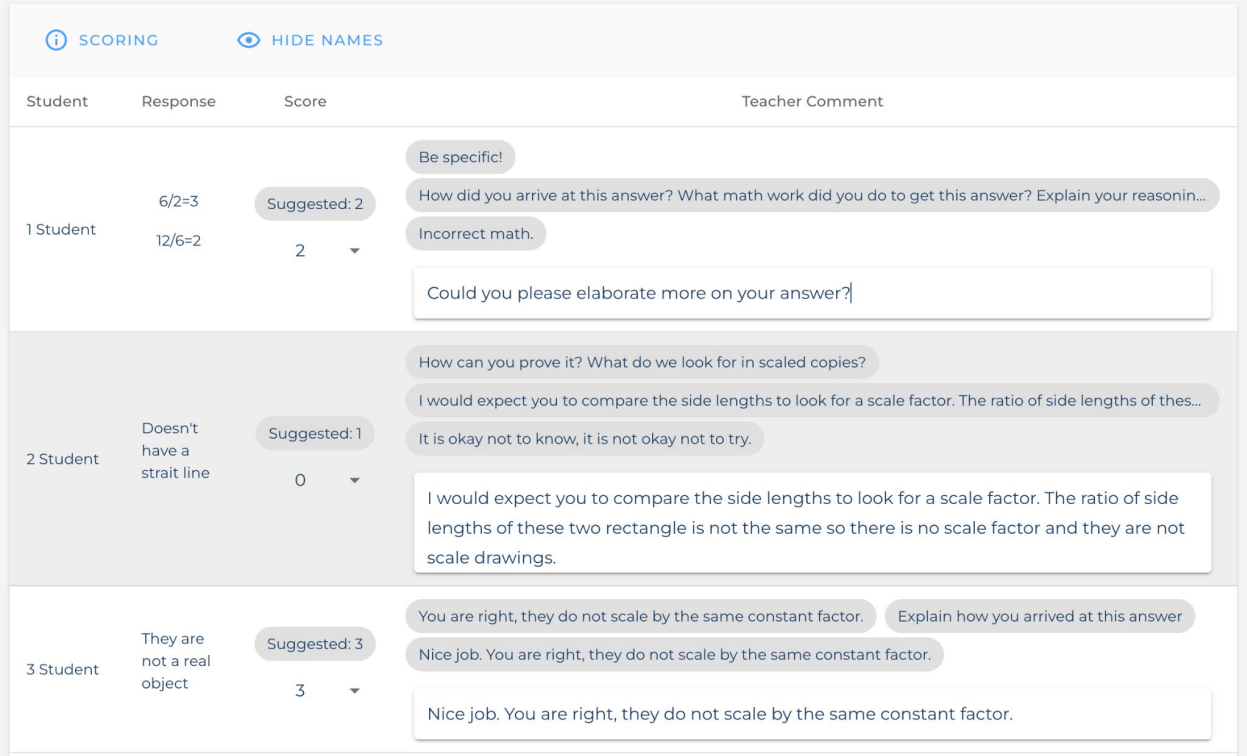


Figure 2. Teacher's View of the Open-response Scoring Page Inside ASSISTments.
Note. **The QUICK-Comments tool suggests automated score and feedback messages to student open responses.**

While the opportunity to analyze the teachers' reaction to the recommendation of the NLP model provides us with an opportunity to improve the model performance, this interaction also provides us with an excellent opportunity to explore the Human-AI interaction, which will help us explore various aspects such as fairness, transparency, and explainability of the model, the influences of the teachers' perception of the AI model, and how the model's performance varies with the change in teacher perception of the model. We explored the presence of possible algorithmic biases caused by the NLP models and did not detect biases in any particular direction (Erickson et al., 2020). Currently, our team is exploring the existence of biases at a fundamental level where the biased grading behavior of teachers might have corrupted the crowdsourced data.

Helping Teachers Use the Authoring Tools and Produce Good Content

Now that we have the crowdsourcing authoring tools, how do we help teachers use them to produce good content? After recruiting teachers to contribute to our crowdsourcing projects, we hosted a series of webinars to provide professional development for them in conjunction with providing guideline documents and tutorials and building a network of teachers. Our effort has been focused on fostering a community with teachers, helping teachers write good content, and orienting teachers with the authoring tools.

Fostering a Community with Teachers

Before diving into the specific crowdsourcing content area, we laid a foundation by fostering a community with teachers. We advocated values that communities hold as important such as being present and engaged,

striving for equity of voice, using strengths-based language and contributing to a safe work environment in which teachers feel comfortable to learn what they do not know. By communicating these values with teachers and having teachers share what they value in a collective workspace, we fostered a community that made teachers feel safe and embraced to work together. We also used a Slack channel to facilitate communication and build up the community. As we fostered a community with teachers, we moved forward with helping them write good content.

Helping Teachers Write Good Content

ASSISTments provides hints, explanations, and common wrong answer feedback for math problems to support students when they need help, address student misconceptions in real time, and increase students' ability to self-direct and enable them to take ownership of their learning in the online environment. We helped teachers understand how ASSISTments works and why it is important to provide student support. As an example, we illustrate how we helped teachers write hints for math problems. We explained what hints are and showed them how hints work in ASSISTments then we provided some example hints and guidelines on hints. Teachers learned about the four steps of writing hints. First, read the unit/module narrative to identify the essential understandings and progression of the unit, vocabulary from the curriculum, and models and strategies from the curriculum. Second, read the entire lesson to identify where it fits in the unit and anticipated misconceptions from the lesson guidance. Third, read the math problems and solve the problems, read the common wrong answers and figure out how the students might have gotten them, and review the answer key for the problem set. Finally, write hints that are specific to each individual problem. Besides the four general steps, we also provided guidelines on how to write good hints, such as using curricular language, focusing on strategy, making hints accessible by using brief and accurate language and simple sentence structure at grade-level. These processes and guidelines focused on the importance of understanding the math content and using strategies that are appropriate in their hints. To further help teachers understand hints, we used Zoom breakout rooms to engage teachers in critiquing some hints in groups and debriefed their thinking and how the guidelines work. Teachers were assigned a task to practice on writing hints for math problems. We hosted follow up webinars to engage teachers in reflecting on and revising their hints to help them write high-quality hints and further develop a network among them.

Orienting Teachers with Authoring Tools

Upon helping teachers gain the knowledge and skills on writing good student support such as hints for math problems, we oriented them with using the authoring tools. We created accounts for them to access the tools and provided tutorial videos for logging into the tools and writing hints in the builder. To increase accessibility of our materials, we also provided text-based instructions for the step-by-step directions of the process of writing hints in the builder. There was a practice assignment in the builder to make teachers feel comfortable using the tool. Teachers were also recommended to use Grammarly to check the readability level of their hints. As this might be the first time some teachers are using Slack, we provided a quick start guide and tutorial videos to help them get familiar with using Slack. Follow-up webinars and technical support were provided to help teachers. We made sure teachers were comfortable with any of the technologies associated with the work they were doing with us.

Vision for the Future of Crowdsourcing for Intelligent Tutoring Systems

Crowdsourcing has gained a great deal of attention from researchers and investments from governments and agencies. It is a promising strategy for creating and improving intelligent tutoring systems. Our vision for the future of crowdsourcing in intelligent tutoring systems centers on four aspects: crowdsourcing and optimizing content in intelligent tutoring systems, leveraging AI to crowdsource live while tutoring is

happening, using crowdsourcing to support personalized learning in intelligent tutoring systems, and providing professional development for crowdsourcing good content.

Crowdsourcing and Optimizing Content

While it is important to crowdsource and gather more content, it is crucial to improve the quality of the content delivered to students. Creating authoring tools that can not only crowdsource content but also optimize the content would be necessary to ensure the quality of the content delivered to a large group of students who use the intelligent tutoring system. We have used randomized controlled trials to compare learning supports and identified the ones that would be more beneficial for students. We envision more randomized controlled trials with crowdsourced content to provide insights for intelligent tutoring systems. Another potential way to optimize crowdsourced content is to build functionality in the authoring tools to allow users to view other users' contributed content and rate on their content. Users' credit accumulates over time as they receive good ratings. Similar to the rating system in Stack Overflow in which users get credit for posting good questions and providing good answers, a rating functionality in the authoring tools for crowdsourcing content for intelligent tutoring systems would provide insights on which users have constantly provided good content.

Leveraging AI to Crowdsource Live while Tutoring is Happening

Many of our current crowdsourcing projects focus on randomized controlled trials to determine which student supports work in which context and for which students. To make crowdsourcing more efficient, we envision creating authoring tools that can crowdsource live while tutoring is happening. Take the following case as an example. A student asks a question in a discussion forum and a tutor provides an explanation. The student is able to use the explanation and solves that problem and the next one successfully. In the background, the AI system logs the question and answer pair, the context of the question, and details about the student's performance, in order to decide how and when it can use the explanation to help other students. The next day, when another student asks a similar question, the system guesses based on NLP and delivers the answer from yesterday to help the student solve the problem. Over time the system accumulates more questions from students and more answers from tutors and becomes better and better at responding to students automatically. This example is an embodiment of our vision for the future of intelligent tutoring systems through crowdsourcing live while tutoring is happening.

Towards Personalization via Crowdsourcing

Personalized learning requires investigating what crowdsourced content is more effective for which groups of students. Randomized controlled trials is a scientific method to examine which content is better, but the requirements for using this method such as independent of samples constrained the analysis for interventions in natural learning environments that do not allow randomization. Also, the lag caused by the data collection and analysis makes crowdsourcing mostly beneficial for future students. We have begun to use reinforcement learning, a machine learning method, to examine what crowdsourced content in ASSISTments is most effective for which groups of students so as to personalize tutoring for students. This machine learning method is not subject to assumptions that have to be met for statistical analysis for randomized controlled trials such as the independence of samples. Crowdsourcing is still in its infant stage moving towards personalization in intelligent tutoring systems. With the large amount and variety of crowdsourced content, new machine learning algorithms will need to be developed to quickly identify conditions for personalization. We envision authoring tools that leverage machine learning and AI to efficiently identify the optimal learning condition for different groups of students and provide personalized

learning during the crowdsourcing process. Although it is challenging, we believe it is possible with more advancements in machine learning and AI and more research endeavors.

Professional Development for Crowdsourcing Good Content

While the gist of crowdsourcing is to aggregate knowledge and wisdom from the community, it is essential to provide professional development opportunities to help teachers produce good content to the best of their ability. First it is important to build a network of teachers to make them feel connected and supported. A safe and inclusive community creates a comfortable environment for teachers to work together. Second, helping teachers understand what is good content and how to produce good content is crucial. Strategies such as providing guidelines, examples, practice opportunities, and reflective activities can be used to facilitate the learning process. While upholding the general guidelines, we should also embrace different mindsets on what is good content, especially that different content might work best for different students. For instance, hints could be concrete with specific steps to solve a problem or abstract with clues to help students figure out the process. Another important strategy we learned from our experience is to provide timely feedback to teachers as they are getting started producing content, which helps keep teachers on the right track. Last but not least, authoring tools should be made easy and straightforward to use. Tutorials and technical support should be provided to help teachers use the authoring tools.

Recommendation for GIFT

The Generalized Intelligent Framework for Tutoring (GIFT) is a framework for intelligent tutoring. Most of the intelligent tutoring systems that have been built have used the approach that first the researcher does a lot of research to guess what the mistake will be that users will do, and what good tutors would say in response. At step two, the tutor is released and all users get the same version of everything until a new version of the software is deployed. The release of the first version is not so much designed to collect data but to deploy a running system. GIFT might want to look into ways of extending the framework to deal with allowing new content to come in over time. An even more dramatic goal would be for GIFT to be able to teach the new content creators if their content is good. We recognize these goals as very "aspirational"; we do not see any existing intelligent framework doing this well. Stack Overflow is a system that gets better as more users use it and thumb up things that worked for them. No existing intelligent tutoring systems operate with the ability of Stack Overflow. We ask ourselves, "What future enhancement to the GIFT framework would allow for more dynamic creation and refinements of tutoring systems?" GIFT should allow for crowdsourcing to collect, vet and run different components of the intelligent tutoring system. So rather than building once and running, we envision that the intelligent tutoring system will get better over time. In theory, if our Bandits could get better, we could learn what features matter and then provide suggestions to teachers as they write hint messages.

Conclusion

Crowdsourcing has risen to show its great potential for creating intelligent tutoring systems that personalize learning. We have taken efforts to develop free and accessible authoring tools for crowdsourcing in a particular intelligent tutoring system called ASSISTments, but there is still a great need for future work. Endeavors on providing authoring tools to capitalize on the knowledge and power of the crowd, harnessing rigorous research methodologies to optimize crowdsourced content and advance the learning sciences, leveraging powerful machine learning and AI techniques to personalize learning, and providing professional development for crowdsourcing good content hold the promise for the future of intelligent tutoring systems to better support learners and teachers.

Acknowledgements

Contribution Note: Cheng, Prihar, Baral and Gurung wrote the manuscript. Prihar, Patitcorn, Haim and Sales did the TeacherASSIST work. Baral, Botelho and Gurung did the QUICK-Comments work. Cristina L. Heffernan and Dr. Heffernan oversaw all of it. Other than the first four authors and last author, everyone is listed alphabetically. We thank the ASSISTments content team for providing professional development and all the participating teachers.

References

Alenezi, H. S., & Faisal, M. H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. *Education and Information Technologies*, *25*(4), 2971-2986. https://doi.org/10.1007/s10639-020-10102-w

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-Based Writing Instruction. *Grantee Submission*. https://eric.ed.gov/?id=ED586512

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ED615565

Botelho, A. F., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (accepted, 2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*.

Botelho, A. F., & Heffernan, N., (2019). Crowdsourcing feedback to support teachers and students. In Sinatra, A.M., Graesser, A.C., Hu, X., Brawner, K., and Rus, V. (Eds.). (2019). *Design recommendations for intelligent tutoring systems: Volume 7 - Self-improving systems*(pp. 101-108). Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9977257-7-3. Retrieved at https://gifttutoring.org/attachments/download/3410/DesignRecommendationsforITS_Volume7_SelfImprovingSystemsBook.pdf on January 23, 2023

Dermeval, D., Paiva, R., Bittencourt, I. I., Vassileva, J., & Borges, D. (2018). Authoring tools for designing intelligent tutoring systems: A systematic review of the literature. *International Journal of Artificial Intelligence in Education*, *28*(3), 336–384. https://doi.org/10.1007/s40593-017-0157-9

Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020, March). The automated grading of student open responses in mathematics. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 615-624).

Floryan, M., & Woolf, B. P. (2013). Authoring Expert Knowledge Bases for Intelligent Tutors through Crowdsourcing. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 640–643). Springer. https://doi.org/10.1007/978-3-642-39112-5_78

Heffernan, N. T., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., & Williams, J. J. (2016). The future of adaptive learning: does the crowd hold the key?. *International Journal of Artificial Intelligence in Education*, *26*(2), 615-644. https://doi.org/10.1007/s40593-016-0094-z

Howe, J. (2006). The rise of crowdsourcing. *Wired, 14*(6), 1-4. Retrieved from https://www.wired.com/2006/06/crowds/ on February 28, 2022.

Jiang, Y., Schlagwein, D., & Benatallah, B. (2018, June). *A Review on Crowdsourcing for Education: State of the Art of Literature and Practice*. In Pacific Asia Conference on Information Systems. 180. Retrieved from https://aisel.aisnet.org/pacis2018/180

Kamath, A., Biswas, A., & Balasubramanian, V. (2016, March). A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1-9). IEEE. https://doi.org/10.1109/WACV.2016.7477618

Khosravi, H., Kitto, K., & Williams, J. J. (2019). RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics, 6*(3), 1-10. *ArXiv:1910.05522[Cs]*. http://arxiv.org/abs/1910.0552

Kim, J. (2015). *Learnersourcing: Improving learning with collective learner activity*. MIT PhD Thesis. Retrieved from http://juhokim.com/files/JuhoKim-Thesis.pdf on March 18, 2022.

Krause, M., Garncarz, T., Song, J., Gerber, E. M., Bailey, B. P., & Dow, S. P. (2017, May). Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI*

Conference on Human Factors in Computing Systems (pp. 4627-4639). https://doi.org/10.1145/3025453.3025883

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, *45*(2), 499-515. https://doi.org/10.3758/s13428-012-0258-1

Patikorn, T., & Heffernan, N. T. (2020, August). Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning@Scale* (pp. 115-124). https://doi.org/10.1145/3386527.3405912

Prihar, E., Patikorn, T., Botelho, A., Sales, A., & Heffernan, N. (2021, June). Toward Personalizing Students' Education with Crowdsourced Tutoring. In *Proceedings of the Eighth ACM Conference on Learning@Scale* (pp. 37-45). https://doi.org/10.1145/3430895.3460130

Roscoe, R. D., Allen, L. K., & McNamara, D. S. (2019). Contrasting writing practice formats in a writing strategy tutoring system. *Journal of Educational Computing Research*, *57*(3), 723-754. https://doi.org/10.1177/0735633118763429

U.S. Department of Education. (2015). *Every Student Succeeds Act (ESSA)*. Retrieved from https://www.ed.gov/essa?src=rn on March 18, 2022.

Weir, S., Kim, J., Gajos, K. Z., & Miller, R. C. (2015, February). Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 405-416). https://doi.org/10.1145/2675133.2675219

Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016, April). Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 379-388). https://doi.org/10.1145/2876034.2876042

BIOGRAPHIES

Editors

Dr. Anne M. Sinatra is a Research Psychologist at the US Army DEVCOM Soldier Center, Simulation and Training Technology Center (STTC) in Orlando, FL. Her research focuses on applying cognitive psychology and human factors principles to computer-based education and adaptive training to enhance learning. She is a member of the research team for the award winning Generalized Intelligent Framework for Tutoring (GIFT) software. She is currently the lead editor of the Design Recommendations for Intelligent Tutoring Systems book series. Dr. Sinatra holds a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

Dr. Arthur C. Graesser is an Emeritus professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis, as well as an Honorary Research Fellow at University of Oxford. His research interests question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, problem solving, memory, emotions, artificial intelligence, computational linguistics, and human-computer interaction. He served as editor of the journal ***Discourse Processes*** and ***Journal of Educational Psychology***, as well as presidents of four societies, including Society for Text and Discourse, the International Society for Artificial Intelligence in Education, and the Federation of Associations in the Behavioral and Brain Sciences. He and his colleagues have developed and tested software in learning, language, and discourse technologies, including those that hold a conversation in natural language and interact with multimedia (such as AutoTutor) and those that analyze text on multiple levels of language and discourse (Coh-Metrix and Question Understanding Aid -- QUAID). He has served on four panels with the National Academy of Sciences and four OECD expert panels on problem solving, namely PIAAC 2011 Problem Solving in Technology Rich Environments, PISA 2012 Complex Problem Solving, PISA 2015 Collaborative Problem Solving (chair), and PIAAC Complex Problem Solving 2021.

Dr. Xiangen Hu is a professor in the Department of Psychology, Department of Electrical and Computer Engineering and Computer Science Department at The University of Memphis (UofM) and senior researcher at the Institute for Intelligent Systems (IIS) at the UofM and is professor and Dean of the School of Psychology at Central China Normal University (CCNU). Dr. Hu received his MS in applied mathematics from Huazhong University of Science and Technology, MA in social sciences and Ph.D. in Cognitive Sciences from the University of California, Irvine. Dr. Hu is the Director of Advanced Distributed Learning (ADL) Partnership Laboratory at the UofM, and is a senior researcher in the Chinese Ministry of Education's Key Laboratory of Adolescent Cyberpsychology and Behavior. Dr. Hu's primary research areas include Mathematical Psychology, Research Design and Statistics, and Cognitive Psychology. More specific research interests include General Processing Tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and advanced distributed learning.

Lisa N. Townsend is a Psychologist who just began working at STTC (SED/DEVCOM SC) in Orlando, FL. Prior to STTC, she spent 27 years as a Research Psychologist at the Naval Air Warfare Center Training Systems Division (NAWCTSD) in Orlando, FL. She has a Master of Science in Industrial/Organizational

Psychology and a Bachelor of Arts in Psychology, both from the University of Central Florida (UCF). She has worked on many diverse teams including those within Research and Development, Technology Transfer, Instructional Systems Design, and Human Systems Integration. Ms. Townsend's areas of expertise involve team training related research, Front End Analyses (FEAs), Training Systems Analyses (TSAs), Instructional Systems Design (ISD), Training Effectiveness Evaluations (TEEs), and the development of training and organization related metrics. Her efforts in these areas have spanned across Services and platforms.

Dr. Vasile Rus is the Jack and Jane Morris Professor of Computer Science and Intelligent Systems at The University of Memphis. He also serves as the Director of the Data Science Center and Program. Dr. Rus' research interests are at the intersection of human, animal, and machine learning; specifically, he is exploring how to use Artificial Intelligence and the data revolution to further our understanding of how people learn, how to improve adaptive instructional systems (AISs), and how to make emerging learning ecologies that include online and blended learning with AISs more effective, efficient, engaging, equitable, relevant, and affordable. Dr. Rus' research has been extensively funded by many federal funding agencies. Currently, he serves as PI on 4 projects funded by NSF and Department of Education and as co-PI on 2 projects, one funded by NSF and one by Department of Defense, for a total amount of funding of more than $11 million. Other accomplishments include 150+ peer-reviewed publications (conference papers, journal articles, book chapters), 3 best paper awards (5 best paper award nominations, all with his student advisees), winner of several research competitions (e.g., on automated Question Answering), and supervising and graduating 10 PhD students and 37 Masters students.

Authors

Dr. LisaRe Babin earned a Ph.D. in Learning and Comparative Experimental Psychology at the University of Montana in 1996 and is currently a research psychologist in Army University's Institutional Research and Assessment Division. Lisa taught psychology classes at Montana State University, SUNY College at Buffalo, as well as the University of Maryland's European Division as an associate professor where she taught military personnel stationed in Iraq, Afghanistan, Bahrain, Kuwait, Germany, Italy, Portugal, Greece, and England. Dr. Babin deployed to Afghanistan from 2010-2011 as a D.A. civilian working with 101st Airborne Division (AASLT) where she engaged with Afghan women to develop local economic opportunities and report to the RC-East Division Commander the perspectives of the local women related to operationally relevant issues, and she recently served as the Gender Advisor to TF Eagle for Operations Allies Welcome. In her current position, Lisa is responsible for managing research programs that address leadership needs for Army training and education. She currently is the lead researcher for ArmyU on the Career Courses' Cognitive Assessment Battery (C3AB). Additionally, she is researching how tacit knowledge can be more readily transferred from experts to novices. Her article entitled "Tacit Knowledge Cultivation as an Essential Component of Developing Experts" was published in the Journal of Military Learning in April 2019. Lastly, she has been working on the development and staffing of the Army Learning Concept 2030-2040 for the past three years.

Sami Baral is a Ph.D. student in the computer science program at Worcester Polytechnic Institute. She works in the ASSISTments lab at WPI, where she works on various projects for building better support for teachers for assessment of students' work. Her work revolves around organizing data of student responses for open-ended questions and building machine learning models utilizing natural language processing techniques to automatically assess and suggest feedback messages to teachers to give to their students.

Dr. Elizabeth "Beth" Biddle is a Boeing Senior Technical Fellow in human performance engineering and training with 18 years' service with The Boeing Company. She currently provides technical leadership in the development of advanced learning, human performance modeling and human engineering processes

and capabilities across the enterprise. Beth's prior Boeing roles include Live Training Lead (Future Combat Systems/Brigade Combat Team Modernization), Capture Team Leader, Live-Virtual-Constructive Technology Research & Development Manager and Principal Investigator. Prior to joining Boeing, Beth had over five years' experience in leading human performance and training research and development activities for academic, government and small business organizations. She has a Ph.D. in Industrial Engineering and Management Systems, specializing in Interactive Training Simulations, from the University of Central Florida; a M.S. in Counseling and Human Development from Troy State University and a B.A. in Psychology from Florida State University. She was awarded the Modeling & Simulation Award in Training by the Defense Modeling & Simulation Office (DMSO) in 2001 and nominated as a Charter Member by the National Training and Simulation Association (NTSA) to receive the Certified Modeling & Simulation Professional (CMSP) certification in 2002. Beth was the Conference Chair for the 2018 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC). She is currently a member of Women in Defense (WID) and served on the Central Florida Chapter's Board of Directors as President 2020-2021.

Dr. Anthony T. Botelho is an Assistant professor of Educational Technology in the College of Education at the University of Florida. He seeks to impact learning by studying aspects of student cognition, behavior, and affect through the application of quantitative methods grounded in learning theory and is passionate about using a human-in-the-loop design approach to build that research into practice.

Dr. Li Cheng is a Research Scientist at Worcester Polytechnic Institute (WPI). She holds a Ph.D. in Curriculum and Instruction with Educational Technology emphasis and a minor in Research and Evaluation Methodology from the University of Florida. She works with Dr. Neil Heffernan, Cristina Heffernan and the team on E-TRIALS, a joint project between WPI and The ASSISTments Foundation that leverages ASSISTments to promote educational research and improve K-12 math learning and teaching.

Dr. Michael C. Dorneich is a Professor in the Industrial and Manufacturing Systems Engineering (IMSE) Department at Iowa State University, and a faculty affiliate of the human computer interaction (HCI) graduate program at Iowa State University. At the University of Illinois at Urbana-Champaign, he earned his Ph.D. in industrial engineering (Human Factors), and MS and BS in Electrical Engineering. His research interests focus on creating joint adaptive human-machine systems that enable people to be effective in the complex and often stressful environments found in aviation, military, robotic, and space applications. Dr. Dorneich has experience conducting research in industry, government labs, and academia. He led programs from DARPA, NASA, FAA, NIH, and UK and EU international projects. His recent work looks at the development of intelligent team tutoring systems, cognitive assistants for space operations, and the development of human-autonomy team frameworks. Prior to joining the faculty at Iowa State University, he worked at Honeywell Laboratories researching adaptive system design and human factors in a variety of domains. He is the author of more than 200 articles and 27 patents. He is an Associate Editor of this journal IEEE Transactions of Human-Machine Systems, and an Editorial Board Member of the Journal of Cognitive Engineering and Decision Making.

Lauren Egerton is a Data Engineer with a background in foreign language and education. She utilizes a diverse skill set to approach natural language tasks with a unique perspective and a deep understanding of language.

Dr. Stephen B. Gilbert is currently Associate Director of Iowa State University's Virtual Reality Application Center and Director of its Human Computer Interaction graduate program. He is also an associate professor in the Industrial and Manufacturing Systems Engineering department. His research interests focus on technology to advance cognition, including intelligent tutoring systems, human-autonomy teaming, and XR usability. He works closely with industry, NSF, and DoD on research contracts

and has also worked in commercial software development and runs his own company. He received a BSE from Princeton in civil engineering and operations research and a Ph.D. from MIT in brain and cognitive sciences.

Kari Glover is a Full Stack Developer for Eduworks Corporation, co-facilitator of the Open Skills Network's Technical Workgroup, and former educator. She has a passion for building technology that connects people through educational needs, talents, learning, and career opportunities

Jim Goodell is an expert on learning technologies and data standards, and is Vice Chair of the IEEE Learning Technology Standards Committee. As Senior Analyst with Quality Information Partners (QIP) he leads standards development for the U.S. Department of Education sponsored Common Education Data Standards (ceds.ed.gov) and works with stakeholders from early learning, K12, postsecondary, and workforce organizations. He chairs the IEEE Adaptive instructional Systems (AIS) Standards Interoperability Subgroup. He serves on the IEEE IC Industry Consortium on Learning Engineering (ICICLE) Steering Committee, co-chaired the first ICICLE conference, and leads the ICICLE Competencies, Curriculum, and Credentials SIG. In 2016, he co-authored *Student-Centered Learning: Functional Requirements for Integrated Systems to Optimize Learning.*

Asish Gurung is a Ph.D. Student in the Computer Science program at Worcester Polytechnic Institute. At WPI, he works in the ASSISTments lab in developing various classroom orchestration tools that are primarily teacher-facing. As part of the research, Ashish leads different research teams that focus on student behavior on online learning platforms that explore the various approaches researchers can take to synthesize student interaction data into information the teachers can leverage. The primary objective of his work is to facilitate effective student-teacher interaction and augment teachers ability to enhance learning experiences.

Aaron Haim is a Ph.D. student in the data science program at Worcester Polytechnic Institute. At WPI he works in the ASSISTments lab, which develops new features for the ASSISTments online learning platform. In the lab he organizes various projects revolving around creating tools and performing analysis to improve the effectiveness of on-demand assistance crowdsourced from educators on problems for students.

Cristina Heffernan is the Executive Director of the ASSISTments Foundation and the co-founder of ASSISTments. In her prior role, she supported the development and advancement of ASSISTments as well as further engaged and trained the community of teachers who utilize ASSISTments as an educational tool in their classrooms. Over the last 15 years, Ms. Heffernan has supported the integration of technology-enhanced teaching with ASSISTments in a multitude of situations. Before developing technology for the classroom she was a middle school math teacher and an instructional coach.

Dr. Neil T. Heffernan is a Professor of Computer Science and Director of the Learning Sciences and Technologies program at Worcester Polytechnic Institute. While completing his Ph.D. in Computer Science at Carnegie Mellon University, Neil incorporated his passion for education and focused on educational technologies. In 2003, Neil and his wife Cristina created the ASSISTments platform as a forever-free service that is currently used by over 20,000 teachers and 500,000 students across the United States for daily classwork and nightly homework. In 2021, ASSISTments was named by WWC as one of three online middle-school math interventions proven to impact student achievement, and has a Tier 1 rating from Evidence for ESSA.

Elaine Kelsey is the Director of Research at Eduworks Corporation, focusing on applications of natural language processing and machine learning in semantic search, text generation, and automated tutoring, with over five years of experience with pretraining and fine-tuning LLMs for these applications. She designed and led the development of Eduworks' automated assessment generation and evaluation technologies. She

has multiple Bachelors and Masters degrees in computer science, linguistics, and molecular biology, and is currently working towards a Ph.D. in Computational Linguistics.

Dr. Susanne P. Lajoie is a Canada Research Chair in Advanced Technologies for Learning in Authentic Settings in the Department of Educational and Counselling Psychology and is an associate member of the Institute for Health Sciences Education at McGill University. She is a Fellow of the Royal Society of Canada, the American Psychological Association and the American Educational Research Association. Dr. Lajoie explores how theories of learning and affect can be used to guide the design of advanced technology rich learning environments to promote learning in medicine.

Dr. Shan Li is an Assistant Professor in the College of Health at Lehigh University. He is also an affiliated faculty in the Department of Education and Human Services at Lehigh University. His overarching research goal is to understand and enhance health professions education (HPE) by designing intelligent learning and training applications, and examining students' learning processes with educational data mining and learning analytics techniques.

Dr. James C. Lester is Distinguished University Professor of Computer Science and director of the Center for Educational Informatics at North Carolina State University. He is Director of the National Science Foundation AI Institute for Engaged Learning. His research centers on transforming education with artificial intelligence. His current work ranges from AI-driven narrative-centered learning environments and virtual agents for learning to multimodal learning analytics, sketch-based learning environments, and computer-supported collaborative learning. He has served as Editor-in-Chief of the International Journal of Artificial Intelligence in Education. He is the recipient of an NSF CAREER Award and the Best Paper Awards at the International Conference on Artificial Intelligence in Education, the ACM International Conference on Intelligent User Interfaces, the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, and the International Conference on User Modeling, Adaptation, and Personalization. His foundational work on pedagogical agents has been recognized with the IFAAMAS Influential Paper Award by the International Federation for Autonomous Agents and Multiagent Systems. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).

Dr. Laura Milham is an Air Force civilian and an applied human-systems scientist, focusing on individual and team training and assessment. As the ADL Initiative Acting Director, Dr. Milham brings over 25 years of experience to oversee the program's execution, guide its strategy, and ensure that it aligns with US DoD and Federal government priorities. Seeing the impact of training at the point of need has served as a guiding force in her career and reinforced her current mission. Her clear objective is to ensure that the ADL Initiative fully leverages the opportunity to optimize ubiquitously distributed training opportunities, from classroom to simulation, from glass houses built with tape in the dirt, to formal, large scale simulation events.

Dr. Wookhee Min is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He earned his Ph.D. in Computer Science from North Carolina State University. His research focuses on creating adaptive training and learning environments using artificial intelligence for student modeling, natural language processing, and procedural content generation. He has served as a Co-PI on two NSF-supported projects, served as Posters and Demos Co-Chair at the 11th International Conference on Educational Data Mining, and co-organized a tutorial, "Deep Learning for Interactive Digital Entertainment," at the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Dr. Min's work has been recognized with the Best Student Paper Award at the Thirteenth International Conference on Educational Data Mining and five US patents.

Dr. Sazzad Nasir is a senior AI scientist at Eduworks Corporation with a background in theoretical physics and neuroscience. He is an expert on state-of-the art AI/ML methods who has worked in both academia and industry. Dr. Nasir received his BS with Honors (ranked first) in physics from the University of Dhaka,

Bangladesh. Subsequently, he earned his Master of Advanced Study (Part III of Mathematical Tripos) with Distinction and Ph.D. in theoretical particle physics from the University of Cambridge, UK. His teaching and research career straddled multiple scientific disciplines. Before joining Eduworks, he was a faculty member at Northwestern University, taught and led research in neuroscience. His research in physics and neuroscience appeared in leading journals. Currently, he has a research affiliation with Haskins Laboratories at Yale University. He has expertise in machine learning, predictive analytics, NLP and computational modeling. He has managed and supervised several doctoral research projects.

Dr. Benjamin Nye is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies (ICT). Ben's research tries to remove barriers to development and adoption of adaptive and interactive learning technology so that they can reach larger numbers of learners. Dr. Nye's research has been recognized for excellence in adaptive and intelligent tutoring systems, cognitive agents, and realistic behavior in training simulations. His research is on scalable learning technologies and design principles that promote learning, with the goal of making effective learning tools more broadly available.

Dr. Thanaporn Patikorn is a lecturer in the Computer Science Department at Rajamangala University of Technology Suvarnabhumi, Thailand. His research interests include artificial intelligence and machine learning applications, crowdsourcing, and data-driven approaches to improve learning management systems.

Ethan Prihar is a Ph.D. student in the data science program at Worcester Polytechnic Institute. At WPI he works in the ASSISTments lab, which develops new features for the ASSISTments online learning platform. In the lab he organizes various projects revolving around collecting data from students and building machine learning models that either predict student behavior or provide students with personalized tutoring.

Dr. Albert "Skip" Rizzo is a Clinical and Neuro- Psychologist, and Director of the University of Southern California Institute for Creative Technologies Medical VR Lab. He is also a research professor in both the USC Dept. of Psychiatry and in the School of Gerontology. Skip conducts research on the design, development and evaluation of VR systems targeting the areas of clinical assessment, treatment and rehabilitation.

Dr. Rebecca L. Robinson earned a Ph.D. in Experimental Psychology at the University of Texas at Arlington where she conducted studies in social influence and behavior, attitudes, decision-making, personality/individual differences, in addition to survey/scale development and psychometrics. In 2020, Dr. Robinson joined the Institutional Research and Assessment Division (IRAD) at the Army University as a Research Psychologist. Prior to ArmyU, she was a contractor with the Bureau of Safety and Environmental Enforcement, Department of the Interior. Dr. Robinson's current efforts include assisting in the development of program evaluations and assessments, evaluating online learning tools, and improving self-regulated learning in early military education. She also has a role supporting students at the Command and General Staff Officer Course (CGSOC) pursuing a Master of Military Art and Science (MMAS) degree.

Dr. Robby Robson is a researcher, entrepreneur, and standards professional known for creative and disruptive innovation in industry and academia. He is the Chief Science Officer and co-founder of Eduworks Corporation, is on several IEEE governance boards, and is currently developing the next generation of competency management, talent analytics, and skills-based talent pipeline solutions.

Dr. Jonathan Rowe is a Senior Research Scientist in the Center for Educational Informatics and an Adjunct Assistant Professor in the Department of Computer Science at North Carolina State University. He is also Managing Director of the National Science Foundation AI Institute for Engaged Learning (EngageAI Institute). His research focuses on artificial intelligence in adaptive learning technologies, with an emphasis

on game-based learning, interactive narrative generation, intelligent tutoring systems, multimodal learning analytics, affective computing, and user modeling. He received Ph.D. and M.S. degrees in Computer Science from North Carolina State University and his B.S. degree in Computer Science from Lafayette College.

Dr. Adam Sales is an assistant professor in the Mathematical Sciences department and Learning Sciences and Technologies Ph.D. program at Worcester Polytechnic Institute. His research interests focus is in methods for causal inference using educational data, especially data from education technology.

Brent Smith is a Software Systems Architect with over 20 years of experience in designing and developing learning technologies for government stakeholders, defining R&D roadmaps to meet organizational objectives, and establishing chains of research that align with strategic goals. As the ADL Initiative R&D Principal, Mr. Smith helps ensure the ADL Initiative research agenda is aligned with its overall strategy.

Dr. Robert A. Sottilare is the Science Director for Intelligent Training at Soar Technology, Inc. He came to SoarTech in 2018 after completing a 35-year federal career in both Army and Navy training science and technology organizations. At the US Army Research Laboratory, he led the adaptive training science and technology program where the focus of his research was automated authoring, instructional management, and analysis tools and methods for intelligent tutoring systems (ITSs) and standards for adaptive instructional systems. He is the father of the Generalized Intelligent Framework for Tutoring (GIFT), an award-winning open source, AI-based adaptive instructional architecture. GIFT has over 2000 users in 76 countries. Dr. Sottilare has a long history as a leader, speaker, and supporter of learning and training sciences forums at the Defense & Homeland Security Simulation, HCII Augmented Cognition, and AI in Education conferences. He is the founding chair of the HCII Adaptive Instructional Systems (AIS) Conference. He is a member of the AI in Education Society, the Florida AI Research Society, the IEEE Computer Society and Standards Association (senior member), the National Defense Industry Association (lifetime member), and the National Training Systems Association. He is currently the IEEE Project 2247 working group chair for the development of standards and recommended practices for AISs. He is a faculty scholar and has been an adjunct professor at the University of Central Florida where he taught a graduate level course in ITS theory and design.

Dr. Andrew Smith is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He earned his Ph.D. in Computer Science from North Carolina State University. His research focuses on utilizing artificial intelligence and machine learning for applications such as adaptive training and learning environments with an emphasis on user modeling, game-based learning, and educational data mining. Dr. Smith brings 15 years of experience in AI research and software development, including 5 years of software development experience in industry.

Dr. Randall Spain is a Research Scientist at the U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center, Simulation and Training Technology Center (STTC). He received his Ph.D. and MS degrees in Human Factors and Experimental Psychology from Old Dominion University. His research focuses on designing, developing, and evaluating adaptive training technologies with a particular emphasis on investigating data-driven models of coaching and feedback to support team training in synthetic training environments, using natural language processing methods to support team communication analytics, and investigating UI/UX principles for intelligent user interfaces. Prior to joining the DEVCOM-Soldier Center, Dr. Spain was a Research Scientist in the Center for Educational Informatics at North Carolina State University where he led research sponsored by the National Institute of Standards and Technology (NIST), the U.S Air Force, and the U.S. Army Research Laboratory evaluating AI-driven training and learning technologies.

Dr. William Swartout is Chief Technology Officer at the USC Institute for Creative Technologies, providing overall direction to the Institute's research programs. He is also a research professor in the Computer Science Department at the USC Viterbi School of Engineering. His research interests include intelligent computer based education, virtual humans, and explainable and trusted AI. Swartout is a Fellow of the AAAI, has served on their Board of Councilors and is past chair of the Special Interest Group on Artificial Intelligence (SIGART) of the Association for Computing Machinery (ACM). In 2009, Swartout received the Robert Engelmore Award from the Association for the Advancement of Artificial Intelligence (AAAI) for seminal contributions to knowledge-based systems and explanation, groundbreaking research on virtual human technologies and their applications, and outstanding service to the artificial intelligence community. He has served as a member of the Air Force Scientific Advisory Board, the Board on Army Science and Technology of the National Academies and the JFCOM Transformation Advisory Group. Prior to helping found the ICT in 1999, Swartout was the Director of the Intelligent Systems Division at the USC Information Sciences Institute. He received his Ph.D. and M.S. in computer science from MIT and his bachelor's degree from Stanford University.

Dr. Eliot Winer is Director of the VRAC (Visualize. Reason, Analyze. Collaborate.) research center and Professor of Mechanical Engineering at Iowa State University. He also has courtesy appointments in the departments of Electrical and Computer Engineering and Aerospace Engineering at ISU. Dr. Winer has over 23 years of experience working in extended reality (XR), 3D computer graphics, machine learning and approximations, and design methods for a variety of uses. He has developed virtual environments for applications from engineered products, manufacturing, surgical procedures, and distance education in rural communities. Dr. Winer received his B.S. in Aerospace Engineering in 1992 from Ohio State University. His M.S. in Mechanical Engineering from the State University of New York at Buffalo in 1994 and earned his Ph.D. from the State University of New York at Buffalo in Mechanical Engineering in 1999. He has worked in industry for many years and started up three companies.

Peggy Wu is an award winning scientist with 20+ years of experience combining cognitive psychology with Artificial Intelligence to advance Human Computer Interactions, Social Computing, Human Machine Trust, & Ethical AI. Applying advanced technologies such as Virtual/Augmented Reality to improve Human+Machine performance on Earth and in Space. Peggy is the recipient of numerous grants from the DoD, NASA, and DoE, and serves in a leadership capacity for a matrix organization of 200+ team members. She serves as a judge XPRIZE, has been an invited expert for congressional staffer and NATO briefings, NASA proposal reviewer, and conducts public engagements and probono consulting to enhance Science and Technology representation through the National Academies of Sciences and the UN. She is an inventor with 20 patents (14 pending) and over 70 peer reviewed publications.

Design Recommendations for Intelligent Tutoring Systems

Volume 11
Professional Career Education

Design Recommendations for Intelligent Tutoring Systems (ITSs) explores the impact of intelligent tutoring system design on education and training. Specifically, this volume focuses on ITSs for professional career education. The types of training that occurs in professional education tend to be more applied, and experiential than in traditional ITSs. The current book highlights general approaches and techniques that can be beneficial to developing ITSs for professional career education. Additionally, the book includes chapters with specific use cases where ITSs have been implemented in diverse domains including medical, teaching, commercial pilot, and manufacturing training.

About the Editors:

- ***Dr. Anne M. Sinatra*** *is a research psychologist at U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center.*
- ***Dr. Arthur C. Graesser*** *is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is a Senior Research Fellow in the Department of Education at the University of Oxford.*
- ***Dr. Xiangen Hu*** *is a professor in the Department of Psychology at the University of Memphis and a professor at Central China Normal University.*
- ***Ms. Lisa N. Townsend*** *is a research psychologist at U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center.*
- ***Dr. Vasile Rus*** *is a professor in the Department of Computer Science at the University of Memphis with a joint appointment in the Institute for Intelligent Systems.*

A Volume in the Adaptive Tutoring Series