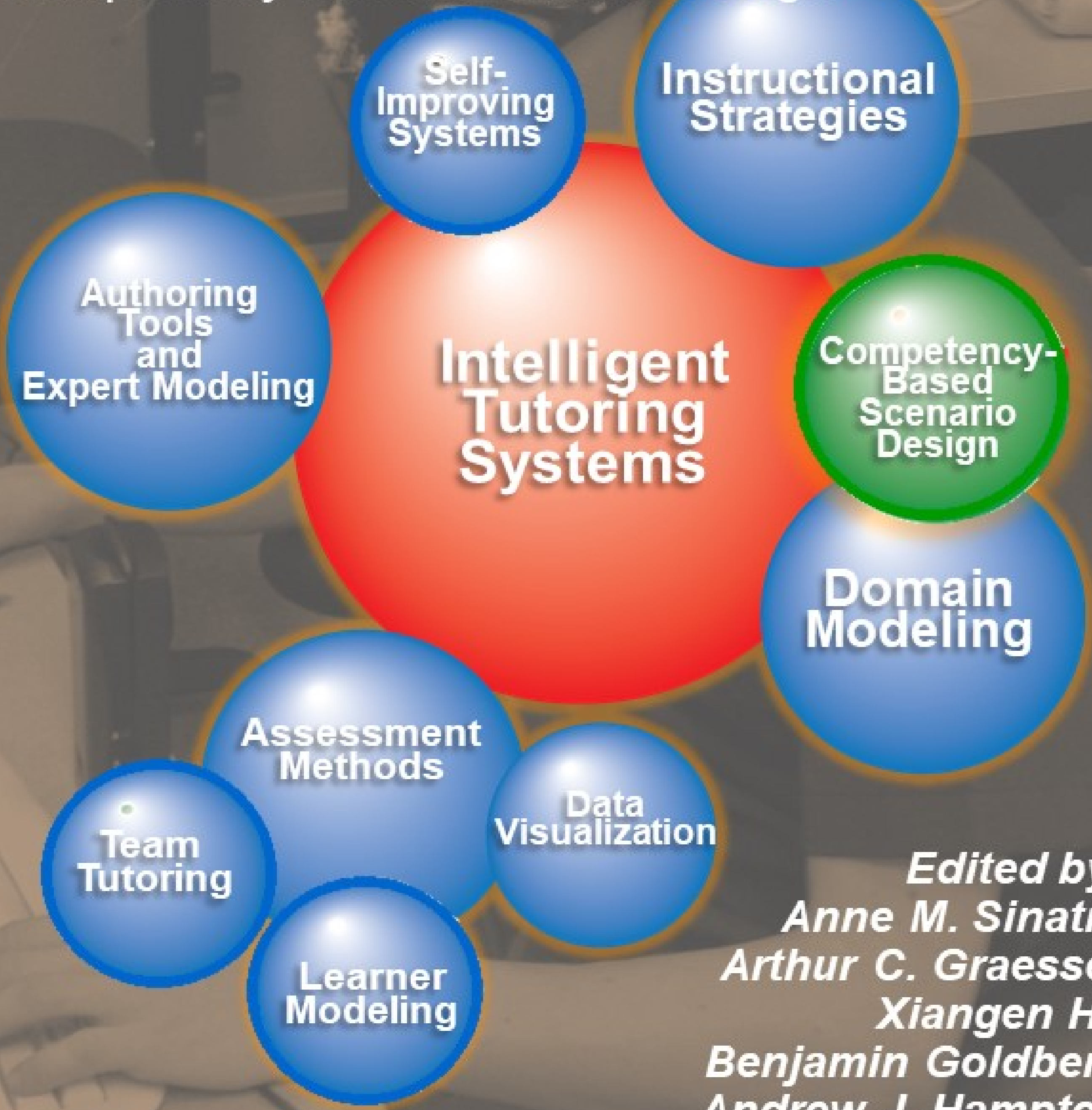
Design Recommendations for Intelligent Tutoring Systems

Volume 9
Competency-Based Scenario Design



Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Benjamin Goldberg
Andrew J. Hampton
Joan H. Johnston

A Book in the Adaptive Tutoring Series

Design Recommendations for Intelligent Tutoring Systems

Volume 9
Competency-Based Scenario Design

Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Benjamin Goldberg
Andrew J. Hampton
Joan H. Johnston

A Book in the Adaptive Tutoring Series

The views expressed herein are those of the authors and do not necessarily reflect the views of the US Army Combat Capabilities Development Command - Soldier Center.

Use of trade names or names of commercial sources is for information only and does not imply endorsement by the US Army Combat Capabilities Development Command - Soldier Center.

This publication is intended to provide accurate information regarding the subject matter addressed herein. The information in this publication is subject to change at any time without notice. The US Army Combat Capabilities Development Command - Soldier Center, nor the authors of the publication, makes any guarantees or warranties concerning the information contained herein.

Printed in the United States of America
First Printing, February 2022

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

International Standard Book Number:
978-0-9977258-1-0

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 COMPETENCY-BASED SCENARIO DESIGN & GIFT

Anne M. Sinatra[1], Arthur C. Graesser[2], Xiangen Hu[2], Benjamin Goldberg[1], Andrew J. Hampton[2], and Joan H. Johnston[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

This book focuses on the topic of competency-based scenario design as it relates to Intelligent Tutoring Systems (ITSs). The current book is the ninth in a series of books that examine key topics in ITSs. The chapters in this book specifically relate the work presented to applications for the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare, Brawner, Goldberg, & Holden, 2012; Sottilare, Brawner, Sinatra, & Johnston, 2017). GIFT is an open-source, domain independent, service-oriented, modular architecture for ITSs. GIFT has specifically been designed to allow for reusability of the GIFT architecture, GIFT tools, and instructional content materials. Further, GIFT has been designed with the goals of reducing the amount of time necessary to author ITSs, and reducing the skill level required for the authoring process. GIFT can be used to create ITSs that can be distributed both locally on a computer and virtually in the Cloud. In addition to creating ITSs, GIFT can be used to examine instructional outcomes, and conduct research.

In addition to this book, the first eight volumes in this series, Learner Modeling (ISBN 978-0-9893923-0-3), Instructional Management (ISBN 978-0-9893923-2-7), Authoring Tools (ISBN 978-0-9893923-6-5), Domain Modeling (978-0-9893923-9-6), Assessment Methods (ISBN 978-0-9977257-2-8), Team Tutoring (ISBN 978-0-9977257-4-2), Self-Improving Systems (978-0-9977257-7-3) and Data Visualization (ISBN 978-0-9977257-8-0) are freely available at www.GIFTtutoring.org.

The topic of this book, Competency-Based Scenario Design is highly relevant to the development of ITSs. Scenarios are information-rich task/problem contexts that are closely aligned with real-world situations that professionals face in their jobs. The tasks/problems exhibit ecological validity rather than stripped-down abstract simplifications. Developers of ITSs and other adaptive instructional systems need to have principled guidance on how to design these scenarios. An example scenario may be a close match to a particular situation in the past, but not be representative of a large range of situations that professionals experience in their job. An example scenario may be very realistic, but not provide reliable and valid assessments of the learners' performance to guide assessments (summative, formative, or stealth). Research teams that build high quality scenarios need to include expertise in the targeted profession, assessment, learning science, and computer science. The current book brings together experts on ITSs to discuss their work as it applies to Competency-Based Scenario Design. We believe that this book can be used as a resource for those who have an interest in developing Scenarios for ITSs, and who want to learn more about how to do so.

GIFT and Expert Workshops

This book series is associated with a series of Expert Workshops that began in 2012. In 2012, the Army Research Laboratory (ARL) along with the University of Memphis developed a series of expert workshops including senior tutoring system scientists from academia, government, and industry to present their work on relevant gaps in ITS research and applications. As part of these workshops the experts also provide suggestions for ways to improve GIFT moving forward. Expert workshops have been held each year since 2012 resulting in published volumes in the ***Design Recommendations for Intelligent Tutoring Systems*** series that are associated with each workshop. In 2018, parts of ARL, including the GIFT team, were reorganized into another organization, Soldier Center. Research into applied adaptive tutoring and team tutoring have continued with Soldier Center. Additionally, the expert workshops and books have continued with topics in line with the relevant research gaps. Table 1 lists the expert workshop topics, the locations of the workshops, as well as the dates of the workshops and associated volume publications.

Table 1. Historical List of Expert Workshops, Locations, Dates, and the Book Publication Output.

Expert Workshop Topic	Expert Workshop Location	Expert Workshop Date	Book Publication
Learner Modeling	Memphis, TN	September 2012	Volume 1 – July 2013
Instructional Management	Memphis, TN	July 2013	Volume 2 – June 2014
Authoring Tools	Pittsburgh, PA	June 2014	Volume 3 – June 2015
Domain Modeling	Orlando, FL	June 2015	Volume 4 – July 2016
Assessment Methods	Princeton, NJ	May 2016	Volume 5 – June 2017
Team Tutoring	Ames, IA	May 2017	Volume 6 – August 2018
Self-Improving Systems	Nashville, TN	May 2018	Volume 7 – October 2019
Data Visualization	Orlando, FL	August 2019	Volume 8 – December 2020
Competency-Based Scenario Design	Virtual	September 2020	Volume 9 – January 2022
Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analysis of Intelligent Tutoring Systems	Virtual	September 2021	Volume 10 – In Progress

Sections of the Book

This book is organized into three sections that cover three related but diverse topics associated with Competency-Based Scenario Design and ITSs:

I. Scenarios for Individual Assessment

II. Scenario Based Training for Groups and Teams

III. Computational and Quantitative Models

Design Recommendations for Intelligent Tutoring Systems: Volume 9 – Competency-Based Scenario Design is intended to be a design resource as well as a community research resource. We believe that Volume 9 can serve as an educational guide for developing ITS scientists and as a roadmap for ITS research opportunities.

References

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.

Section I – Scenarios for Individual Assessment

Dr. Arthur C. Graesser and Dr. Anne M. Sinatra, Eds.

CHAPTER 1 – INTRODUCTION TO SCENARIOS FOR INDIVIDUAL ASSESSMENT

Arthur C. Graesser[1] and Anne M. Sinatra[2]
University of Memphis[1], U.S. Army DEVCOM - Soldier Center[2]

Core Ideas

The chapters in this section focus on assessment of individuals in competency-based scenario design. Individuals are assigned tasks in the assessment that are very similar to those tasks they need to complete in real-world settings. The assessment tasks involve authentic scenarios rather than decontextualized tasks that are many abstractions away from the real tasks they need to perform in a job position or as a citizen. Assessments are designed to assess whether an individual is competent in completing the tasks and/or what level of proficiency they exhibit.

There are many stages in developing learning and assessment tasks in competency-based scenario design. An initial step is to define the theoretical construct (e.g., leadership, cognitive flexibility) or practical scope of the job position (e.g., aircraft electrician, webmaster) being assessed. A second step is to identify the set of knowledge, skills, attitudes, and abilities (KSAAs) that are needed or desired to achieve competent performance in achieving goals. A third step is to identify tasks in the real world that appear to have face validity in being authentic scenarios and that require these KSAAs. A fourth step is to design similar scenarios for assessment and that cover important relevant KSAAs. A fifth step is to measure the test-takers' performance and develop a scale that identifies thresholds of competence and levels of proficiency. Competency-based scenario design requires expertise in psychology, measurement, scenario design, and experience in relevant real-world activities.

Individual Chapters

The chapter by ***Sottilare*** discusses scenario design and development processes for competency-based training at scale in military contexts. Common competency-based assessment methods include pretests, in-situ assessments, and computational methods when there is longitudinal data available in the learning record store. There is a particular focus on adaptive scenario-based training with digital technologies that uses a series of scenarios and events in scenarios to develop or maintain the ability to complete a task. Scenarios and events are adaptive by changing their sequence, level of difficulty, or training content. However, making such decisions on adaptivity involves some challenges that require complex modeling and expertise in instructional design.

The chapter by ***Owens*** and ***Goldberg*** presents a Competency-Based Experiential-Expertise (CBEE) model that was developed during a decade of research in Army and Navy training and performance projects. CBEE is now part of a US Army science and technology research project to develop advanced training management tools in conjunction with the Army's Synthetic Training Environment. It uses Artificial Intelligence (AI) supported evaluation tools and standards for classifying levels of competence that promote learning-on-demand at a point-of-need. Learning and competence evaluation occurs through live, full- or semi-synthetic learning environments that emulate real-world settings and conditions that expose trainee mistakes and learning-by- doing in an experiential learning model.

The chapter by ***Baker, Choi,*** and ***O'Neil*** presents a Training Assessment Framework (TAF) that provides an architecture for designing, developing, and providing valid assessments of training and education. There are innovative tools for scenario-based performance assessment and analysis of relevant features in the assessment. The chapter discusses a project with the Naval Education and Training Command that supported the development and use of a TAF to create a coherent structure for designing end-of-course assessments that also considers formative assessments during training, analytic models of computer-based instruction, and feedback to instructors, trainees, and test developers. Clear performance assessments can be designed for complex practice environments, simulations, and other intelligent tutoring systems (ITSs). A recommendation for the future is to develop a theory and ontology of situations that can help operationally define transfer of training to new situations, including the modification of content, setting, roles, and goals.

The chapter by ***Sabatini*** reviews scenario-based assessment and how it can be designed for ITSs. Design principles are presented based on scenario-based assessments developed and tested in a large-scale Reading for Understanding initiative funded by the Institute of Educational Science. Evidence-centered design is a core technique that guided the process. It is also important to consider constraints in both the research/development phase and the final operational use. The principles guiding scenario-based assessment include a goal (purpose, mission) for the participant to achieve over the course of the assessment, a coherent collection of materials relevant to achieving the goal, use cases in assessment (formative, summative, stealth) that triangulate strengths and weaknesses in performance, links between prior knowledge and performance, and ideally optimization of interest, motivation, and engagement.

The chapter by ***Kyllonen, Graesser, Haviland, Robbins,*** and ***Williams*** explores the process and advantages of implementing soft skills training in GIFT. Soft skills refer to social, emotional, and self-management skills associated with success in school and work. The chapter describes a soft skill training system developed at Educational Testing Service that could potentially be integrated with GIFT in ways analogous to the training of hard skills (e.g., verbal and quantitative literacy, science, particular technical skills). Unlike hard skills, soft skills are rarely explicitly taught and assessed in schools. The training systems include, among other activities, scenarios with open-ended problems in workplace, school, community, and personal settings that present an issue or problem that calls for deliberation on a resolution.

Chapter 2 – Considerations for Adaptive Competency-Based Scenario Design and Development at Scale

Robert A. Sottilare
Soar Technology, Inc.

Introduction

This chapter focuses on the scenario design and development processes for competency-based training. We emphasize the application of these processes for training (learning a specific task) versus education (learning of concepts that may be broadly applied). Competency-based training provides a framework for instruction and the assessment of learning based on a set of predetermined competencies (abilities required to complete or achieve something) where measures of assessment are tied to real-world performance standards and outcomes (Lytras et al., 2010).

In thinking about the scenario design process for *adaptive competency-based training*, the designer/developer should consider the modeling of learners and teams, the use of those models in adapting scenarios in real-time and the evaluation of instructional decisions to reinforce better decision-making in future scenarios. To this end, this chapter explores scenario design and development processes. Next, we define *adaptive scenario-based training* and the contexts in which competency modeling provides opportunities to tailor scenario dimensions.

First, the broad term of *adaptive instruction* has been defined as the use of various instructional strategies and resources (e.g., content, content delivery methods - intelligent multimedia instruction (IMI)) to provide learning experiences that are tailored to the needs, goals, preferences and interests of individual learners or teams (Wang, 1980; Sottilare et al., 2018; Sottilare & Brawner, 2018). Therefore, *adaptive scenario-based training* is adaptive instruction that uses a series of events (scenarios) to develop or maintain the ability to complete a task or set of tasks. The events may be adapted by changing their sequence, their level of difficulty, or by providing alternate content, but how does competency fit into the adaptive instructional decision-making process? In the next section of this chapter, we explore competency modeling and challenges in designing and developing adaptive competency-based training.

Competency Assessment Methods – Challenges and Limitations

Now that we have defined a scope for the design and development of adaptive competency-based training, our next task is to identify the process and challenges associated with competency determination and how it is used to tailor training scenarios. Vygotsky (1978) identified the zone of proximal development (ZPD) as a set of proximal skills (a range of abilities) that an individual can perform with assistance from a tutor or peer, but cannot yet perform completely on their own. An individual's proximal skills are those they are close to mastering. By focusing instruction on the development of proximal skills, the instruction emphasizes objectives that are attainable in the near term and this aids learner engagement. However, the ability to use ZPD as a design guideline is directly dependent on the adaptive instructional system's (AIS's) ability to accurately assess learner or team competency. How is competency modeled today? Competency may be inferred through various methods, but is most often determined using *pretests*, an *in-situ assessment*, or a *computational method* used to assess previous learning and experience.

Pretests as Competency Assessments

The use of well-designed pretests or competency assessments are valid and reliable. They can measure knowledge and skills required to perform a task or successfully work in a specific job. For example, a pretest might be administered to determine if a learner has the requisite skills to enroll in a course or an organization might administer a certification test to ensure that employees operating a specific type of equipment possess the required knowledge and skills. It is important that these tests pose practical problems that reflect the knowledge and skill needed to be successful and efficient.

The limitations associated with using pretests is their validity and the availability of the time required to complete the assessment. Face validity of pretest is the extent to which the test measures the competency that it was designed to measure. The availability of time required for the assessment should be viewed in terms of return on investment (ROI). If there is a return in accelerating learning by eliminating lessons or modules where the learner is already proficient, there is time returned for the pretest time investment. If there are no planned adaptations to recover the time, then the designer may determine that the pretest is unnecessary.

In-Situ or Real-time Competency Assessments

Next, we examine in-situ or real-time competency assessments. It might be useful to determine learner competency to provide more efficient instruction and accelerate learning. Learner competency might be assessed during instruction and real-time decisions made by the instructor to determine what lessons can be skipped or if remediation to more fundamental lessons are required. Machine-based tutoring provides an advantage in being able to track real-time performance and adapt recommendations for future instructional experiences based on that performance. Human tutors with large classes or student populations may not be able to afford the time to make the interim competency assessments required to make real-time recommendations or decisions about learning pathways.

Computational Competency Assessment Methods

Finally, we examine more longitudinal methods of assessing competency. Today, the experience application program interface (xAPI; Sottilare et al., 2017), a statement of achievement, is widely used for capturing military training accomplishments. The xAPI is an e-learning software specification that records all types of learning experiences in a learning record store (LRS; Hruska et al., 2015). While the xAPI as a method of capturing learning achievements has been quite successful, it is somewhat limited in its ability to accurately assess competency.

The xAPI statements are written to the LRS where the statements pile up over time. The statements are time stamped so they could be used with a decay model to compute when the learner needs refresher training. Some logic or computational model is needed to make sense of the xAPI statements as they are often unable to stand alone as competency assessment models. xAPI statements may not be granular enough to determine whether a learner or team of learners is at a sufficient level of competency for the machine-based tutor to make recommendations about their specific learning path. For example, a learner or team completing a scenario may not be sufficient evidence to determine their proficiency because only a single achievement statement was generated.

Now that we have examined competency assessment methods, we will discuss the process, challenges and limitations of scenario design and development in the context of adaptive training.

Scenario Design and Development – Challenges and Limitations

Scenario-based training (SBT) takes place in immersive training environments where learners are exposed to realistic work challenges and conditions (Dunne et al., 2010; Magerko & Laird, 2002). SBT illustrates a time-tested "learn-by-doing approach" for military training. The goal of SBT is to improve learner or team performance and build competency through iteratively more difficult experiences and application of each learner's knowledge and skills to a broad variety of job conditions.

This section provides discussion about the training scenario design process which requires a high degree of knowledge of the domain of instruction and defined learning objectives and measures of assessment (Noori et al., 2017). The myriad of design process steps outlined in the training literature can be distilled down to a few simple heuristics: 1) know your audience, their goals and learning gaps, 2) define your learning objectives, 3) select appropriate activities (e.g., skill, reasoning, decision-making, optimization) to support learning objectives and overcome knowledge gaps, and 4) tie expected scenario outcomes and measures of assessment to a sequence of activities (events).

Military training departments usually provide a set of standard scenarios that are tied to specific objectives so that inexperienced teams can learn to perform in increasingly complex environments and experienced teams can maintain their performance edge. Scenarios are typically authored by training department members, battlemasters or instructors. Each base scenario may be adapted manually by the battlemaster in real-time to keep each team engaged with an appropriate level of difficulty. The battlemaster can adjust the simulation events or activities to reduce or increase complexity if the scenario is determined to be too difficult or too easy for a given team.

While the process may be simple to understand, the process of authoring scenarios for adaptive training experiences is anything but simple. Adaptive instruction, by its nature, requires additional content and more frequent assessment of individual learners/teams to identify opportunities to tailor SBT. Automation should be considered when designing adaptive SBT. Zook et al. (2012) identified several advantages to incorporating automated scenario generation tools and methodologies for use in SBT. A primary advantage is that automated scenario generation develops training scenarios with a broader diversity in a shorter time than human authors can produce. The *combinatorial optimization approach* that was part of Zook's approach provided both a diversity and a quality set of scenarios that are specifically tailored to each learner's needs and abilities as per the ZPD (Vygotsky, 1978).

The approach used by Rowe et al. (2018), used deep reinforcement learning methods to dynamically generate training scenarios specifically configured to support defined learning objectives and tailored to the goals, learning gaps, and abilities (e.g., prior knowledge) of individual learners. Folsom-Kovarik et al. (2019) describe the development of a scenario variation tool based on the novelty search genetic algorithm (Lehman & Stanley, 2011). Genetic algorithms act on a population of prospective solutions which evolve through mutation or crossover operations to create new prospective solutions through an iterative process. The solutions with the highest fitness ratings are selected for reproduction as the population evolves from one generation to the next. Novelty search guides evolution by novelty alone without explicitly specified goals. There is debate whether using novelty as a measure as a fitness or a genetic algorithm provides better results than a goal-directed fitness function.

The challenge of which approaches produce the best results is centered around the selection of evaluation metrics. Zook et al (2012) proposed a set of evaluation metrics that included replayability, tailoring to individual learners, and adaptation to changing conditions in the environment. Replayability focuses on a SBT system's ability to generate numerous, but distinct variations of existing scenarios. This is important to be able to exercise the learner's knowledge and skills and avoid circumstances where the learner can anticipate events and 'game' the system. Tailoring is a response by the SBT system to the needs and changing conditions of each individual learner. Tailoring has been shown to be highly effective in engaging learners. Finally, adaptation includes changes to events or scenario difficulty to align the individual learner or team with the knowledge and skill to be successful in a given scenario per ZPD.

Now that we have discussed competency assessment methods, the scenario development process, and automation options, it is time to capture major considerations in the scenario design and development process.

Considerations for Scenario Design and Development at Scale

Following along with the scenario generation process described above, scenario designers and developers should consider steps in the process, the source and acquisition of data, how data will be used to support adaptation, methods and approaches to scenario design and development at scale, and the limitations of various approaches and how these limitations affect desired outcomes. Scenarios provide the learner activities needed to acquire/maintain/exercise the knowledge and skills outlined in the learning objectives.

Know Your Training Audience

It is necessary to model individual learner or team competency as a basis for designing effective new scenarios or selecting existing scenarios. Consideration should be given to understanding what data is needed to determine competency and how competency will influence scenario adaptations. Additional consideration should be given to methods to infer competency. As noted, the xAPI achievement statements are becoming a military and IEEE standard. While this provides a record of achievements, these statements may not be sufficient to assess competency. Additional data (and context) may be needed to accurately infer individual learner or team competency. Automation should be an option for capturing learner behaviors as well as achievements at scale. Automation can reduce the workload required to model competency, but another consideration is the development and validation of a computational competency model (Nursikuwagus et al., 2018) that can be applied to a variety of instructional domains.

Define Your Learning Objectives

Learning objectives are composed of three elements: 1) conditions under which the task is to be performed, 2) an observable learner behavior that indicates accomplishment of the task, and 3) a measure that validates how well the learner can perform the task. Authoring tools that allow developers to define learning objectives and provide a guided process for tying learning objectives to activities will greatly reduce the work required to develop effective scenarios. It will be important to develop a repeatable authoring process that validates the scenario as 1) adequate to support the defined learning objectives, 2) composed of activities that influence learner behaviors required to assess the learning objectives, and 3) adequate to identify data sources required to assess learning. To scale across multiple learning experiences (e.g., a curriculum or learning pathway), a mechanism is needed to understand the scenario learning objectives relative to other objectives in the curriculum.

Selecting Appropriate Learning Activities

Learning activities are selected to create the set of conditions needed to acquire the knowledge and skill defined by the learning objectives. Learning activities are intended to stimulate experiential learning, critical thinking, collaboration or analysis. Learning activities are also intended to prompt observable learner behaviors to support measures of assessment. Examples of active learning activities are activities where learners participate in simulations, serious games, problem-solving exercises, and decision exercises.

Some learning activities include opportunities for deliberate practice of skills and application of knowledge in relevant environments. The ability to scale learning activities is the ability of the system to capture achievements below the course and lesson level at the activity level (e.g., assessment of problem-solving activities).

Tying Outcomes, Activities and Measures Together

As noted earlier, it is important to identify what learning outcomes (objectives) are targeted, what activities will stimulate appropriate learner behaviors, and how the progress of the learner or team will be measured.

Recommended Next Steps

In this section, we provide specific recommended next steps to achieve desired outcomes for the Generalized Intelligent Framework for Tutoring (GIFT) to support adaptive SBT at scale:

- Design – Understand and model the relationship between learner competency, learning objectives, learning activities and scenario content to design effective adaptive SBT
- Authoring - Automate or simplify the scenario authoring process to expand the available conditions in which learners can apply their relevant knowledge and skill
- Development – Select activities that engage the learner and exercise their skills in a measurable way
- Scaling – Adaptive SBT systems are complex capabilities that require special expertise to produce, and the skills needed can be drastically reduced along with cost/time through the use of guided learning, templates, reports, and automation

References

Dunne, R., Schatz, S., Fiore, S. M., Martin, G., & Nicholson, D. (2010, September). Scenario-based training: scenario complexity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 27, pp. 2238-2242). Sage CA: Los Angeles, CA: SAGE Publications.

Folsom-Kovarik, J. T., Rowe, J., Brawner, K., & Lester, J. (2019). TOWARD AUTOMATED SCENARIO GENERATION WITH GIFT. *Design Recommendations for Intelligent Tutoring Systems*, 109.

Hruska, M., Medford, A., & Murphy, J. (2015). Learning Ecosystems Using the Generalized Intelligent Framework for Tutoring (GIFT) and the Experience API (xAPI). In *AIED Workshops*.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, *19*(2), 189-223.

Lytras, M. D., De Pablos, P. O., Avison, D., Sipior, J., Jin, Q., Leal Filho, W., ... & Horner, D. G. (Eds.). (2010). *Technology Enhanced Learning: Quality of Teaching and Educational Reform: 1st International Conference, TECH-EDUCATION 2010, Athens, Greece, May 19-21, 2010. Proceedings* (Vol. 73). Springer Science & Business Media.

Magerko, B., & Laird, J. (2002). Towards building an interactive, scenario-based training simulator. In *Proceedings of the Behavior and Representation and Computer-Generated Forces Conference* (pp. 15-34).

Noori, N. S., Wang, Y., Comes, T., Schwarz, P., & Lukosch, H. K. (2017, May). Behind the Scenes of Scenario-Based Training: Understanding Scenario Design and Requirements in High-Risk and Uncertain Environments. In *ISCRAM*.

Nursikuwagus, A., Melian, L., & Permatasari, D. (2018, August). Computational model of student competency analysis in fuzzy topsis method. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012095). IOP Publishing.

Rowe, J., Smith, A., Pokorny, B., Mott, B., & Lester, J. (2018, May). Toward automated scenario generation with deep reinforcement learning in GIFT. In *Proceedings of the Sixth Annual GIFT User Symposium* (pp. 65-74).

Sottilare, R., & Brawner, K. (2018, June). Component interaction within the Generalized Intelligent Framework for Tutoring (GIFT) as a model for adaptive instructional system standards. In Proceedings of the *Adaptive Instructional System (AIS) Standards Workshop of the 14th International Conference of the Intelligent Tutoring Systems (ITS) Conference, Montreal, Quebec, Canada*.

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.

Sottilare, R. A., Long, R. A., & Goldberg, B. S. (2017, April). Enhancing the Experience Application Program Interface (xAPI) to improve domain competency modeling for adaptive instruction. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 265-268).

Zook, A., Lee-Urban, S., Riedl, M. O., Holden, H. K., Sottilare, R. A., & Brawner, K. W. (2012, May). Automated scenario generation: toward tailored and optimized military training in virtual environments. In *Proceedings of the international conference on the foundations of digital games* (pp. 164-171).

Vygotsky, L. S. (1978). Zone of proximal development: A new approach. *Mind in society: The development of higher psychological processes*, 84-91.

Wang, M. C. (1980). Adaptive instruction: Building on diversity. *Theory into practice*, *19*(2), 122-128.

Chapter 3 – Competency-Based Experiential-Expertise

Kevin Owens[1] and Benjamin Goldberg[2]
[1]Applied Research Laboratories: The University of Texas at Austin (ARL: UT); [2]U.S. Army DEVCOM Soldier Center, Orlando, FL

Introduction

In this chapter we deliberate on a learning model referred to as Competency-Based Experiential-Expertise (CBEE). CBEE is a result of methodical applied research conducted from 2009 to 2017 by the Applied Research Laboratories: The University of Texas at Austin (ARL: UT) for the US Navy on improving surface ship tactical operator and team-based performance. This research was based on post-event analysis of significant real-time performance data from multiple years of US Navy at-sea live exercises and real-world events, as well as experiential classroom training experiments.

CBEE is based on requirements that are extensions of similar past efforts to improve individual competence (James, 2015; Jaschik, 2017; Linked In, 2021; Marcus, 2020; McClelland, 1973; Nodine, 2016; Pew Research Center, 2017; Tavangar, 2014) but is more focused on team-based performance contexts which most work performance is in. CBEE is also part of a US Army science and technology research project being conducted to explore advanced training management tools in conjunction with the design of the Army's future Synthetic Learning Environment (STE) (Rosman, 2020). In addition, CBEE is being considered as part of the US Army's future learning concept for training and education (US Army Training and Doctrine Command, 2017), and the DoD Advanced Distributed Learning vision to modernize learning (Walcutt & Schatz, 2019) using ubiquitous web-based learning architectures and Adaptive Learning Systems (ALSs) similar to the US Army's Generalized Intelligent Framework for Tutoring (GIFT) product (GIFT, 2020).

In contrast to today's typical subject-matter formatted curriculum that began in the mid-19th century, CBEE is a more 21st century focused learning construct that builds what is called expertise in specific tasks through experiences that a person may perform in many different roles, in different teams and perhaps in different domains over their lifetime. This is actually how today's volunteer military structure works; personnel conduct *permanent change of station* (PCS) moves from duty-station to duty-station after two to four year "tours" in which they focus on different missions and/or assignment mostly using the same knowledge, skills and tasks but sometimes having to learn entirely new knowledge, skill-sets and tasks. The other characteristic that CBEE extols is for one to have any experiential expertise, they have to be objectively asserted as being competent at a specific level based on recent and valid data.

To support this learning model, complex competency development structures are required to support each of the domains in which experience and expertise are to be measured (Biech, 2008; Gilbert, 2007; Spencer & Spencer, 1993) along with finding ways to incorporate century-old philosophies of experiential education and practices (Itin, 1999). The CBEE model incorporates the adult-learning philosophy called andragogy (Knowles, 2015) in order to more rapidly and efficiently develop and sustain competence. To support the competency structure, cloud-based, standards enforced, competency frameworks will provide ubiquitous, reusable and machine-readable data structures that provide both an authoritative definition of organizational learning and performance requirements. CBEE is focused on objective evaluation by using data informed artificial intelligence (AI) supported evaluation tools and math-models for classifying levels of competence. The math-model will calculate a three-dimensional level of competence: (1) *how-well* one performs against a set of threshold-criterion, (2) *how-hard* the conditions were when they performed, and (3) *how-often* one has performed a competency within a set window of time (i.e. their experience).

In support of evaluating these competence dimensions, CBEE uses data standards and competency management services (CMS) that are ubiquitous across everyday information networks. Experiential learning data is delivered, managed and facilitated through CMSs (CASS project, 2021), with credentialing services (Credential Engine) established to store, track and support data of a team or person's competence across their lifetime. By employing a more ubiquitous andragogical approach around how humans naturally learn (Kamenetz, 2020; thecriticalthinkingchild.com, 2020; Young, 2020) and using technology to support self, instructor, or trainer led facilitated learning across local centers, schools, work-environment or online, CBEE promotes learning on-demand, and at a point-of-need.

CBEE assumes future individual and team competence development practices will not just employ locally accessed or online multi-media content. Nor, will it limit competence evaluation services and accrediting to traditional written test-based assessments (Merzenic, 2013; National Research Council, 2001; Nugroho, 2020). Instead, learning and competence evaluation occurs through live, full or semi-synthetic learning-environments that emulate real-world settings and conditions, and encourage mistakes and learning-by-doing using an experiential learning model. This approach has already been tested in various military, space and pilot training programs, as well as advanced academic learning efforts that incorporate virtual engagements (Board of Behavioral and Social Sciences and Education, 2018). These mediums benefit learning by allowing learners to build competence through reflection on noted gaps and mistakes, and repetition across multiple simulated experiences.

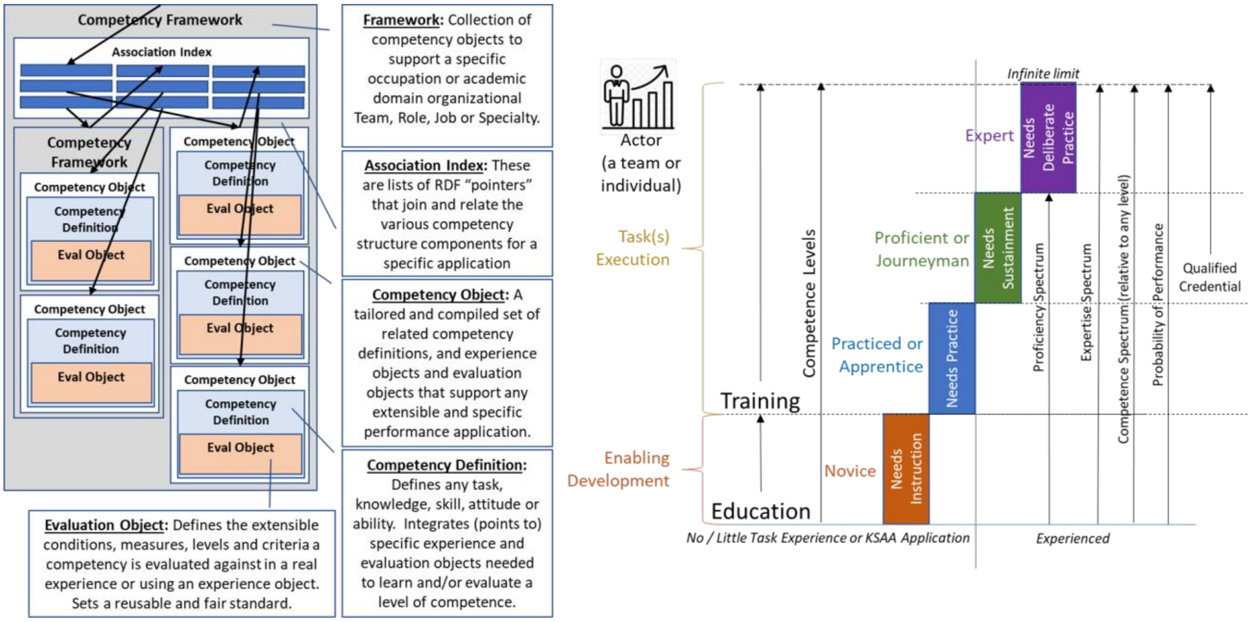
In the following sections we will provide a cursory discussion of the ontology required to incorporate CBEE, which includes reusable competency definitions, frameworks, and indexes, and tailored evaluation standards aligned to competency registries that are accessible on demand through the world-wide-web. We will then discuss the experiential learning model based on the work of past pioneers, and how that model develops the state (or level) of competence around what is called experiential-expertise. To conclude, we will briefly demonstrate the proposed CBEE architecture using the open source GIFT product. We will then summarize and make recommendations of future research needed to support the CBEE model.

Discussion

The CBEE Ontology and Conceptual Structures

An ontology is critical so that anyone trying to research, design or implement a new technology or endeavor is speaking the same semantic and syntactic language beyond just human language. This means terms, information and data having the same meaning, as well as having the relationships with other terms, information and data so that they are meaningful to the context of effort. An ontology also defines how entities of information are grouped to basic categories and their purpose. Figure 1 provides a summary illustration of this ontology.

At the left side of Figure 1 is an example of a competency structure. A competency structure is a little different from that of a typical task structure in that it is made up of multiple "entities" that capture and represent "classes" of performance definition and standards that can be "reused" and tailored for different performance and task contexts. This ensures performance requirements can be unique but follow a similar ontology (for machine readability) and to ensure comparable standards of performance for cross-domain collaboration and teamwork. Competency structures can be used for both academic and occupational/operational learning domains to ensure there is more consistency across these domains, and across associated institutions and organizations implementing a CBEE strategy. This approach allows teams and people to be evaluated and credentialed fairly and consistently as they apply their competence from academic to occupational domains or from one occupational job to another.



Framework: Collection of competency objects to support a specific occupation or academic domain organizational Team, Role, Job or Specialty.

Association Index: These are lists of RDF "pointers" that join and relate the various competency structure components for a specific application

Competency Object: A tailored and compiled set of related competency definitions, and experience objects and evaluation objects that support any extensible and specific performance application.

Competency Definition: Defines any task, knowledge, skill, attitude or ability. Integrates (points to) specific experience and evaluation objects needed to learn and/or evaluate a level of competence.

Evaluation Object: Defines the extensible conditions, measures, levels and criteria a competency is evaluated against in a real experience or using an experience object. Sets a reusable and fair standard.

Figure 1. The Competency-Based Experiential Expertise (CBEE) Ontology

Also shown on the left-side of Figure 1 is the CBEE competency structure (or model) consisting of six basic entity types: Frameworks, an association index, competency objects, competency definitions, and evaluation objects (see Figure 1 for descriptions of each). This structure is based on standards being defined by the Institute of Electrical and Electronic Engineers (IEEE) 1484.20.1 working group (IEEE, 2021) but refined to meet the needs of experiential-expertise. Entity relationships and their actual tangible representation will exist in what is defined as the Resource Description Framework (RDF) that is used to store and access structures like these over the world-wide-web, and are often written in what is called a Javascript Object Notation (JSON) text format for both storage and network transfer.

As shown at the right-side of Figure 1, the terms and meaning used in the competency level structures are a modified version of the Dreyfus model of skill acquisition (Dreyfus, 2004) and intended to provide a logical competence structure and taxonomy whether using the 13th-century apprentice model or a more modern taxonomy of expertise. The term "actor" - from the unified modeling language (UML) standard – is used to represent a team of people (a work group) or an individual performer (who performs a function in a team role, a job or in a specialty). Today the newest form of an actor may be an AI-based machine service or "assistant" that will be required to have minimum competence in low-level competencies; allowing humans to focus on more higher-level functions. This evaluation process and corresponding model structures will be discussed next.

Corresponding Competency Structures to Performance Structures

Frameworks are the natural clustering of competencies as they apply to organizational structures within an occupation or a learning event. In CBEE, frameworks should be established from existing organizational servers and/or academic courses and classes, and then the individual roles, jobs or specialties within those organizational entities. As noted in the right-side of Figure 1, frameworks can be within frameworks which is often the case the higher the team is in an organizational structure. This translation is how to implement a competency structure into Human Resource / Talent Management software and makes competencies better understood by managers and personnel whom are associated with hiring, promotion, and payroll management. Competency objects are any task one needs to learn, perform and evaluate or measure to attain or sustain a level of competence for a given team, role or job as determined by the framework.

CBEE Competency Evaluation Process

Figure 2 below shows the activity flow of an Evaluation Object. At the very left-side of the figure are the enabling elements needed to "start" any competency evaluation process: the performance *experience(s)* (be it live performance, exercise-based performance or lesson-based performance) with assessment events designed within them, and *actors* (team, individual, machine) that must respond to prompts from the assessment events, using one or more competencies.

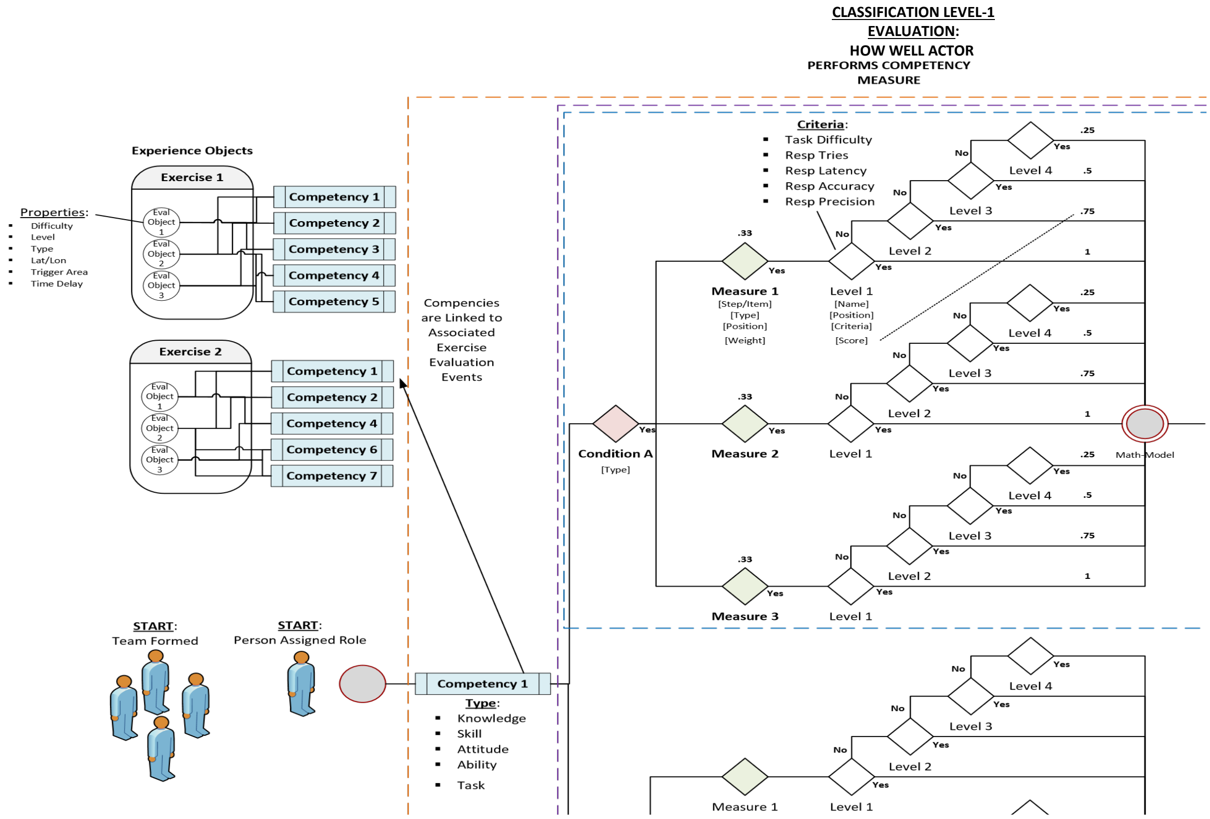


Figure 2. Competency Evaluation Level 1 Process

As shown in Figure 2, any competency will be pre-planned and naturally prompted within an experience during one or more assessment events. A competency can be associated with one or more specific conditions or use-cases that defines *how-hard* it is to perform a competency that not only gives context to the performance but gives its evaluation more weight in calculating competence. Conditions also indicate the level of competence needed to perform (for KSAAs, "conditions" may be the tasks they support). Which competency condition is used will be pre-defined in a designed experience and/or analyzed and identified in unscripted recorded live performance.

From each condition, an evaluation will take a path to one or more measurement points that will be different for each condition. This difference is necessary because different conditions may enhance or inhibit the same performance levels due to their setting (e.g., performing in a humid jungle at night vs. a brisk open field during the day) or prioritize what competencies are more critical over others at a given moment. Competency measures are the *how-well* an actor performs, and can be a task or skill-step, phase, knowledge-item or decision that is performed physically or cognitively. Measures can be one of two types: *measures of performance* (MOP) that are data sampled during an actor's real-time competency performance or *measures of effectiveness* (MOE) that are the result or outcome data sampled after an actor's competency performance. These measures establish a domain model (Professional Credential Services Inc., 2021) that can be applied and re-configured across all conditions, enabling consistent competence evaluation.

Next is the adjudication of what competence-level the actor's performance equates to. The number of competence-levels used in a competency evaluation object can vary but CBEE will use the four-level ontology stipulated earlier in Figure 1. Which level of competence an actor is evaluated at will be determined by a two-phase process that consists of (1) an AI supported manual classification process at the point of performance and then (2) an objective math-model discussed more below, based on data filtered through set criterion that is tailored to the type of competency and its unique measurement needs (e.g., a weapon competency, with a marksmanship measure will use criterion based on precision and accuracy). Once an actor's condition-based, measured levels of expertise are classified, they can be weighted or scored accordingly, and reported to a performance registry like a learning resource store (LRS).

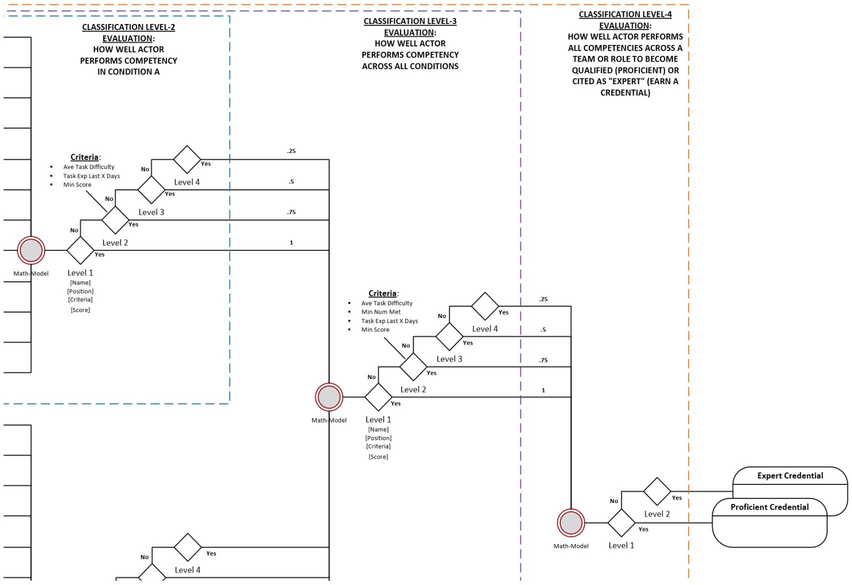


Figure 3. Competency Evaluation Level 2-4 Process

From that point in the evaluation process when a task has been evaluated following performance (manually with support of AI- based measurement and recommendation), a series of more complex classification-levels occur as shown in Figure 3. These later stage classification-levels will occur within a CMS that is outside the adaptive instructional system (AIS). As shown in Figure 3, each classification-level aggregates the previous classification-level outcomes until they ultimately determine a current state (percentage) and level (as defined in Figure 1) of competence. This ultimately determines, as applicable, if some form of credential is awarded.

Each competence-level adjudication process is done within a math-model shown in Figure 4. This model uses accumulated experiential data from previous and recently encountered live or synthetic "experience events" (xEvents) that are associated with specific metadata (task, mission, role), paradata (method data was collected), and what we will herein call "condata" (a combined series of context and conditions the performance was done in as described later). All this data is then multiplied by a longitudinal factor that accounts for performance repetition and/or decay over time (representing a team's cohesion or a person's neurologic state). This math-model can be in the form of a modified Rasch model (Rasch, 1960) or any other similar predictive performance math-model.

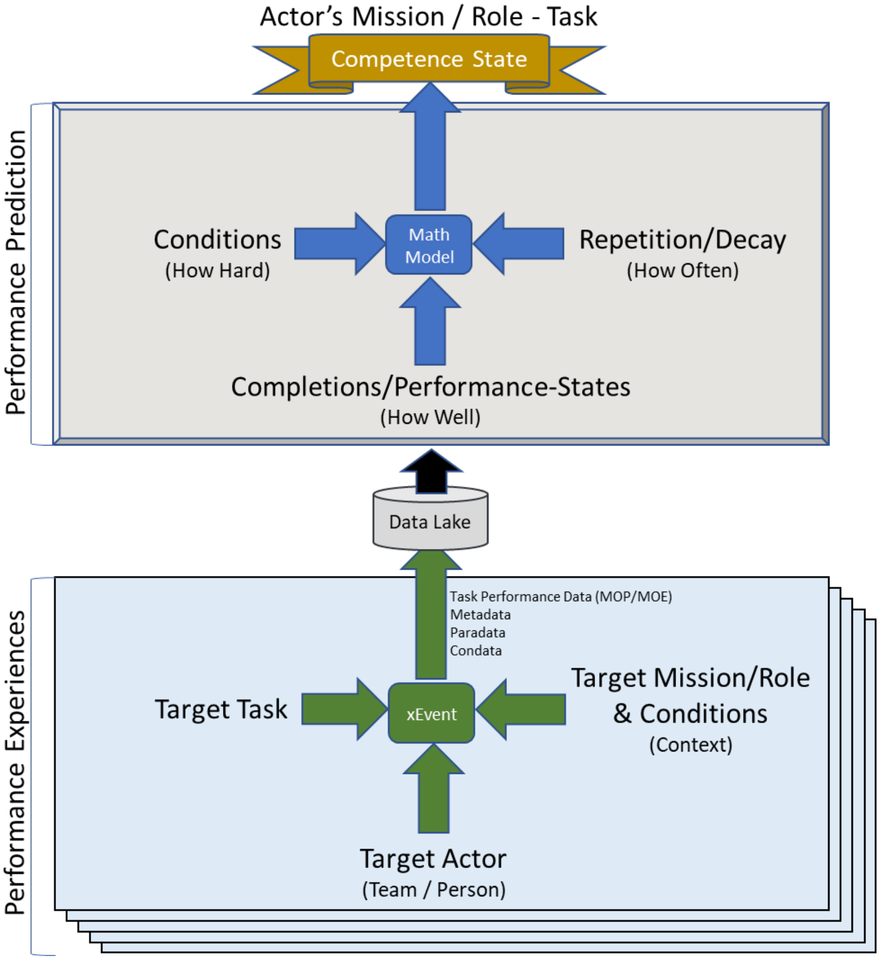


Figure 4. Experiential Expertise (Competence) Evaluation Model

As shown in Figure 4, each time an experience event occurs and specific task measure outcomes are reported in an actor's record, those levels are objectively calculated by the math function located within the CMS. It should be noted that each competency being evaluated is directly related to the three parts of the performance experience noted in Figure 4: the target task, a specific actor (team or person), and a context the task was performed in. It is critical that the range of xEvent conditions become part of the standard for a specific level of competence since that, along with temporal specifications are determining factors. As described later, the progress of one's competence level can be represented in a competency dashboard through representations such as color or a bar-chart, and a competency state in the form of a probability of performing the competency at an expert level.

In CBEE, one's experiential-expertise (competence) in a given mission/role task is presumed to follow a natural sigmoid curve of development. This pattern of growth is identical to other natural biological capability development including the brain's natural neuronal development. At a micro-level, when an actor is first motivated to develop a competency, they begin performing at a *novice* level because they lack the neuronal or muscular structure to perform at any better level. As the actor begins to practice a task over time and across different conditions of difficulty, their cognitive and/or behavior expertise increases as they become *practiced*. As they continue to practice they are able to perform the task more rapidly (through automated response). That along with their increased self-efficacy allows them to reach a minimum *proficient* level of predictable performance, for the specific role or mission. As the actor continues to practice (and is measured), in more different conditions and with more difficulty, their level of expertise increases into the highest domain of competence – an *expert*, thus they are predicted to perform competently regardless of condition. When in this domain of expertise, their performance improvement begins to level off - requiring significant deliberate practice in more challenging conditions to increase it any more. A similar process occurs for developing expert teams in that each team-member needs to qualify in their respective roles, as well as develop teamwork, and then all members must learn to work together to best accomplish the actual team-tasks as well. As the team performs more and more together they will begin to rapidly increase in each of these elements until they reach an expert level which is when each member can essentially anticipate each other's actions and communication is minimized; all of this happens on a sigmoid curve as well.

Methods

The Experiential-Expertise Process

Experiential-expertise is another rung on the evolution of the original thought that began with the broader philosophical term "experiential education" coined in the early 20th-century by Dr. John Dewey (Dewey, 1938). Experiential education was a push against the overly didactic, pedagogical methods of education that were pervasive then, and are still pervasive today. Next came research by Dr. David Kolb that was called "experiential learning" (Kolb, 1984) that took experiential education and focused it on the individual learning process itself within the educational system; however still focused within the academic context (illustrated in Figure 5 below). Standing on these past principles and efforts, experiential-expertise is the next evolutionary inflection-point in the experiential narrative.

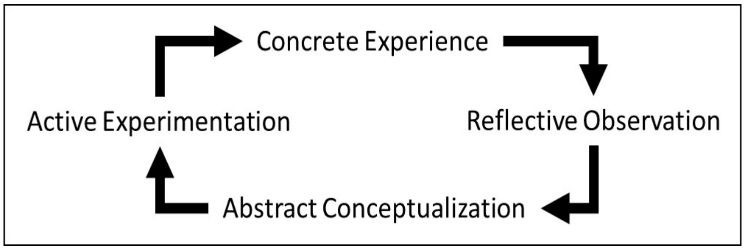


Figure 5. The Original Experiential Learning Model

The basic premise of experiential-expertise is that academic learning and professional occupational learning are both part of the same common social mechanism to develop the future workforce, government employees and even military membership. To produce that common learning framework, the employment of modern ubiquitous internet-based competencies is the most efficient and effective way of doing that. Experiential-expertise suggests that in all modern learning, the ultimate goal is the application of what is learned in experiential practical endeavors. This application includes the collection of data and evaluation of how well what one learned is *capable* of being performed in the future - i.e., competence. Unlike traditional evaluation methods that reference measures against a minimum standard or criterion, Experiential-expertise references all performance against the highest standard baseline - measured as *expertise*, as defined by the continuously collected raw performance data across a domain.

Experiential-expertise is developed in context to real prompts, demands and supported by “just-in-time” and “just-enough” inquiry (resulting in multiple interleaved KSAA learning episodes or “micro-learning” (Alagumalai et al., 2005)), as well as performing in different conditions, over many iterations. The type of performance data collected and used to measure an actor’s performance expertise includes: system actions, video, kinematics, eye-tracking, voice, and positions. As an actor begins to practice a competency over and over (and apply KSAAs) in varying conditions, their expertise and efficacy increases. Observations of modern video game players learning new role-playing or first-person performance games show a similar process. The same is true for kids learning to ride a bike for the first time, the latter also requiring previous expertise in more fundamental competencies such as balance, inertia, and mechanical energy transfer; all usually learned through experiential means. Thus, CBEE asserts, with support from neuroscience, this is how humans best learn, especially in the more technical, technology based future workforce.

CBEE extends the basic experiential learning model in Figure 5 within a competence structure by integrating it in the evaluation math model described above that adds the evidence required for objective *expertise assertion*. This modified model executes as follows (see Figure 6):

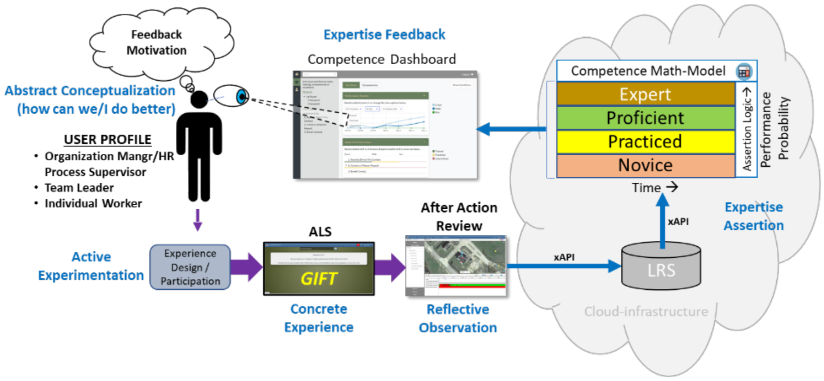


Figure 6. CBEE Application Use-Case

(1) *Concrete Experience.* This phase begins when an actor cognitively and/or physically interacts with prompts or orders to perform a task (a competency) in a given condition, and within a live or synthetic learning environment. These are termed experience events, and are the units of experience used to calculate specific targeted task competence. These events also provide the actor more structure with which to categorize the performance better in memory.

(2) *Reflective Observation*. The recorded concrete experience data can be played back to the instructor or trainer and the actor immediately after their performance using data analysis and playback tools. This reflective process can and will likely occur again later as part of all future performance; it is in fact the learning product. In this phase, the actor is given the opportunity to consider what parts of their experience performance worked and what needed improvement. These reflective points are highlighted by ALS automated measures as well as manual subject matter expert measures which then are used as data for expertise assertion.

(3) *Expertise Assertion*. This phase is added to Kolb’s original model and modifies it to the experiential-expertise moniker. As described earlier, this phase is when the CMS math-model begins using the past performance measured outcome trends across multiple concrete experience

phase outputs, and is represented as an expertise state. These inferences are then visualized in an online competency indicator format, accessible from appropriately configured devices.

(4) *Expertise Feedback*. This phase is also added to the Kolb model to provide the summative feedback one needs to understand their gaps with data to support the levels of competence provided. The actor or leader or manager can now view the competency levels and states of interest to help define what improvement strategy is needed. They can also compare competence against others' in similar positions. This phase is also critical to mitigate what is called the Dunning-Kruger effect which is the false-assumption of expertise/proficiency without such feedback being present.

(5) *Abstract Conceptualization*. With the feedback provided, actors, leaders or managers can now conduct abstract conceptualization of improvement. This not only develops targeted goals but can be attained by receiving intelligent recommendations (from AI within the Adaptive Learning System) on strategies to improve a competence level in a future performance.

(6) *Active Experimentation*. Based on the goal developed from the previous phase, an experience (novel or repetition in performance) is sought or produced in a cyclical pattern to improve. This results in experimenting with new conditions that will help the actor attain the desired level of competence.

Here are the key enabling capabilities to be successful in employing experiential expertise (as facilitated through future learning technology):

- Set competency definitions and standards are established with declared criteria.
- Raw performance data is collected to support evaluation and feedback to the actor.
- Actors have the necessary core competencies that enable associated tasks or skills to be performed as part of experiential learning.
- Actors are *willing* to be actively involved in the learning experience (meaning they are interested in the task being performed – thus the need for andragogical learning).
- The actor is able to reflect on what was learned through the experience (facilitated through data).
- The actor is provided tools to analyze a past performance experience and to conceptualize future alternative performance.

Recommendations and Future Research

This chapter discussed the basic ontology of CBEE, the concept of the basic competency structure and how it corresponds to real occupation or academic structures and performance, how evaluation occurs, the experiential learning model, and demonstrated a use-case of CBEE based experiential learning. Several lines of research are still needed to support and improve the CBEE model.

These lines of research include:

- Establish a CBEE course template in GIFT that enforces the five phases of interaction. Produce better synergy between a CMS and available content in an ecosystem environment, supporting the abstract conceptualization phase.
- How to describe the competency elements a GIFT lesson supports must be improved.

- Building better GIFT mechanisms for building a competency state from a single interaction.
- Better technologies to manage, update, and sustain competency structures.
- Improved methods of providing competence feedback and awareness following experiential learning or performance.
- Improved synthetic and live data collection technologies and translation algorithms.
- Improved AI models that support the automatic detection and classification of activity into competency-based real-time evaluation algorithms.
- Improved research on classifying teamwork expertise and the factors that define expert from novice teams.
- Technology to integrate this learning model and its data collection needs into everyday life, from personal devices, wearable technology, and even fixed technology (televisions, security cameras and sensors).

References

Alagumalai, S., Curtis, D.D. & Hungi, N. (2005). Applied Rasch Measurement: A Book of Exemplars. Springer-Klumer.

Biech, E., ed. (2008). ASTD Handbook for Workplace Learning Professionals. First ed. Danvers, MA.

Board of Behavioral and Social Sciences and Education (2018). How People Learn II: Learners, Contexts and Cultures. National Academies Press: Washington DC.

CASS project (2021). Competency and Skills System (CaSS). Available at: https://cassproject.org/. Retrieved on: 10 January 2021.

Credential Engine. Available at: https://credentialengine.org/. Retrieved: 2021-03-15.

Dewey, J (1938). Experience & Education. New York, NY: Kappa Delta Pi.

Dreyfus, S. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology and Society, 24*(3), 177–181.

GIFT (2020). Generalized Intelligent Framework for Tutoring (GIFT) Documentation. Available at: https://gifttutoring.org/projects/gift/wiki/Overview#Background. Retrieved 2021-01-16

Gilbert, T.F. (2007). Human Competence: Engineering Worthy Performance. Pfeiffer, San Francisco, CA.

IEEE P1484.20.1 Working Group (2021). P1484.20.1-2021™/D1 Draft Standard for Learning Technology Data Model for Reusable Competency Definitions. Available at: https://standards.ieee.org/content/ieee-standards/en/standard/1484_20_1-2007.html . Retrieved: 2021-02-11.

Itin, C. M. (1999). Reasserting the philosophy of experiential education as a vehicle for change in the 21st century. *The Journal of Physical Education, 22*(2), 91-98.

Jaschik, S. (2017). A Plan to Kill High School Transcripts … and Transform College Admissions found at: https://www.insidehighered.com/news/2017/05/10/top-private-high-schools-start-campaign-kill-traditionaltranscripts-and-change . Retrieved: 2021-02-04

James, G (2015). Colleges Aren't Preparing Students for the Workforce: What This Means for Recruiters. Available at: https://business.linkedin.com/talent-solutions/blog/2015/07/colleges-arent-preparing-studentsfor-the-workforce-what-this-means-for-recruiters

Kamenetz, A. (2020). Survey Shows Big Remote Learning Gaps For Low-Income And Special Needs Children. Available at: https://www.npr.org/sections/coronavirus-live-updates/2020/05/27/862705225/surveyshows-big-remote-learning-gaps-for-low-income-and-special-needs-children. NPR. Retrieved on: 202103-08.

Kolb, D. (1984). Experiential Learning: experience as the source of learning and development. Englewood Cliffs, NJ: Prentice Hall.

Knowles, M.S. et al. (2015). The Adult Learner (8th Edition). Routledge: London and New York.

Linked In (2021). 12 Jobs You'll Be Recruiting for in 2030. Available at: https://business.linkedin.com/talentsolutions/blog/future-of-recruiting/2018/12-jobs-you-will-be-recruiting-for-in-2030. Retrieved: 2021-01-05

Marcus, J. (2020). How Technology Is Changing the Future of Higher Education. Found at: New York Times: https://www.nytimes.com/2020/02/20/education/learning/education-technology.html. Retrieved: 2021-02-09

McClelland, D.C. (1973). Testing for competence rather than for intelligence. *American Psychologist, 28*, 1-14.

Merzenic, M (2013). Soft-Wired: How the New Science of Brain Plasticity Can Change Your Life. Parnassus Publishing, San Francisco, CA.

National Credentialing Solutions (nCred) (2021). Available at: https://nationalcredentialing.com/. Accessed on: 2021-03-21

National Research Council (2001). How People Learn I: Brain, Mind, Experience, and School: Expanded Edition. National Academies Press: Washington DC.

Nodine, T.R. (2016). How did we get here? A brief history of competency-based higher education in the United States. Found at: https://doi.org/10.1002/cbe2.1004 . Retrieved: 2021-02-09

Nugroho, D., et al. (2020). COVID-19 Trends, Promising Practices and Gaps in Remote Learning for PrePrimary Education. Available at: https://www.unicef-irc.org/publications/1166-covid-19-trends-promisingpractices-and-gaps-in-remote-learning-for-pre-primary-education.html. UNICEF. Accessed on 2021-14-02

Pew Research Center (2017). The Future of Jobs and Jobs Training. Available at https://www.pewresearch.org/internet/2017/05/03/the-future-of-jobs-and-jobs-training/ . Retrieved: 2021-01-05

Professional Credential Services, Inc. (2021). Available at: https://www.pcshq.com/?page=pcshistory. Accessed on: 2021-03-21

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests.(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rosman, J. (2020). The Synthetic Training Environment. Available at: https://www.ausa.org/sites/default/files/publications/SL-20-6-The-Synthetic-Training-Environment.pdf. Association of the US Army.

Spencer, L.M. & Spencer, S.M. (1993). Competence at Work: Models for Superior Performance. Wiley and Sons Inc.

Tavangar, H. (2014). The Out of Eden Walk: An Experiential Learning Journey from the Virtual to the Real. Available at: https://www.edutopia.org/blog/out-of-eden-experiential-learning-homa-tavangar . Edutopia, January 3, 2014. Retrieved on: 2021-01-16

thecriticalthinkingchild.com (2020). Closing Learning Gaps During Remote Learning. Available at: https://www.thecriticalthinkingchild.com/closing-learning-gaps-during-remote-learning/. Accessed on 2021-03-28.

US Army Training and Doctrine Command (2017). The US Army Learning Concept for Training and Education (2020-2040). Available at: https://adminpubs.tradoc.army.mil/pamphlets/TP525-8-2.pdf . Retrieved on 14 Feb 2021.

Walcutt, J.J. & Schatz, S. (Eds.) (2019). Modernizing Learning: Building the Future Learning Ecosystem. Washington, DC: Government Publishing Office.

Young, J.R. (2020). Sudden Shift to Online Learning Revealed Gaps in Digital Literacy, Study Finds. Available at: https://www.edsurge.com/news/2020-10-01-sudden-shift-to-online-learning-revealed-gaps-in-digital-literacy-study-finds. EdSurge.com. Accessed on 2021-03-08.

Chapter 4 – The Training Assessment Framework: Innovative Tools Using Scenario-Based Assessment and Feature Analysis

Eva L. Baker[1], Kilchan Choi[1], and Harold F. O'Neil[2]
[1]University of California, Los Angeles/CRESST; [2] University of Southern California/CRESST

Introduction

This chapter describes advances using the Training Assessment Framework (TAF), a conceptual architecture for the design, development, and establishment of validity for assessment of training and education in which innovative tools for design and validity were developed. This includes scenario-based performance assessment and feature analysis. The project was conducted with the support of the Naval Education and Training Command (NETC) over three years. This project was developed to build reusable guidance, prototypes, methods, and tools useful for Navy training assessment. Research was not a principal purpose of this project, although comparative solutions to assessment and methodology problems were explored and evaluated as needed to meet the project requirements. In addition to discussing the overall project, this chapter also includes additional detail for two specific methodological components of the project: (1) scenario-based performance assessment (PA), an innovative assessment format used throughout our research and (2) the application of a new approach to validity, feature analysis (FA) (Baker & Choi, 2019; Chung & Redman, 2015). Their applicability to intelligent tutoring systems (ITSs) is relevant to systems whose goals include assessment of open-ended, complex learning. For example, the Generalized Intelligent Framework for Tutoring (GIFT; Goldberg, 2021) provides such opportunities.

The term, 'Training Assessment Framework', evolved to describe the overall project as well as the particular architecture that was the basis of the prototype assessments for end-of-course examinations at the A-Schools at Naval Education and Training Command. The chapter considers our larger aspirations for the use of the TAF as a general assessment development tool as well as the goals, methods, and outcomes achieved over the three years of the project focused on technical training.

Goals of the TAF Project

A set of ambitious goals motivated the project, with proximal goals to create A-School assessment prototypes to exemplify the objectives. A-Schools in the Navy provide initial technical training for various Navy jobs (i.e., ratings) for new enlisted Sailors. In an attempt to strengthen previous design and development assessment models, the TAF project was launched with the idea to create a structure to support the quality of end-of-course assessments for each rating. Another goal was to consider the assessment *system* in use in A-Schools and to explore the value of a common framework to increase the coherence among the end-of-course assessments of different ratings. Our focus was to create assessments with high fidelity to the goals of particular courses tied to job training by developing a framework and applying it across ratings. If a common framework were workable, it could serve to give the A-School end-of-course assessment system greater coherence overall as well as to provide a basis for judging the comparability of assessments of different content.

In addition, the framework was conceived to serve multiple purposes of different but related assessments. To support the findings of end-of-course assessment, we wished to have impact as well on testing conducted within courses to support learning, such as interim or formative assessments, analytic models applied to computer-based instruction, or the provision of feedback to instructors and other test developers. For the most part, we were unable to meet this goal in the Naval Education and Training Command project as instructional systems were already in place and could not be modified to conform to the TAF provisions. An additional training use for the framework was to develop the assessments that could be used to generate measures of transfer and application, that is, to document the degree to which trainees could demonstrate their learning under varied conditions or requirements similar to the modifications they would naturally encounter when they were sent to duty stations with different contexts and expected roles. Happily, transfer and application are currently a part of a recently awarded project.

The goal of clarifying the development of performance assessments and employing them in Navy training settings grew in its importance over the course of the project. For instance, we were asked to develop a prototype performance assessment for use in the Navy-Wide Advancement Examination. We agreed because we were interested in the robustness of the TAF across purposes.

The TAF encompassed not only design parameters for the development of assessments, it also focused on a range of protocols intended to support quality inferences about tasks, items, and tests, as well as exploring validity for various training purposes (AERA, APA, & NCME, 2014). These validity protocols resulted in sub-goals addressing design and execution of the formative evaluation procedures involving both qualitative and quantitative data collection procedures and interpretations. Constraints include involving administration time, varying lengths of instructional time, and different sizes and schedules of training cohorts. These conditions required the addition of objectives focused on the methodology needed to draw conclusions about quality, such as the reliability and validity of items, as much prior work has focused on large-data sets. As Naval Education and Training Command used instructional courses with limited numbers of trainees, the project methodology team needed to devise validity approaches that reflected the sampling constraints and include them in the TAF itself.

Description of the TAF as an Assessment Tool

This description focuses at the outset on the overall TAF project and then provides detail on the design and use of performance assessments, promotion examinations, and feature analysis. The methodological description considers the design, development, and revision of the TAF over the life of the project. The TAF was designed with multiple assessment purposes in mind and with built-in attention to the quality and validity of the assessments emanating from the design. As a broad template, we used a software design document to structure the Training Assessment Framework (ISO/IEC/IEEE, 2011) that included a wide range of stakeholders, purposes, and requirements intended to clarify the framework.

The overall plan was to create the TAF, and to use it to guide the generation of prototype tasks and items for two very different ratings—Fire Controlman (FC) and Damage Controlman (DC)—in an effort to determine the robustness of TAF for skill and content requirements. Beginning with earlier versions of model-based assessment (Linn et al., 1991; Baker & Niemi, 1996; Baker, 1997, 2007; Baker & Chung, 2002) our approach embodied design features that went beyond the usual exclusive focus on content. Following approaches developed by Bloom (1956), Gagne (1974), Gagne and Briggs (1979), and Anderson and Krathwohl (2001), we first began to design with the consideration of cognitive demands. Cognitive demands, or required thinking or processing, involves types of learning and performance such as comprehension of facts and principles, procedural learning, problem-solving, search and analysis, and systems thinking. These cognitive demands are always to be embedded in the content (subject matter) and skill objectives for each rating of interest. A second precept of the TAF was determining and explicating the range of tasks or item formats that could be used to elicit appropriate responses to the cognitive demands and choosing the best options for the goals. Although multiple-choice was the dominant assessment format at Naval Education and Training Command, in the first year of the project we explored alternative ways to assess understanding of facts and principles as well as procedural learning using available item templates from Questionmark Perception, a computer-based system that Naval Education and Training Command was using to create and administer examination items (Questionmark Computing, 2013). As our main interest was furthering the fidelity of the assessment both to the courses and to the jobs to which the Sailors would perform following A-School, we also were committed to Naval Education and Training Command to create and implement innovative approaches to assessment. We created specifications for performance assessments designed to measure procedural learning, such as following a set of rules or principles to complete a task, and problem-solving that allows a more open-ended response. The design of performance assessments also required attention to scoring rubrics, a topic to which we will return.

We also involved subject-matter experts (SMEs) throughout the process, who included instructors or other individuals made available by Naval Education and Training Command or veterans with recent experience in the ratings of interest to assist us on TAF components. Subject-matter experts provided significant input to the development of domain content and skills. Creating the content model for each rating was a key requirement. We adopted an approach using an ontology for each rating to describe and depict key content knowledge and principles and their structural and functional relationships to one another. Input came from Navy curriculum and requirements documents as well as SMEs. These ontologies underwent revisions in content and representation over the project, fixing on a set of network depictions that included the importance of ideas (nodes) determined by the frequency and importance with which the nodes were connected to other nodes. The connections in the form of directional links, and each proposition, node-link-node, showed what the relationship was, including links such as *of part or type of, causes, precedes*, for example, Chung et al., 2003; O'Neil and Chung, 2011; Nye et al., 2018; Baker, Choi, Kao et al., 2019. The ontology functioned not only to highlight the content and relationships of greatest importance, but also to bound the limits of the domain, in other words, specifying which content was fair game for assessment and which was "out of bounds." SME input to and review of the parameters of the content model as well as its internal relationships were essential. The content model as represented by the ontology also served to clarify the validity focus of the project and emphasize attention to the explicated domain rather than a broad construct. This distinction had implications for the multiple validity studies conducted here as well as for validity methods in general.

<u>Formative Evaluation</u>. Two types of formative evaluation were implemented: qualitative and quantitative. The first evaluation involved the close inspection of the actual items, task, and examination. Following the development of draft assessments using the TAF requirements of cognitive demands, task demands and responses, and content models, drafts were reviewed systematically by SMEs to determine the importance and clarity of the assessment task or item. In this review and revise stage, we used the qualitative ratings of item and tasks characteristics described in the feature analysis section. These reviews resulted in revisions prior to administration to trainees. In addition, the methodology of feature analysis, where items and tasks were rated to see the extent to which they actually exhibited desired TAF elements, i.e., cognitive demands, task requirements and a content model. In addition, tasks and items were reviewed to determine their reading difficulty.

Quantitative formative assessment required data collection from trainees under actual conditions of use. There were multiple reasons to collect empirical data on the assessments. They included verifying the usability of the assessments by the trainees. Part of the usability involved the function of the technology platform for the assessment (iPads) and the comfort and ease with which the examinees engaged with the assessments provided on them. As we collected process data for each examinee, in addition to direct feedback, we were able to evaluate usability, including the time spent on various tasks. Because some of the item types were unfamiliar to the examinees and required new types of responses, we wished to assure that good information was collected. An overriding concern of our on-site data collection was our ability to fit into the existing classroom and school environments, minimizing burden to all participants. Because considerable attention was given to assuring the data collection team understood Navy protocols and comported themselves as they were part of the Naval Education Training Command team, we received excellent cooperation throughout. In addition, usability concerns extended to how well we could negotiate the technology environment in place at the sites. Our requirements were to avoid official Navy technology systems, upload examinee data, and otherwise conduct the trials with efficiency and clarity. To do so we provided our own devices and servers to simplify administration and uploading.

To conduct quantitatively-oriented formative evaluation, we developed designs intended to address specific purposes. For example, pretest-posttest designs were used, but each rating brought difficulties in implementation. Some ratings were relatively short in duration, e.g., weeks instead of months. The school week was Monday-Friday from 7:30am-4:00pm which would allow the comparison of the same trainees on a pretest and posttest basis, in order to determine whether the assessments were sensitive to instruction. If no changes were seen, the measures would be reviewed to assure that they were responsive to the delivered instruction. Another model used a posttest-only design that compared new trainees with those ready to graduate and was employed for the Fire Controlman rating that required several months of instruction. Because the length of each project contract (one year) would not allow us to follow up with the same group in the Fire Controlman rating, different cohorts before and after instruction provided data. Obviously, we had no control over the assignment of trainees to groups, and because randomization was not possible, there was a strong likelihood that the groups may have differed *a priori* from one another. A usual adjustment to be made in non-equivalent groups requires the collection of detailed background information to serve as covariates, but limitations on the collection of such information were in place. Thus, growth attributed to instruction could be approximated at best, and the question of instructional validity was not clearly resolved.

Nevertheless, the conduct of empirical trials also gave us the opportunity to administer the assessment to instructors or Naval Education and Training Command SMEs to determine their performance. The original purpose was to contrast expert performance with that of novices, an approach we have used in the past to set criteria for scoring rubrics in open-ended assessments. However, in this project, in addition to determining the ability of the measures to distinguish between expert and novice groups, we used expert performance as a means to set a quantitative standard based on their expertise. This approach contrasts with other training environments, where the performance standard, or cut-score for passing, is arbitrarily set at 80 or 90%. This type of criterion, if applied to the whole test operationally, considers each item as approximately equally difficult and drawn from the same construct. It is clear that the 80% criterion can be achieved by manipulating the difficulty of the examination or allowing multiple trials to reach the criterion, or both. In our case, we had multi-dimensional components of the assessment. We decided to use the average of expert performance on the assessment as the criterion level. Therefore, we reported trainee performance as a proportion of the experts' scores. As with much real-world development, this solution also presented difficulties. For example, for some ratings, instructors only taught a subset of goals as part of an instructional team so they would not be fully proficient in all course components.

Additional information was gathered in formal data collection using revised assessments, in particular, results of short measures of self-efficacy and anxiety, to determine whether trainees were uncomfortable with unfamiliar assessment formats (Baker, Choi, Iseli et al., 2019).

Description of Scenario-based Performance Assessments

Performance assessments as they were formulated in the TAF project were high fidelity tasks synthesizing important learning in the ratings under study. While performance can be stimulated by text or even brief instructions, for instance, to write an essay on a particular topic or process, our approach involved selecting tasks that required combining sets of skills, portraying them in a scenario or setting that directly mapped to significant course requirements, and including, in some efforts, requests for explanations of why the examinees made their decisions. The development of performance tasks followed the TAF sequence described above and they were included in the testing administration. Performance assessments were generated based on cognitive demands, particularly those that involved multiple steps, such as problem-solving, procedural learning, and search, all of which were applied to the content of the rating. Two notes are important. First, we were able to use in the scenarios partial art and software support created for development on a parallel project: The Navy Life Game (Koenig et al., 2020). This project presented examples of rating tasks in a game format in order to familiarize individuals with various Navy jobs. This project allowed us to import some of the scenario aspects that contributed to verisimilitude of the game tasks. For example, we used operationally functional work stations and resources for the Fire Controlman rating or settings involving casualties (fire and flood) and their remedies used by the Damage Controlman rating. Another distinction of performance assessment development is the need to create and implement scoring rubrics for constructed, or open-ended, responses, using automated approaches.

How did our work differ from traditional performance assessment? For most performance assessment development, great attention is given to the creation of the one and only perfect situation to elicit responses, and far less thought devoted to how performance should be judged and evaluated. Unique performance assessments each would have its own special scoring approach. As a result, to support coherent instruction and assessment we have identified principles that would be used in scoring, but differentially exemplified as appropriate to particular content. To clarify and operationalize scoring, and to get agreement from SMEs and measurement experts, we decided to include multiple explicit steps in the performance task to provide support for the examinee as well as to give us a strategy to develop and validate scoring rules. Given that the performance assessments were derived from the TAF design and would be subject to revision based on qualitative and quantitative data, we focused both on the TAF requirements and the quality of representations presented to the examinees. Although rubric development is usually one of the more challenging components of performance assessment development, in the Naval Education and Training Command cases it was relatively easy. For one thing, there is clearly a Navy “way” or “ways” of doing complex tasks, and respondents are not encouraged to free-lance unless the situation is previously un-encountered.

Another opportunity developed in the performance assessment area. We were asked to consider the use of the TAF performance assessment model and to create performance assessment prototypes for the Personnel Specialist promotion to E-5 and E-6 (Baker, Koenig, & O’Neil, 2019; Choi et al., 2019). For promotion, performance on the Navy-Wide Advancement Examination is an important element but not the single criterion for promotion. The current exam uses 175 multiple-choice items. To keep within the examination time limits for administration, the number of multiple-choice test items would need to be reduced substantially to make time for the longer performance assessment task. Thus, a corollary objective was to determine what number and selection approach of Navy-Wide Advancement Examination multiple-choice items would achieve an equivalent level of prediction as the longer version of the exam. The goal of clarifying the development of performance assessments and employing them in Navy training settings grew in its importance over the course of the project. Thus, an additional task was to determine the reduced number and source of multiple-choice items that could predict Navy-Wide Advancement Examination performance as well as the 175-item set.

A more general development challenge, largely unsolved in scientific literature and practice, is the generation of performance assessments that are coherent and from which one could infer a common set of skills. Creating performance assessments as unique objects limits their exchangeability and potential use for multiple purposes. Instead of administering the same performance assessment twice on a pretest and posttest, one would want to be able to document that the designs used in each were comparable. One of the more interesting challenges involves how one generates scenarios for use in performance assessments. In the Navy examples, these were happily limited to particular objectives. For instance, tasks for the Damage Controlman rating might require navigation to find the location of the problem, the number of team members involved, and the numbers of concurrent casualties, such as different types of fires, floods, or causalities that involved injuries. The rules for action in the assessment are specified in instruction based on prior risk assessment. In developing the performance assessment for the Personnel Specialist rating, we needed to decide what actions were required for the trainee to understand the personnel request, find the correct form, fill it out carefully and error free, and send an email to execute actions.

In the past, we have used simple combinations to generate the characteristics of performance assessments that would cover the high probability situations. For instance, a set of parameters might include four core scenarios, three cognitive demands, two task options, and four response options that would generate a large number of potential combinations as a design pool for the number of performance assessments. They can rapidly be edited and subsets selected for testing.

The reactions to the performance assessments were positive across the range of leadership, management, instruction, and trainees. Operationally, the performance assessments presented a few technical problems. The first was to determine their weighting within the test which also contained a number of comprehensive multiple-choice items related to knowledge of and about the course content. Given that the administration time was fixed, we decided to score the performance assessments partly on their challenge, weighted by the proportion of time they took. In our first-year trials, the performance assessments were considerably longer and we determined that we could better use the time by presenting additional shorter performance assessments to the examinees.

Description of Feature Analysis

One of the relatively recent methods to assess the quality of assessments used in the Naval Education Training Command TAF project is feature analysis (Chung & Redman, 2015; Kao et al., 2018; Baker & Choi, 2019). Feature analysis (FA) combines qualitative and quantitative data to characterize items or task components and their empirical impact on test performance in order to support development and to contribute to validity inferences about assessments. Our notion was that FA could address “functional” validity, that is, how attributes of items and tasks affected performance. To begin, qualitative ratings by items and tasks are conducted by teams of trained raters who understand assessment parameters and the relevant content domain(s). The ratings are tagged in the data (assuming a sufficiently large sample) and examined in terms of their relationship to examinee performance. The results show the association of features to different types and levels of examinee accomplishments. Findings may have implications for assessment design and revision of the instructional intervention, administration, and scoring. FA can be conceived as serving both prospective and forensic views of validity, with prospective uses focused on design and development of assessments and forensic uses examining analyses following the administration of the measure under planned conditions.

FA was used in the development of the Naval Education Training Command A-School end-of-course. FA was used principally to verify that TAF elements could be observed in the developed tasks and items for Naval Education and Training Command ratings. Where shortfalls were found using FA, tasks and items were revised. In particular, FA was useful in performance assessments in which the scoring rules affect cognitive demands, for instance, comparing problem solving with procedural tasks. In each of the cycles of design, revision, and further development for each rating, feature analysis was employed with agreement of raters being documented. Our experience with the TAF suggests that FA should be considered as a methodological tool key to the development of assessments.

The results of the TAF at Naval Education and Training Command were documented in a range of reports (Baker, Choi, Kao et al., 2019; Baker et al., 2021). They document that the TAF functions as intended and can assist in the generation of assessments useful in training. We realized that our contribution to assessment included not only the design parameters for cognitive demands, content models, and formative evaluation, but that our approach to creating a coherent design strategy across ratings also resulted in a “proof of concept" in the ratings that we used. The ratings varied in terms of requirements and content, ranging from technological, emergency-oriented, and logistics content. Furthermore, our design and implementation of performance assessments showed their utility, and resulted in positive responses by Navy personnel from senior Navy leadership to trainees. The idea that tests can show greater fidelity to real tasks was demonstrated, along with reliability and validity evidence. An additional set of results focused on methods or infrastructure for application in assessment. First, the refinement of content modeling can be used for both instructional design and assessment purposes.

Discussion and Design Suggestions for GIFT

In this section, we raise some continuing issues relevant to assessment and not fully resolved by our experience but that would serve as possible design enhancements to the Generalized Intelligent Framework for Tutoring (Goldberg, 2021). Our attempt to create performance assessments that are coherent and from which one could infer a common set of skills was documented by prototypes tried on limited sample sizes related to the size of Sailor training cohorts. Stronger evidence should be obtained by collecting data over multiple cohorts and expanding the number of ratings to which the TAF applies. Second, our goal of developing exchangeable performance assessments within a domain has not been achieved in this project, in part because performance assessments require greater time to administer and trainee time is greatly limited. Nonetheless if performance assessments are conceived as unique objects, without exchangeability and potential use for multiple purposes, they will be of limited use. We would want to document that common forms could be developed so we could monitor performance gains over multiple occasions. That is only possible with a design for a set of performance assessments measuring the same cognitive and content requirements with evidence of their comparability.

In the past, we have used simple combinations to generate the characteristics of a set performance assessments that would cover the high probability situations. For instance, a set of parameters for a set of performance assessments might include choices among four core scenarios, three cognitive demands, two task options, and four response options. These combinations would generate a large number of potential assessments as a pool for creating the desired number of performance assessments. In addition, if one were to add a specific number of processes to monitor, the size of the pool would increase substantially. In order for this activity to be feasible, software supporting the performance environment would need to be designed so its modules could be reused or modestly adapted. Otherwise the programming costs would be too high.

We also believe that FA methodology is useful in a number of ways pertinent to the design of GIFT for training. First, FA can be used alone as a qualitative method in order to confirm that the intended characteristics of items and tasks, as specified in standards, frameworks, and the assessment design, actually can be observed by raters in the items and tasks. A lack of confirmation between intentions and actual observations of tasks and items should result in the revisions of the examination components and, perhaps, of rater training as well. Second, by adding the data component following empirical trials with sufficient numbers, the FA can show interactions among item and task features and characteristics of the examinee population. These findings have implications for fairness inferences about the examination and the interventions. Third, the linking of features with performance data can also reflect specific relationships of exam elements to the degree and type of effects of particular interventions, using pre-post or other multi-occasion designs. A more direct focus on GIFT intervention attributes is a fourth use of FA, where relationships between rated characteristics of the intervention are compared to findings in the data, both overall and on particular features. Such findings suggest how the intervention might be strengthened to achieve desired goals. The FA process can be applied to a single measure or to measures intended to serve as a source of criterion-oriented validity inferences. Data can be explained by the similarities and differences found during the rating process. This approach may as well serve as a more general approach to analyze both the instructional impact of the intervention and the instructional sensitivity of the assessment. If FA can be shown to disambiguate the concepts of instructional sensitivity of the measure and the characteristics of the interventions, these views of validity address the contribution of assessment elements to validity inferences about performance, particularly when the assessment is thought to be useful for program evaluation and to monitor changes in learning. Fifth, FA, when conducted in combination with think-aloud or interview protocols, can uncover task and item attributes not specifically intended by the design. For instance, in one set of studies, we found that particular items required an unplanned high cognitive load for examinees were previously unnoticed.

Rather than a global judgment, FA emphasizes a more analytical approach that relies on data from mixed methods to provide evidence of validity. Our notion was that FA could address functional validity, that is, how attributes of tasks affected performance. The resulting analyses can identify tasks or items that are regarded as difficult or easy, make comparisons following an intervention in order to determine which features contribute to change over occasions, and to recommend to assessment developers features that might be included or excluded based on the data. FA inferences also relate to the quality of instruction; for instance, where instruction is irrelevant or ineffective, one might see no score improvements. One approach that has been taken (Chung & Redman, 2015) involves also coding the intervention with elements that overlap features rated on the assessments. When those data are examined together, it is possible to infer characteristics that lead to desired performance changes. In such a case, these findings are referred back to the instructional or intervention developer for revision.

Recommendations and Future Research

The outcomes and limitations of the TAF work may have implications for ITSs, particularly those that focus on complex learning. Rules for the design of clear performance assessments have application to complex practice environments, simulations, and other ITSs approaches. One important consideration is the degree to which outcome measures used in ITS implementations have validity evidence beyond content or alignment with a set of standards. Second, it is clear that well-designed practice environments, such as those often found in ITSs, can be adapted to serve as stand-alone assessments. Third, reviewing the rules for advancement or adaptation of learning environments may profit from the additional nuance associated with particular TAF elements. Assessment has, since the second half of the 20th century, attempted to employ team or unit environments to determine competency or readiness. The use of simulated characters with interactive capabilities could permit the assessment of multiple roles and conditions. In our own planned work, we are finding ways to compress the time of performance assessments, and apply them to a number of transfer and application situations at various distances from the original training environment. Our plan is to develop and refine a theory of situations that can help codify what is meant by transfer operationally, including the modification of content, setting, role, and goal. If such an analysis can be embodied in a general ontology, it could be applied to a range of content essential for efficient learning. In this age of rapid and unexpected change, assessment must address a wide range of potential situations for which skills and approaches may need to be adapted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). Standards for educational and psychological testing. Washington, DC: Author.

Anderson, L. W., & Krathwohl, D. R. (Eds). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York, NY: Longman.

Baker, E. L. (1997). Model-based performance assessment. Theory Into Practice, 36(4), 247-254.

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. Educational Assessment (Special Issue), 12(3&4), 179-194.

Baker, E. L., & Choi, K. (2019, April5–9). Feature analysis approach: Uses for national and international assessments [Paper presentation]. Annual Meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Baker, E. L., Choi, K., Iseli, M., Wang, J., Kao, J. C., & O'Neil, H. F. (2019). Library of assessments for PS Navy rating and revision/expansion of FC/DC items (Deliverable Item No. 003 to funder). Los Angeles: University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Choi, K., Kao, J. C., & O'Neil, H. F. (2019). Final report (Deliverable Item No. 001 to funder). Los Angeles: University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Choi, K., & O'Neil, H. F. (2021). Final report: Training Assessment Framework (Deliverable Item No. 00003 to funder). Los Angeles: University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Chung, G. (2002, September). Model-based assessment and marine marksmanship example. Presentation to the Office of Naval Research, ONR-NETC Meeting, University of California, Los Angeles.

Baker, E. L., Koenig, A., O'Neil, H. F. (2019). Navy Life Game and demand-based advancement exam final report (Deliverable Item No. 015 to funder). Los Angeles: University of California, Los Angeles, CRESST

Baker, E. L., & Niemi, D. (1996). School and program evaluation. In D. Berliner & R. Calfee (Eds.), Handbook of educational psychology (pp. 926-944). New York: Simon & Schuster Macmillan.

Bloom, B. S. (Ed.). (1956). Taxonomy of Educational Objectives. Vol. 1: Cognitive Domain. New York, NY: McKay.

Choi, K., Baker, E. L., Lee, J. J., Iseli, M. R., Koenig, A. D., & O'Neil, H. F. (2019). Alpha version of advancement exam (Deliverable Item No. 004a to funder). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Baker, E. L., Brill, D. G., Sinha, R., Saadat, F., & Bewley, W. L. (2003). Automated assessment of domain knowledge with online knowledge mapping. Proceedings of the I/ITSEC, 25, 1168–1179.

Chung, G. K. W. K., & Redman, E. H. (2015). Feature analysis framework – final (Deliverable to PBS KIDS). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Gagne, R. M. (1974). Task analysis—Its relation to content analysis. Educational Psychology, 11(1), 11–18.

Gagne, R. M., & Briggs, L. J. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart & Winston.

Goldberg, B. (Ed.) (2021, March). GIFT. Proceedings of the 9th Annual Generalized Intelligent Framework for Tutoring Users Symposium. Orlando, FL: US Army Combat Capabilities Development Command – Soldier Center.

ISO/IEC/IEEE. (2011). Systems and software engineering—Architecture description. Retrieved from https://www.iso.org/standard/50508.html.

Kao, J. C., Choi, K., Rivera, N. M., Madni, A., & Cai, L. (2018, December). Using feature analysis to examine career readiness in high school assessments (CRESST Rep. 858). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koenig, A. D., Baker, E. L., & O'Neil, H. F. (2020). Final report (Deliverable Item No. 003 to funder). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L., Baker, E. L., Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher 20(8) 15-21.

Nye, B., Baker, E. L., & Choi, K. (2018). Final report for force modernization: Training Assessment Framework (Report to funder). University of California, Los Angeles, CRESST.

O'Neil, H. F., & Chung, G. K. W. K. (2011, April 8–12). Use of knowledge mapping in computer-based assessment [Paper presentation]. Annual meeting of the American Educational Research Association, New Orleans, LA, United States.

Questionmark Computing Ltd. (2013). Questionmark Perception Version 5 getting started guide. https://support.questionmark.com/sites/default/files/PDF/v5_getting_started.pdf

Acknowledgements

This chapter was supported in part by funding from the Naval Education Training Command and by contract number OPM2615D0001 (PO 180815-1) from the Office of Personnel Management (OPM) via PowerTrain Inc. to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles and a subcontract to the University of Southern California. The findings and opinions expressed in this chapter are those of the authors and do not necessarily reflect the positions or policies of OPM PowerTrain, CRESST, or NETC. Special thanks to Ms. Joanne Michiuye and Mr. Rob Sitter for their excellent assistance in preparation of this manuscript.

Chapter 5 – Scenario-Based Assessment: Design, Development, and Research

John Sabatini
Institute for Intelligent Systems, University of Memphis

Introduction

This chapter describes scenario-based assessment (SBA) and thoughts about how they can be integrated into the Generalized Intelligent Framework for Tutoring (GIFT) intelligent tutoring system (ITS) designs to measure formative and summative outcomes in the context of "Competency Based Scenario Design and Intelligent Tutoring Systems." Volume 5 of this series "Design Recommendations for Intelligent Tutoring Systems (Volume 5): Assessment Methods", (Sottilare et al., 2017) addressed much of the technical, measurement, and psychometric machinery required to design, implement, and evaluate the psychometric properties of a scenario-based assessment strategy. We do not repeat that content here. Instead, we focus on the concepts and principles of "scenario-based" design.

In searching the aforementioned volume, the term "scenario" appeared some 90 times across eight chapters, but only two or three chapters discuss the specific sense of SBA we describe here (Katz et al., 2017; Mislevy & Yan, 2017; Zapata-Rivera et al., 2017). The sense of SBA used here was derived from an Educational Testing Service Research and Design (R&D) program called "Cognitively Based Assessment of, for, and as Learning" (CBAL) (Bennett, 2011). CBAL was a research initiative that focused on assessment in K-12 settings in English Language Arts (ELA), mathematics, and science (Bennett, 2010, 2011; Bennett & Gitomer, 2009; Sabatini et al., 2011). The CBAL ELA competency and key practice models (and associated learning progressions) were based on syntheses of the literature of reading, writing, thinking, and their connections (e.g., Deane et al., 2011; O'Reilly et al., 2015). A key goal of CBAL was to integrate the research from learning sciences to make assessments meaningful for instruction. Multiple prototype ELA summative and formative assessments were developed and evaluated; thus, building interpretive and validity arguments for their value and utility (e.g., Bennett, 2011).

The subsequent Global, Integrated Scenario-Based Assessment (GISA) was developed with a primary focus on benchmark or summative applications, across kindergarten through 12th grade as part of the Reading for Understanding initiative (Pearson et al., 2020; Sabatini et al., 2018). The GISA framework and design relied on web-based delivery and principles from the learning science and text and discourse, discussed in three framework documents (O'Reilly & Sabatini, 2013; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013). A major goal of the project was to design SBAs to be feasible and practical, while maintaining adequate psychometric properties (Sabatini et al., 2020). While Reading for Understanding focused entirely on reading literacy, we generalize the SBA principles for assessment in other domains.

SBAs typically include a range of principles and techniques that distinguish them from other types of assessments: (1) they provide an authentic purpose for reading, (2) they place reading in context for completing a set of interrelated activities that may move from more guided to independent performance, (3) items tend to require the integration and evaluation of a wide range of diverse sources and, (4) in many cases, items provide scaffolds (e.g., a graphic organizer for an analysis of text structures) and guidelines (e.g., tips for summary writing) to help better understand and model the target performance in the assessment (O'Reilly & Sabatini, 2013). SBAs also include items that model the social aspects of literacy and learning, such as engaging with peers or a teacher to clarify understanding in reading, reviewing and evaluating peer writing. Using these principles, SBAs may broaden the range of interactions, perspectives, and information a test taker is exposed to on a topic. Ultimately, the key aims of SBAs are to measure 21st century competencies, while simultaneously supporting skill development and, in the cases that our team developed, instructional usefulness.

Design and Development Principles for Generalized SBA design

O'Reilly and Sabatini (2013) described key features of SBA and performance moderators. Key elements include the following.

- Establish a *purpose* or goal (or mission) for the participant to achieve over the course of the assessment. This purpose or assessment narrative in turn is used to guide the participant in planning and deciding what is relevant to focus their attention and performance upon, setting situational standards of coherence (van den Broek et al., 2011).

- Provide a *coherent collection of materials* useful or required to achieve the goals of the assessment narrative. In a typical comprehension assessment, to serve the goal of sampling a range of text types and genres, the passages a student encounters are randomly chosen, administered in a random sequence, and have no interconnections to each other. However, in most life scenarios, the materials one encounters or chooses are relevant to the purpose and goals. The individual builds, shares, fact checks, evaluates, and troubleshoots as they move through a sequence of events towards achieving their goal. Consequently, the design of the assessment builds on the logic of such a sequence. This is not to say that random or irrelevant materials will not be encountered, but part of competency is to discard or minimize one's distraction when encountering irrelevant content. As we focused on reading literacy derived from school-based contexts, scenarios we constructed involved students forming study groups, planning presentations in public forums, or producing websites to inform others about a critical issue (e.g., Sabatini et al., 2020).

- In formative (and sometimes summative) use cases, *triangulate strengths and weaknesses* in performance. That is, in an ITS, one is rarely only interested in evidence of proficiency, but also rather in diagnosing competencies where the learner might need to learn or practice further. We used multiple methods, but typically we would start with a full performance task, then break it down into subcomponents in subsequent assessment tasks (e.g., Sabatini et al., 2014). If the individual performed adequately on the full performance task, we would expect this level of performance would be validated by their also performing well on the subcomponents. On the other hand, if there were breakdowns in the full performance, the subsequent tasks helped us to gather evidence regarding which subcomponent competencies may be weak and negatively impacting the integrated performance. One could in this manner identify individuals that were adequate in all the subcomponents, just not yet capable of integrating them, and vice versa.

- Promote *collaboration* including distributed and collective understanding. This principle partially stemmed from our recognition of point of view and perspective taking as a critical dimension of higher order reasoning and performance in our reading literacy proficiency construct definition. As noted, our context was literacy and with billions of websites of information and misinformation, often contested and contradictory, the skills required to first understand sources, their credibility, and intentions of authors/publishers is often key to understanding (Braasch et al., 2018; Magliano et al., 2018; Rouet et al., 2017; Sabatini et al., 2018).

- Include and explore learner *relevant prior or background knowledge* as it relates to performance (O’Reilly et al., 2019a, b). We evaluated student knowledge using multiple techniques. One especially quick and productive method was to use Natural Language Processing (NLP) techniques to identify topically related vocabulary, then have learners perform a simple sorting task as to whether the terms were related or not to a specific topic. The terms often extended beyond those used in the assessment itself and ranged from simple, foundational terms to rarer, technical terminology. We consistently found low to moderate correlations with outcome score performance, and could use this information to better interpret student outcome scores. The technique also proved to be highly reliable without taking much extra time and effort from the examinee (O’Reilly et al., 2019b). By giving learners the option “I don’t know”, we also were able to probe their metaknowledge, and found a consistent relationship between performance scores and efficient use of the “I don’t know” option, suggesting that student’s awareness of their own knowledge (or gaps) was conducive to better learning of subsequent content (O’Reilly et al., 2019a). We have recently extended this research into non-academic domains (Wang et al., 2021).

- As feasible, promote *interest, motivation, and engagement*. Given that the alternative reading comprehension testing procedure at the time we were conducting our research was traditional, multiple choice standardized tests, the bar was set very low for applying this principle at the outset of our research program. Nonetheless, we developed multiple techniques that served more nuanced strategies that aligned with our framework and constructs. We attempted to balance several competing sources of construct irrelevant variance. Because traditional passage comprehension formats tend to produce test anxiety in some groups, we transformed the look and feel of item delivery, though often retaining a multiple-choice item type. More centrally, we engaged students with the use of the simulated agents (teachers, peers) that were co-participants in the assessment narrative (Graesser, 2016; Graesser et al., 2014; Graesser, Forsyth et al., 2017; So et al., 2015). The simulated agents we used in the SBAs were static, that is, pictures of students or teachers, who spoke via text chat-style bubbles. They were not adaptive – every student saw the same agent communications regardless of how they responded to questions. The static agents served multiple, other specific purposes, such as directing attention towards relevant content, modeling expected responses, scaffolding steps not yet introduced or known to be beyond the participant’s level, as well as producing common errors that the participant might be asked to help correct. To the extent we were successful in allowing students to suspend their sense of disbelief, the approach helped them maintain the performative aspects such that students gave us their best efforts for themselves and their peers. At the same time, this gave many of the scenarios the feel of a learning vs. assessment context, hopefully diffusing some of the anxiety typical of standardized test settings (Sena et al., 2007).

A side principle to consider regards awareness and planning for constraints of the test performance settings. Our design team reasoned through aspects of feasibility of implementation and scoring, scalability, and maintenance of psychometric rigor, first in the research context and ultimately for application in schools. For example, we chose to make the entire system web-administered, which conferred multiple benefits to our project and ultimate scalability goals like remote recruitment, ease of administration, data collection, scoring, the implementation of complex, randomized designs within and across schools, and a natural environment for using digital sources. The parallel here regards GIFT architecture and the specifics of the SBA targets one might design. For example, scalability may or may not be an issue if the target skill set is designed for a small subpopulation. Another constraint was to limit the test length to around 45-50 minutes, a typical classroom period in the United States. This limited duration made it easier for us to recruit schools and collect student data.

We also limited the use of written constructed response (CR) items to questions we believed could be scored using automated processes. We primarily focused on paraphrase, summary, and some short-answer explanations. These item types are important cognitive reading strategies (McNamara & Magliano, 2009), hence, worth teaching to by instructors, as well as rich sources of comprehension evidence. This approach served several aims simultaneously: 1) less student time on individual CR items (which are often also effort intensive) and 2) amenable to automated scoring. Of course, advances in NLP (e.g., Rus et al., 2017), data mining, and machine learning are improving at a rapid rate. The issue remains whether the response type itself (in this case writing) is authentic to the task setting. These examples are meant to highlight the more general principle of thinking about system or practical constraints during the design process, to be better able to determine their impact on validity of inferences later.

I and my colleagues (Mislevy & Sabatini, 2012; O'Reilly et al., 2014; Sabatini et al., 2020) consistently used design methodologies such as evidence-centered design (ECD) to enact the "argument" underlying our assessments (Mislevy, 2007, 2018; Mislevy & Yan, 2017). In straightforward terms, ECD consists of three major components. First, it forefronts defining what learner/trainee variables (knowledge, skills, strategies, dispositions) are essential, as well as how they interact with each other, i.e., a student model. Second, it requires identifying tasks that would require that the learner use/display those skills. Ideally, these tasks will be close in proximity and authenticity to the actual use cases where they would be applied. Finally, there must be specific identification of the evidence that can be used to evaluate the level or proficiency of the learner skills. Evidence comes in many, many forms. It can be as simple as correct/incorrect answers to multiple choice questions to complex, multidimensional judgements based on performance during the scenario. I agree with Zapata-Rivera et al. (2017) that formalizing what counts as evidence to support claims of student model proficiency is essential. It is what transforms an ITS experience into an assessment. That is, in a well-designed ITS, the evidence to support claims of learning or proficiency are likely already collected in the responses to tasks. Consequently, it is not too difficult to imagine merging the learning and assessment functions in the design of the ITS. Many measurement/statistical models are available for creating scores from embedded items that possess properties that support valid inference for measuring growth, stability of learning gains, or achievement (Sabatini et al., 2019).

Critical issues for the future of SBA design in ITSs

Embedded scenario-based assessments. The key distinction between a learning versus assessment environment is the inferences one wishes to make from the data. In an ITS, the designers typically think of process or performance data as part of a feedback loop for optimizing adaptively the instructional/learning experience. It is truly not much of a leap (if there is any distance at all) to the concepts of formative assessment or progress monitoring. Formative assessment is also sometimes called learning-based assessment and entails gathering evidence that would support a teaching action that enhances learning (Black & Wiliam, 1998; Heritage, 2008). Some express formative assessment as moment to moment, in that the immediacy of a student benefiting from a prior learning episode is often what is assessed. Progress monitoring is typically conducted after a longer period of time learning and targets a larger 'chunk size' of learning. Many progress monitoring tools use as a metric their predictiveness of whether the individual is on a trajectory of achieving expected outcomes; thus, they are calibrated with correlations or probabilities of passing outcome tests (O'Reilly et al., 2012; Santi & Vaughn, 2007). For this reason, tasks that resemble outcome achievement tasks are typically used.

The logic of these approaches can be applied to SBAs (Mislevy & Sabatini, 2012; O'Reilly et al., 2012), that are embedded within ITSs (Shute & Kim, 2014). Two primary design considerations are paramount. First, designing progressively more authentic SBA task sets across the course of learning. That is, while the learning content itself may be designed to optimize knowledge and skill development, the scenario tasks should be focused on the application of those skills in a scenario that more closely matches authentic use of the skills in context. Second, the desire that the learners apply their skill in real-time performance needs to be communicated and signaled to the learners. Otherwise, they may behave as if they are in learning or practice mode and not give their best performance. This may be especially important when game-based environments have been used extensively for learning/training, and thus the internal mindset of the participant may still be in that modality. Of course, this should always be an element considered when examining validation evidence of a performance in a simulation or assessment versus real life application.

Collaborative/team-based assessments. Increasingly, we are concerned with assessing the performance of a group or team (Graesser, Dowell et al., 2017; Olivieri et al., 2019). Historically, assessment techniques have evolved to measure individual proficiency/competency, not group level performance. Strong work is emerging in building constructs and frameworks to examine collaborative processes, such as collaborative problem solving (Graesser et al., 2018) or collaborative, critical discussions (Johnson, 2015). However, the breadth of this research is still limited in examples and domains.

One way to move forward with addressing the assessment problem is to consider three overlapping, but distinct measurement aims, frameworks, constructs:

1. How proficient was the team/group in their scenario performance?
2. How well did each individual perform in the team/group based scenario?
3. How proficient is the individual in supporting the team/group in their performance?

In the first question, the unit of analysis is the group or team performance, not necessarily how well any individual participated or contributed to the team performance. The second focuses on the individual's contribution to that performance, though not necessarily on whether the team did well overall. And the third focuses most closely on the notion of collaboration or teamwork, in that the individual's own contribution may be seen via how well they helped others and the team to perform better.

Each approach has its design and measurement challenges, and it is unclear which might be most valuable in the context of GIFT, ITSs, or particular training/learning goals. For example, if the goal is simply to form the most effective, efficient teams, then the first approach might be the most useful in forming strong teams, under the assumption one can then keep those teams together as a unit in actual settings. On the other hand, if the goal is to train a set of individuals who are strong in enhancing the performance of any team they join, then the framework surrounding the third question may be optimal.

It would be interesting to try to measure all three and examine the intercorrelations or predictive value of one approach over another. It may turn out that one of these approaches is a better assessment predictor across the three, or pairwise. Another analysis angle would be to consider frameworks that optimize selection and development around roles/responsibilities within collaborative teams, such as leadership, building/maintaining moral, fostering effective communication, technical skills, and so forth.

Applying lessons learned from SBA to GIFT

The application of SBA to GIFT largely depends on the instructional design and learning objectives of the GIFT training or course. For example, if the instructional design is reinforcement learning at a discrete subskill level, then there is likely a large gap between the acquisition of the knowledge and skills, and the complex application of those skills in real world settings. In such, well designed SBAs help reveal whether the transfer of skills to applied settings is naturally occurring, or whether the instruction needs to be enhanced (at least for some trainees) to ensure transfer. For instructional designs that progressively build and integrate subcomponents of skills into scenario-based instruction, SBAs may be redundant with the scenarios that occur late in the course itself. Or the SBAs can be designed as near or far transfer or challenge applications to evaluate proficiency or identify relative strengths and weaknesses. When used before a course, SBAs can help the trainee by previewing and modeling the expected skill learning they should possess by course end.

In the course of an intelligent tutoring course, the focus of the trainee is on learning, with activities including instructional content delivery, problem solving tasks, and adaptive feedback targeting the building of knowledge or skills. This is not necessarily the same as using the skills to achieve a goal. In an SBA, the focus should first be on the performance or application of those skills in achieving a realistic goal. Secondarily, it may be on identifying strengths and weaknesses in performing the subgoals (or more generally, subcomponent competencies).

To capture the likelihood of the candidate performing adequately in a complex scenario or environment, the essential elements of that environment should be represented in the SBA, with some aspect of real-time problem solving incorporated as feasible. In the case of reading, we described this as a coherent collection of materials and aligned tasks. Also, if the expected performance of the skill would be time-based in the world, it would be best if it is also time-based in the assessment. This practice contrasts the traditional assessment concept of sampling randomly, but independently across a domain; or relaxing real time solutions in favor of providing immediate feedback to help in learning content. In the SBA, the concept of purpose-driven performance should drive the selection of materials in a scenario and the sequence and timing of the events. Again, scenario-based learning may already simulate these conditions in the ITS. When that is not optimal, using SBAs as the performance ‘practice’ and evaluation tool may be a recommended use of SBA with GIFT.

Because an assessment collapses events in time and space, some reduction in the complexity of the environment is likely necessary. However, this can also result in gaps in the coherence of the experience in comparison to authentic settings. It may also create ambiguity or confusion for participants in what is expected or how to respond. In our SBAs, we used collaborative agents to scaffold, model, and provide guiding information, as needed. In most cases, it is realistic to expect social context and support to be available – and for some it reduces the anxiety accompanying traditional test formats. These agents may differ from those in the ITS in that they must be designed to not interfere with the measurement goals – collecting evidence of participant proficiencies on specific tasks.

Finally, we find it useful to explore a learner’s relevant background knowledge to assessment performance, both construct relevant and irrelevant. If an assessment is being used as an outcome measure after training or completing an ITS course, then the relevant learner records collected during the ITS may serve this function. Strong prior knowledge or skills in any area reduce cognitive load during the assessment and should enhance the quality and sophistication of responses to complex tasks. Low relevant knowledge may increase cognitive load or require compensatory processes to complete tasks. In any case, interpretation of SBA scores can be enhanced by knowing what types and level of prior knowledge was available to the learner.

Summary

In this chapter, I reviewed SBA and how it might be designed in the context of ITSs. I presented design principles derived from the design of global, integrated, SBAs developed and implemented as part of the Reading for Understanding initiative. I presented evidence-centered design as a core technique for guiding the process, as well as the importance of considering the constraints, both in the research/development phase, and in final operational use. Finally, I briefly discussed a couple critical future issues in ITS scenario-based design: embedded or stealth assessment and collaborative/team-based assessment.

References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. Measurement, 8(2–3), 70–91.

Bennett, R. E. (2011). CBAL: Results from piloting innovative K–12 assessments. ETS Research Report Series, 2011(1), i–39.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In Educational assessment in the 21st century (pp. 43–61). Springer.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan, v80 n2 p139-44 Oct 1998.

Braasch, J. L., Braaten, I., & McCrudden, M. T. (2018). Handbook of multiple source use. Routledge.

Deane, P., Sabatini, J., & O'Reilly, T. (2011). English language arts literacy framework. Education Testing Service.

Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. International Journal of Artificial Intelligence in Education, 26(1), 124–132. https://doi.org/10.1007/s40593-015-0086-4

Graesser, A. C., Dowell, N., & Clewley, D. (2017). Assessing collaborative problem solving through conversational agents. In Innovative assessment of collaboration (pp. 65–80). Springer.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. Psychological Science in the Public Interest, 19(2), 59–92.

Graesser, A., Forsyth, C., & Lehman, B. (2017). Two Heads May be Better than One: Learning from Computer Agents in Conversational Trialogues. Teachers College Record, 119(3), 1–20.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. Current Directions in Psychological Science, 23(5), 374–380.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Paper prepared for the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO).

Johnson, D. W. (2015). Constructive controversy: Theory, research, practice. Cambridge University Press.

Katz, I. R., LaMar, M. M., Spain, R., Zapata-Rivera, J. D., Baird, J.-A., & Greiff, S. (2017). Validity CHAPTER 18–Issues and Concerns for Technology-based Performance Assessments. Design Recommendations for Intelligent Tutoring System-Volume 5: Assessment Methods, 5, 209.

Magliano, J. P., McCrudden, M. T., Rouet, J.-F., & Sabatini, J. (2018). The modern reader: Should changes to how we read affect research and theory? In The Routledge handbook of discourse processes, 2nd ed. (pp. 343–361). Routledge/Taylor & Francis Group.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. Psychology of Learning and Motivation, 51, 297–384.

Mislevy, R.J. (2007). Validity by design. Educational Researcher, 36, 463–469.

Mislevy, Robert J. (2018). Sociocognitive foundations of educational measurement. Routledge.

Mislevy, R. J., & Sabatini, J. (2012). How research on reading and research on assessment are transforming reading assessment (or if they aren't, how they ought to). In J. P. Sabatini, E. R. Albro & T. O'Reilly (Eds.), Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences. Lanham, MD: Rowman and Littlefield.

Mislevy, Robert J., & Yan, D. (2017). Evidence-Centered CHAPTER 10–Assessment Design and Probability-Based Inference to Support the Generalized Intelligent Framework forTutoring (GIFT). Design Recommendations for Intelligent Tutoring System-Volume 5: Assessment Methods, 5, 101.

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. International Journal of Testing, 19(3), 270–300.

O'Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under an RTI framework. Journal of Reading Psychology, 33, 162–189.

O'Reilly, TT, Deane, P., & Sabatini, J. (2015). Building and Sharing Knowledge Key Practice: What Do You Know, What Don't You Know, What Did You Learn? Research Report ETS RR–15-24. Princeton, NJ: Educational Testing Service.

O'Reilly, T., & Sabatini, J. (2013). Reading for understanding: How performance moderators and scenarios impact assessment design. Princeton, NJ: Educational Testing Service.

O'Reilly, Tenaha, Sabatini, J., & Wang, Z. (2019a). What You Don't Know Won't Hurt You, Unless You Don't Know You're Wrong. Reading Psychology, 40(7), 638–677. https://doi.org/10.1080/02702711.2019.1658668

O'Reilly, Tenaha, Wang, Z., & Sabatini, J. (2019b). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. Psychological Science, 30(9), 1344–1351.

O'Reilly, Tenaha, Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. Educational Psychology Review, 26(3), 403–424.

Pearson, P. D., Palincsar, A. S., Biancarosa, G., & Berman, A. I. (2020). Reaping the Rewards of the Reading for Understanding Initiative. National Academy of Education.

Rouet, J.-F., Britt, M. A., & Durik, A. M. (2017). RESOLV: Readers' representation of reading contexts and tasks. Educational Psychologist, 52(3), 200–215.

Rus, V., Olney, A. M., Foltz, P. W., & Hu, X. (2017). –Automated Assessment of Learner-Generated Natural Language Responses. Design Recommendations for Intelligent Tutoring Systems, 155.

Sabatini, J., Bennett, R. E., & Deane, P. (2011). Four years of cognitively based assessment of, for, and as learning (CBAL): Learning about through course assessment (TCA). Invitational research symposium on through-course assessments, Atlanta, GA.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, L. Cutting, & P. McCardle (Eds.), Unraveling the behavioral, neurobiological, and genetic components of reading comprehension (pp. 100-111). Baltimore, MD: Brookes Publishing, Inc.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design (RR-13-30). Princeton, NJ: Educational Testing Service.

Sabatini, J., O'Reilly, T., Halderman, L. & Bruce, K. (2014). Broadening the Scope of Reading Comprehension using Scenario-based Assessments: Preliminary Findings and Challenges. L'Année psychologique/Topics in Cognitive Psychology, 114, 693-723.

Sabatini, J., O'Reilly, T., Wang, Z., & Dreier, K. (2018). Scenario-based assessment of multiple source use. In J. L. G. Braasch, I. Braten, & M. T. McCrudden (Eds.), The handbook of multiple source use (pp. 447–465). Taylor & Francis/Routledge.

Sabatini, J., O'Reilly, T., Weeks, J. & Wang, Z. (2020). Engineering a 21st Century Reading Comprehension Assessment System Utilizing Scenario-based Assessment Techniques. International Journal of Testing, 1-23. DOI: 10.1080/15305058.2018.1551224

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Chao, & Steinberg, J. (2019). SARA reading components tests, RISE forms: Test design and technical adequacy, 3rd Edition (ETS RR-15-32). Princeton, NJ: ETS. doi:10.1002/ets2.12076

Santi, K., & Vaughn, S. (2007). Progress monitoring: An integral part of instruction. Reading and Writing, 20(6), 535–537.

Sena, J. D. W., Lowe, P. A., & Lee, S. W. (2007). Significant Predictors of Test Anxiety Among Students With and Without Learning Disabilities. Journal of Learning Disabilities, 40(4), 360–376.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In Handbook of research on educational communications and technology (pp. 311–321). Springer.

So, Y., Zapata-Rivera, D., Cho, Y., Luce, C., & Battistini, L. (2015). Using trialogues to measure English language skills. Journal of Educational Technology & Society, 18(2), 21–32.

Sottilare, R., Graesser, A., Hu, X., & Goodwin, G. (2017). Design Recommendations for Intelligent Tutoring System-Volume 5: Assessment Methods (Vol. 5). US Army Research Laboratory.

Van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., & White, M. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In M. McCrudden, J. P. Magliano, & G. Schraw (Eds.), Text relevance and learning from text (pp. 123–140). Information Age Publishing.

Wang, Z., O'Reilly, T., Sabatini, J., McCarthy, K. S., & McNamara, D. S. (2021). A tale of two tests: The role of topic and general academic knowledge in traditional versus contemporary scenario-based reading. Learning and Instruction, 73, 101462.

Zapata-Rivera, D., Brawner, K., Jackson, G. T., & Katz, I. R. (2017). Reusing Evidence in Assessment and Intelligent Tutors. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), Design Recommendations for Intelligent Tutoring Systems (Vol. 5, pp. 125–136). US Army Research Laboratory.

Chapter 6 – Implementing Soft Skills Training in GIFT

Patrick C. Kyllonen[1], Arthur C. Graesser[2], Sara B. Haviland[1], Steven B. Robbins[1], & Kevin Williams[1]
[1]Educational Testing Service, Princeton, NJ; [2]University of Memphis

Introduction

The purpose of this chapter is to explore the steps it would take and the advantages it would afford to implement soft skills training in the Generalized Intelligent Framework for Tutoring (GIFT) system. Soft skills are not always thought of as competencies in the same way that hard skills are, and we address that by discussing soft skills and soft skills training generally, contrasting soft skills with hard skills. We outline the details of a soft skills training system developed at Educational Testing Service (ETS). The ETS system is not a commercial product but is a research effort and is currently being used in various settings including as a component of a community college course designed to develop students' career skills. Then we briefly discuss learning management systems (LMSs), one of which (Blackboard Learning) currently hosts ETS's soft skills training. We discuss GIFT and its capabilities and features and discuss how GIFT might enable certain capabilities in tutoring soft skills that are not currently available, at least commercially. We also speculate on ways to imagine how soft skills training might evolve to be more comparable to hard skills training with respect to GIFT concepts like the pedagogical and domain modules. This chapter is not intended to serve as a step-by-step how-to guide for implementing soft skills training, but instead is designed to elicit discussion about improving soft skills training from concepts and lessons learned in hard skills training.

Soft Skills

The term *soft skills* refers to social, emotional, and self-management skills associated with success in school and work. Soft skills are alternatively referred to as work styles, behavioral skills, interpersonal and intrapersonal skills, 21st century skills, noncognitive skills, character skills, social and emotional skills, intangibles, or hard-to-measure skills. Soft skills may be contrasted with hard skills, such as verbal and quantitative ability or achievement, prose and quantitative literacy, and technical skills, such as information technology, mathematical skills, language skills, or the ability to use various tools. This is an important distinction because traditionally, hard skills are explicitly taught in schools or in training programs and evaluated using standardized tests of one's knowledge or mastery of those skills, whereas that has not conventionally been the case with soft skills. In fact, soft skills are sometimes referred to as those that are not explicitly taught (nor assessed), the ones "they don't teach you at Harvard Business school" (McCormack, 1984).

Despite a widespread belief that soft skills are fixed and not malleable, there is an established and growing literature on personality change (Roberts et al., 2017; Soto et al., 2011), including studies on the malleability of social and emotional skills in K-12 (Corcoran et al., 2018; Duncan et al., 2017; Durlak et al, 2011; Mahoney et al., 2018) and higher education (National Academies of Sciences, Engineering, and Medicine, 2017). Soft skills training, such as for leadership and social skills, is routinely conducted in workplace settings (Arthur et al., 2003; Martin-Raugh, Williams et al., 2020).

There are many soft-skills constructs (National Research Council, 2012), but probably the most prominent framework or taxonomy is the Big Five model of personality (John & Srivastava, 1999), or a related scheme that expands the number of key dimensions to something higher than five (Condon, 2018; Condon et al., 2021; Drasgow et al., 2012; Saucier & Iurino, 2020). What these systems have in common is that the basis for identifying dimensions is a factor analysis of responses to self- (or other-) descriptive statements presented typically in a Likert scale format (e.g., strongly disagree to strongly agree) or sometimes using pairwise statement comparisons or other statement ranking methods ("select or rank the statements for how well they describe you"). Given the prominence of the Big Five framework, the demonstrated susceptibility of Big Five factors to developmental growth (Roberts et al., 2006), and its acknowledged importance in predicting outcomes across a variety of school (Poropat, 2009) and workforce settings (Salgado & Tauriz, 2014; Salgado et al., 2015), it is somewhat surprising that there are few demonstrations of successful interventions targeting Big 5 factors. In their comprehensive review of the malleability of noncognitive constructs, in which they identified 39 meta-analyses, Martin-Raugh, Williams et al. (2020) found only a couple that arguably targeted Big Five type factors (Roberts et al., 2017; Vanhove et al., 2016). This paucity may be due to skepticism about the potential efficacy of personality interventions (and the high cost of evaluating such interventions), the lack of evidence for the causal (as opposed to correlational) role of personality on outcomes (Mottus et al., 2020), or to the opportunity cost of intervening on general factors rather than more specific targets such as communication skills, leadership, and emotions, which Martin-Raugh, Williams et al. (2020) found were more common. Nevertheless, it would seem to be important to explore the efficacy of soft skills training targeting general soft skills like the Big 5, the ones that have been shown in meta-analyses to be the most important in the sense of predicting a wide variety of outcomes (Poropat, 2009; Salgado & Tauriz, 2014; Salgado et al., 2015).[1]

ETS Soft Skills Training History

ETS conducted research on soft skills training beginning with the development of the Work Readiness Strength Assessment and Training System (WRSATS) (Shore et al., 2016), a soft-skills training system targeting community colleges, programs in career pathways, workforce and workplace training, and adult secondary education (ACE). The system comprised three components, (a) a work readiness strengths assessment, which measured 13 personality factors from which six competency composites were derived (initiative and perseverance, responsibility, flexibility and resilience, teamwork and citizenship, customer service orientation, and problem solving and ingenuity), (b) individual work readiness strength profiles, or score reports, which provided feedback to individuals on those competencies, and (c) the work readiness training system, a set of eight, approximately one-hour learning modules, one for each competency and an introduction and conclusion.

Work Readiness Training: Instructional Modules. The training system was designed to address several issues and best practices in learning, particularly from the adult learner perspective (Knowles et al., 2015). Specifically, modules were designed to make clear to learners the importance and real-life (practical) significance and relevance of the material and to encourage learners to relate the material to their life experiences in concrete and tangible ways for immediate application. Modules were designed to address both extrinsic (e.g., promotion, salary) and intrinsic (self-esteem) incentives. Thus, each module began with presenting a set of concrete objectives, followed by (a) a *warmup activity* in which students discussed the importance of the competency in a workplace setting, (b) a *presentation of new materials and concepts* (including interactive media), (c) a *practice session* involving those concepts, based on both group and

[1] Personality is another term for soft skills, although there are schools of thought that attempt to differentiate them, for example, by considering personality an umbrella term, by distinguishing skills from thoughts, beliefs, and behaviors, or by focusing on skills versus personality assessment strategy differences. These discussions go beyond the scope of this chapter. Here we use the terms mostly interchangeably.

individual work, (d) an *evaluation* (self- and peer-) and test on knowledge of the concepts taught, and (e) an *application*, a reflection activity based on a self-evaluation.

Work Readiness Strengths Competency Assessment and Score Report. The attribute and competency assessment is a multi-dimensional forced-choice self-assessment method for evaluating students (Naemi et al., 2014). The assessment measured 13 factors (diligence, dependability, organization, self-discipline, assertiveness, friendliness, collaboration, generosity, stability, optimism, creativity, intellectual orientation, and inquisitiveness) through the presentation of a series of statement pairs requiring respondents to choose the statement "most like you." An item response theory approach developed by Stark et al. (2005) was used for scoring responses to the statement pairs to infer levels on the 13 factors (see Naemi et al., 2014). The approach involved the initial calibration of Likert responses using the generalized graded unfolding model. Then Bayesian modal scoring of pair statement preferences was based on a multi-unidimensional pairwise preference model. This approach reduces response style biases and social desirability effects found in Likert ratings because statements in pairs are matched on social desirability.

Training the Trainers. The training system (WRSATS) was designed to be administered by human instructors, and so there was a plan for training the trainers, which involved an introductory webinar and two two-hour workshops, to introduce trainers to the underlying research and the training program and its features. The training session also involved mock lessons and roleplays. The training was supplemented by an instructional guide and training videos.

Pilot Testing. This system was tested with 19 instructors and 300 students (from adult education and community colleges) and lessons learned were incorporated into revisions.

Workforce Assessment for Development. The WRATS system was upgraded for commercial use as part of ETS's Workforce suite (which included Workforce Assessment for Job Fit and Workforce Assessment for Development). That program was discontinued in 2020, but the ETS soft skills development system continued to be maintained, features added, and several implementations are underway or are being planned.

Learning Management Systems (LMSs)

ETS's various soft skills training systems, as described in the previous section, have been implemented on different platforms. The current system, ETS Essential Skills for Success Training (ESST) resides on a learning management system (LMS) platform, Blackboard Learn, hosted by a community college. In its current iteration it is designed to be introduced at two moments in the college pathway: a more general set of skills at entry (e.g., integrated into a first-year experience course) and a more advanced set of skills later as part of experiential learning (e.g., internships or clinicals).

LMSs are platforms designed to be used by schools (or organizations), teachers (or trainers), and students (or employees) for elearning with components ranging from class management to course content and evaluation. Two of the systems most widely used in the U.S.are Blackboard Learn and Instructure's Canvas (and Bridge systems for the corporate market). Blackboard Learn is widely used in colleges and universities in the U.S. It comprises a set of capabilities such as course management (for posting due dates, syllabi, grades, maintaining student profiles, enabling announcements), student monitoring (posting and enabling students to take tests and quizzes, posting and completing assignments), course content (articles, assignments, lessons or learning modules, media library for videos), peer-to-peer and student-teacher interactions (discussion threads, email, real-time chats). Other systems such as Canvas, which is primarily used in business contexts, and Moodle, which is open-source, provide generally similar features and capabilities, and in some cases can work together. LMSs do not have intelligent tutoring capabilities such as what is offered through the Generalized Intelligent Framework for Tutoring (GIFT) platform.

The Generalized Intelligent Framework for Tutoring (GIFT) as a Platform for Soft Skills Training

The Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare et al., 2012; Sottilare et al., 2017) is an “empirically based, service-oriented framework of tools, methods and standards to make it easier to author computer-based tutoring systems (CBTS), manage instruction and assess the effects of CBTS, components and methodologies” (ARL, 2015a, Description para.). This description places GIFT into the category of LMSs. However, GIFT’s emphasis on tutoring, and methodologies for evaluating students and learning offer significant additional capabilities beyond those associated with typical LMSs (Graesser et al., 2016; Sottilare et al., 2018).

Intelligent tutors or intelligent tutoring systems (ITSs) are adaptive instructional systems that provide learners with curricular content tailored to their current knowledge and skill level. Since Barr et al.’s (1976) Basic Instructional Program (BIP) and Anderson et al.’s (1985) ITSs based on the Adaptive Control of Thought (ACT) theory, ITSs have generally comprised domain (expert) models, student models, and a pedagogical model. A domain model is a representation of the curriculum, that is, the curricular elements, or the content and lessons to be covered. A student model is a representation of what students know at a given time with respect to the curriculum; for example, which concepts have been mastered or not. A pedagogical model is a set of rules for presenting curricular elements or problems to students based on their current knowledge state; that is, it is responsible for the adaptivity of the instruction.

As a general framework therefore, GIFT includes (a) domain, (b) learner, and (c) pedagogical (or tutor), modules. It also adds a fourth, (d) a sensor module to accommodate additional technology enabled sensing capabilities, to measure emotional changes. There is also an interface component, which displays feedback and system states to the student or teacher. It also involves sensing input in various modalities. There are also auxiliary components such as a Gateway module for interoperability, and an LMS module, which is part of the learner module, to track a user’s training history.[2] All these modules are defined in the GIFT FAQ Glossary.

This background was designed to provide an overview of the ETS soft skills training system and the GIFT system. We now describe a current version of the ETS soft skills training system and GIFT functionality in more depth to determine what it would take to implement soft skills training in GIFT and what GIFT might have to offer for doing so.

ETS’s Soft Skills Training System

The ETS Soft Skills training system includes articles, videos, closed- and open-ended assessments, both formative and summative in nature, opportunities for interacting and role-playing with classmates, and rubrics for evaluating interactions. The content of the training is specified in instructional objectives, and concerns declarative knowledge regarding six distinctive competencies (*Initiative and Perseverance*, *Responsibility*, *Flexibility and Resilience*, *Teamwork and Citizenship*, *Customer Service Orientation*, *Problem Solving and Ingenuity*), explained with respect to more basic behavioral dimensions (also referred to here as subconstructs), knowledge of how to recognize the relevance of those competencies in daily interaction contexts, and skill in exercising those competencies in specific interaction situations such as interviews. The soft skills competencies are broad and habitual in nature, and therefore difficult to modify without knowledge of the competency and practice in exercising it in situations. The ETS Soft Skills

[2] GIFT also includes three components to facilitate research which are (e) the Survey Author Tool, (f) user tracking in MySQL, and (g) the capability for writing compiled data to a .csv file for analysis.

training system is designed to provide such knowledge and practice, although it is primarily designed to be used in an online-classroom blended-learning environment, which includes a face-to-face classroom experience, supervised by an instructor, with exercises performed with other students. Educational Testing Service's (ETS, 2018a) *WorkFORCE® Program for Career Development: Using Your Results* is a set of materials designed for students to interpret their score reports and prepare for the training program. ETS's (2018b) *WorkFORCE® Program for Career Development: Instructional Guide* is designed for instructors. Haviland et al. (2021) presents results from case studies involving the administration of soft skills training in two community colleges. Here we provide an overview of some of the key features of the system.

There are eight modules (*Introduction*; the six competency modules listed above; and *Completion*). Each module includes a standard set of components, as follows:

Pre-work articles provide background on soft skills and cover the key points in the lessons. For example, there might be an article on time management (for *Initiative and Perseverance*), stress management (for *Flexibility and Resilience*) or workforce diversity (for *Teamwork and Citizenship*).

Guiding questions and objectives activates learners' personal experiences, provides cues, questions, and advance organizers, and sets goals and learning objectives for the course. For example, an objective could be to learn to spot opportunities for managing one's time more efficiently or to become aware of strategies for overcoming lack of motivation for completing work (*Initiative and Perseverance*), or learning to work well with people whose opinions, values, and backgrounds are different from one's own (*Teamwork and Citizenship*). Learning and establishing these objectives can be accomplished in small groups reporting back to the whole class, or in a brainstorming session to generate strategies for working towards the objectives.

Covering the basics introduces new concepts, defines the behavioral competency, and presents a video illustrating the competency. For example, *Initiative and Perseverance* is defined in the system as work context behaviors associated with approaching job duties, acting as a self-starter, and completing tasks efficiently. *Responsibility* is defined as conducting oneself with accountability and excellence; working in a focused, organized manner, following safety and other regulatory rules, and demonstrating appropriate workplace behavior. Video vignettes illustrate these concepts in workplace settings, and learners are instructed to take notes while viewing. They are given multiple-choice knowledge checks to check comprehension, which are automatically scored. The vignettes may also present questions about the competency for discussion. One strategy is to have students in the class take positions and argue on one side or the other of the discussion topics.

Workplace scenarios present open-ended audio (with transcript) problems and exercises, eight for each module, in workplace, school, community, and personal settings. Scenarios are four to six sentences in length, describing a setting, an issue or problem, then, optionally, a resolution. Some are designed to elicit a discussion about the competency invoked in the scenario; some are designed to choose between different ways to handle a situation. For example, a scenario might discuss how a worker handled a situation in which the boss was out of the office or the worker was given a new and difficult task. The scenario is intended to elicit a discussion about what the best way to handle the situation might be. Scenarios sometimes are open-ended, and sometimes provide effective and ineffective resolutions, which then ask learners to choose between them, allowing for practice and skill demonstration (based on the material in the *Pre-work* and *Covering the basics*). An example scenario (*Teamwork and Citizenship* module) is as follows:

> *A classmate and I had to work on a long-term project. Unfortunately, she and I had had some disagreements in the past, and we had very different ways of doing things. We decided to work independently and only come together when absolutely necessary. When we did work together,*

> *most of the time was spent arguing or criticizing one another's work. We got the project done so even though we had an unconventional working style, we still finished the work.*

Workplace stories present vignettes in which several workers are disagreeing on an approach to a problem, some of which elicit open-ended discussion, and some presenting alternative solutions, and the learner indicates which best demonstrates the target competency. Some of these are enacted through role-playing (each team member assumes the role of one of the workers depicted). Following the role-plays, a filmed vignette might depict different approaches to the situation acted out and students choose the better resolution. Four stories are featured per module. An example story (for the *Responsibility* module) is as follows:

> *Parisa supervises a group of fifteen people, including Kelly, Lorne, and Matt. During a department meeting, Parisa singles out Matt for his contributions to an important project. She awards him a $200 bonus for his extra efforts. Matt is surprised, but he graciously accepts the recognition. A moment later, he whispers to Lorne that he has no idea what project Parisa is talking about. Kelly knows that Matt is getting credit for her work. Nevertheless, she feels uncomfortable speaking up.*
>
> *When acting out the story, consider this: How could the situation be resolved in a respectful, productive way?*
>
> *After your story, answer this with your class: What elements of effective or ineffective responsibility are demonstrated in this resolution?*

Workplace stories also include rubrics which indicate novice, developing, and expert levels of the subconstructs (e.g., for *Responsibility*, subconstructs are work ethic, dependability, self-discipline, and orderliness). For example, for the *Responsibility* subconstruct *Self-discipline*, novice level is "members are easily distracted. They rush through their work and their work may be careless;" and expert level is "members are focused on the task. They don't rush through their work and are careful and thorough." Rubrics are used for scoring the teams in the stories into the 3 expertise levels and include open-ended "constructive criticism" blanks for evaluators to complete based on the vignettes.

ACE your interview refers to an acronym helpful for addressing behavioral interview questions, such as "Tell me about a time when you …". Following the acronym, the interviewee first describes the Activity (or Action), describes the Consequences of that activity, and describes the lasting Effects on the interviewee and the organization. This component allows students to practice responses, for example, by being paired with a classmate to practice. An interview scoring rubric is also provided.

Reflection and goal setting is a concluding component that reviews the main ideas presented, allows learners to reflect on the competency, and suggests that learners generate short-, intermediate-, and long-term goals for developing the competency. This component also presents a journal opportunity in that students can record their reflections and goals in their journal. Students also choose someone such as a current employer or other significant parties who can help monitor their development and provide feedback.

Knowledge Check comprises 10 multiple-choice or true-false questions to evaluate students' understanding of the competency. They are presented as audio- or video-based scenarios (along with a transcript) in a situational judgment type format. These are presented after each module.

Journal is an online (or paper-and-pencil) journal that serves as a place to record reflections and objectives pertaining to the competency. The notes recorded are useful for later consultation and for reflection during the *Reflection and goal-setting* component.

FACETS is a behavioral assessment designed to measure students' social, emotional, and self-management skills (Naemi et al., 2014). It is like the assessment used in the Work Readiness Strength Assessment and Training System (WRSATS) (described in the *ETS Soft Skills Training History* section). It can be used initially or after training to provide students feedback on their behavioral dimension strengths and weaknesses to identify areas to capitalize on (for example, in career decision making) and areas for improvement. Dimensions measured by FACETS can be categorized into the six competencies (*Initiative and Perseverance, Responsibility, Flexibility and Resilience, Teamwork and Citizenship, Customer Service Orientation, and Problem Solving and Ingenuity*). Competency levels are determined as unit-weighted composites of subsets of the 13 FACETS dimension scores.

The system also includes some *behaviorally anchored rating scales (BARS)* (Klieger et al., 2018) designed to have an employer (e.g., in an externship setting) or instructor evaluate students' progress.

Prospects for Implementing ETS Soft Skills Training in GIFT

In this section we consider how ETS's soft skills training system could be implemented in GIFT. GIFT provides a set of tools and methods that can be used to author computer-based tutoring systems. Courses can be authored in either a local or cloud environment. The local environment, requiring a download of the GIFT software to run under Windows (https://gifttutoring.org Downloads), contains the full set of features and options including sensor interfaces, monitoring tools, and developer documentation. The GIFT Cloud/Virtual Open campus (https://cloud.gifttutoring.org) does not require a download. It is useful because it contains sample courses that can be taken to get a sense for GIFT features (https://cloud.gifttutoring.org/dashboard/#takeacourse). To author a course, it is necessary to register on the GIFT site to obtain the GIFT software, which is explained in a Quick Start Guide (Ososky, 2017).

Constructing a GIFT course

An adaptive GIFT course (Domain Session) can be developed with the Domain Authoring Tool (DAT) to create a Domain Knowledge File (DKF). The DKF contains the rules for assessment, strategies, and actions, and is read by the Domain and Pedagogical Modules (the workings of these modules are explained in more detail, below). The course can have regular Exercises (formative or interim assessments) and a final Evaluation (end-of-course summative assessment). Real time analyses, such as of a conversation or of performance in a simulator, are referred to as *Assessments.*

A Course Creator contains authoring tools to create and edit GIFT courses. This is done by using Course Objects (e.g., information from a file, slide show, Powerpoint, survey test). Administration tools include a Survey Authoring System (SAS), Event Report Tools (ERT), and import and export tools. They also include the DAT and DKF (described in the previous paragraph), Sensor Configuration Authoring Tool (SCAT), Learner Configuration Authoring Tool (LCAT), Course Authoring Tool (CAT), Metadata Authoring Tool (MAT), and Pedagogy Configuration Authoring Tool (PCAT).

It would seem possible to author or adapt the ETS Soft Skills Training system into a GIFT environment using the administration tools described above. There are some questions that could be addressed while doing so. Would it be easier to use a Blackboard implementation or to reauthor using GIFT? What would be the added value of GIFT? To what extent would adaptivity provided by GIFT lead to better student learning? What separates the GIFT approach from standard LMSs? In what follows, we consider the key GIFT modules—domain, learner, pedagogical, and sensor modules—and what issues might arise in implementing ETS Soft Skills Training in GIFT.

Domain Module

In GIFT, the domain module contains the domain content for training and "is responsible for providing assessments of user ability/knowledge and responding to instructional strategy requests." (ARL, 2015b, Domain Module para.). It assesses learner performance with respect to standards, communicates learner performance to the learner module, and implements requests from the pedagogical module (instructional intervention or feedback, scenario adaptation, requests for performance assessment) as appropriate. The DKF is read at the start of a session; it contains domain-specific rules for performance assessment.

At a high level, much of the domain content of the ETS system—articles, videos, assessments, rubrics—seems compatible with GIFT hosting. This would probably be the most straight-forward part of using GIFT for soft skills training.

Learner Module

In GIFT, the learner module maintains the learner state, which is "communicated via messages to the pedagogical module" (ARL, 2015b, Learner State para.).

Learners are evaluated in the ETS system by being administered a personality assessment (FACETS or something similar) at pretest then a set of behaviorally anchored scales (Klieger et al., 2018) after completing the course, which are completed by employers to measure their competency levels on the six competencies. During the course, they also are given various assessments throughout to check their mastery of the declarative content of the material (e.g., the degree to which they understand the definitions of the six competencies), and multiple-choice knowledge checks. Learner competency levels and knowledge levels could be represented in the learner module based on these assessments. In addition, there are other opportunities for learners to display mastery of the module concepts, such as in the mock interview, in role-playing exercises (in workplace stories), in evaluations of the vignettes (characterizing vignette participants by proficiency level in workplace stories), and in reactions to workplace scenarios. These other opportunities are not scored in the current system, although in principle they could be. There would need to be some development in first human scoring of these more subjective responses, and then subsequently, some real-time natural language processing (NLP) analyses of auditory or written open-ended responses. This could be accomplished using methods ETS (2021) has developed for such applications. Combining information from the full array of assessments given in the ETS system, many of which are not now scored, to get a more complete picture of an individual student's competency knowledge level (i.e., the "learner state"), would seem to be a useful activity for adding instructional value to the ETS system.

Pedagogical Module

In GIFT, the pedagogical module "use(s) information about the learner's state to select instructional strategies that better influence learning" (ARL, 2015b, Pedagogical Module para.), including feedback, hints (on failed problems), varying levels of hints based on requests, instructional intervention, and further performance assessment. The information on the learner's state is based on trainee performance and is taken from the learner module.

The current ETS system does not adapt to learners' states other than through a somewhat arbitrary sequencing of instruction (the *Introduction* and *Summary* modules are naturally sequenced, of course). The system is typically taught as part of a class by a human instructor. GIFT hosting of the ETS system could prompt discussions about the potential benefits of adaptation and enable a self-instruction approach. GIFT appears to allow for many kinds of pedagogical approaches and rules for providing feedback, hints, and additional assessment.

Exploiting the pedagogical module potentially could provide real additional value to the ETS system with GIFT methods. The instructional examples discussed in GIFT applications tend to fall into two categories. Traditional applications use Gagne's (Gagne et al., 1992) and Merrill's (2012) instructional system design frameworks (the US Air Force's Instructional System Design [ISD] approach is consistent with these), along with Bloom's taxonomy (Anderson & Krathwohl, 2005) and Chi and Wylie's (2014) useful Interactive-Constructive-Active-Passive (ICAP) cognitive engagement framework.

The other category is the use of Bayesian knowledge tracing (BKT) (Anderson et al., 1995; Corbett & Anderson, 1995). BKT relies on a cognitive model of learning to evaluate student actions against the model to infer the learner's state, enabling the adaptation of instruction to that state. As van de Sande (2013) points out, there are two forms of the model, a hidden Markov model (HMM) form and a knowledge tracing algorithm form. The HMM models the process of successfully applying a skill to solve problem *j* (an item, an attempt, a problem), $P(C_j)$, as a function of the probabilities of initially possessing the skill prior to the attempt, $P(L_0)$*;* guessing correctly, $P(G)$; slipping (making a mistake even though the skill is known), $P(S)$*;* and learning the skill during this attempt, $P(T)$, the hidden variable in the HMM. Van de Sande (2013) showed that the model can be rewritten as an exponential learning model with 3 parameters, $P(S)$, $P(T)$ (which can be rewritten to a learning rate, β), and an A parameter that represents the probability that the skill is not yet learned. The knowledge tracing algorithm differs from the HMM in that it uses student performance on an item *j* (correct or incorrect) to update the conditional probability that the student has learned the skill given a set of performances on previous attempts, $P(Lj|Oj)$*,* where *Oj* is the history of successes and failures (corrects and incorrect attempts) to that point *j*. This algorithm also incorporates $P(S)$ and $P(G)$*.*

Because BKT models item responses and there is a well-developed science (psychometrics) and theory (item response theory, IRT) for modeling item responses, it is somewhat curious that only recently have there been attempts to apply IRT to the problem BKT is designed to solve. This may be due to different literatures and histories, with BKT emanating from learning theory and experimental psychology, and IRT from psychometrics and correlational psychology. Alternatively, it could be that the relevance of IRT, which targets the measurement of stable attributes, to the measurement of dynamically changing skills was not obvious. In any event by now there have been several attempts to do so, which are informative.

IRT is a general, flexible framework and there are many IRT models, parameter estimation and model fit methods, and applications (van der Linden, 2016a; 2016b; 2018). IRT models item responses as a function of both person and item effects, which are generally assumed to be stable parameters. This differs fundamentally from BKT, which assumes that items do not vary in their difficulty, but that persons change (learn) during the session. IRT datasets tend to have many more persons than items ($n > p$) whereas BKT applications often have more items than persons (known as the $p >> n$ problem in machine learning). IRT is mostly unidimensional, although multidimensional IRT models are available, typically for low dimensional applications, whereas in BKT the implicit assumption is that there are many skill dimensions.

Khajah et al. (2014), noted the limitations of IRT (assumes stable traits) and BKT (assumes equal item difficulties) for modeling ITS response data. They proposed a hybrid model, which replaces the emission (guesses and slips) probabilities in the HMM with an IRT model, and they used both an Expectation Maximization (EM) and a Bayesian estimation technique (Markov chain Monte Carlo slice sampling) to estimate model parameters. Comparing BKT, IRT, and the hybrid model on 4 ITS datasets using a cross-validation approach, they found no differences in the smallest dataset (N = 59 students), superior IRT performance in the middle-sized datasets (N = 66, 110), and superior hybrid performance in the largest dataset (N = 333). They suggest that IRT can model responses from ITS/learning datasets because despite ITS personalization, items tend to be determistically ordered and IRT accommodates order effects as item parameter differences and student learning is then reflected in item effects.

In their comprehensive comparison of BKT and IRT, Deonovic et al. (2018) review various extensions to the BKT and IRT models which move them closer to each other. These include dynamic IRT models with multiple ability parameters for persons at different time points and IRT models that allow dependencies between item responses to accommodate repeated item attempts (e.g., after hints). However, Deonovic et al. (2018) also point out that both models are limited in how they account for the educational experience per se. BKT's learning parameter allows personalization that depends almost entirely on initial conditions, leading Deonovic et al. to refer to BKT as a ballistic model, more similar to firing a canon than to flying a plane. And IRT, which is designed for cross-sectional data, can explain differences between people but does not have much to offer in suggesting remedies. They suggest that network psychometrics models may offer a way to integrate learning and assessment, but this is mostly a theoretical argument at this time.

There are other developments worth mentioning. One is the use of cognitive diagnostic models (CDM) for BKT (Wang et al., 2018). A CDM is a model of item responses in which items are tagged with a set of attributes (e.g., skill requirements) in a Q matrix, making them useful when items mix skills requirements, which is the typical case. Wang et al. (2018) combine a hidden Markov model like the one used in BKT with the CDM framework. This provides a benefit of being able to track the growth of multiple skills and accommodates covariates to model the HMM skill transitions.

Finally, there are developments in using machine learning methods for knowledge tracing as represented in the research called deep knowledge tracing (DKT) (Piech et al., 2015). DKT uses Recurrent Neural Networks (RNN), a neural network that incorporates time dynamics so that previous outputs can be used as inputs; which is suitable for knowledge tracing because the learner's history is incorporated in the modeling. RNNs are widely used in natural language processing and speech recognition. Piech et al. (2015) adapt them for the knowledge tracing problem; student responses (x_t) are inputs to the network and the predictions (y_t) are the probabilities of getting each item correct. The network "discovers" dependencies (conditional influences) between items (in the best-case scenario, into interpretable clusters) avoiding the need for experts to tag items with the skill that is being exercised by that item, an expensive process required in traditional BKT. This method requires lots of data, and thus the RNN approach is designed for large-scale online learning. Also, RNN results are not always readily interpretable (Ding & Larson, 2019). Nevertheless, as data sets accumulate, and with the flexibility of the general deep learning approach for accommodating performance data from all kinds of exercises and inputs, a deep learning approach seems to be particularly promising (Wilson et al., 2016). This may be especially true in the realm of soft skills, given the tendency, realized in the ETS system, to incorporate diverse instructional and assessment approaches.

The developments discussed here vary in their applicability and usefulness, and there is still much to be learned regarding the benefits of applying these models for personalizing instruction. It is also the case that much of the work in this area has focused on the development of hard skills, specifically ones that are more naturally discrete, sequenced, and cumulative. However, it would be useful to experiment with some of these approaches for tracking student growth in the accumulation of soft skills.

Sensor Module

The sensor module includes interfaces to support sensors to measure electrodermal activity (EDA) (or galvanic skin response, GSR). EDA (or GSR) is often interpreted as a measure of emotional state or shifts in emotion, motivation, attention, and preferences (Mendes, 2009).

The current implementation of the ETS system does not use measures of emotion although there is interest in this area for measuring student engagement (Mota & Picard, 2003). Engagement measurement and affect sensing is an active research area in intelligent tutoring (Chen et al., 2021; D'Mello et al., 2007) and assessment (Halderman et al., 2021).

Discussion

As this brief review of the ETS soft skills training system suggests, training soft skills is not fundamentally different from training hard skills. Thus, in principle, it should be quite possible to develop a soft skills training system in GIFT, which could take advantage of GIFT's "tools, methods and standards" (ARL, 2015b, Generalized Intelligent Framework for Tutoring [GIFT] para.) for building a useful, robust system. Nevertheless, as Martin-Raugh, Williams et al. (2020) noted, there has been a dearth of systematic attempts to train soft skills, particularly compared to hard skills, and a question is why that has been the case. There may be several explanations.

First, the recognition of soft skills as an important predictor of and result of education, comparable to hard skills, is relatively recent (Heckman & Kautz, 2012). Thus, there has not been a long-standing focus on training soft skills due to the relatively recent awareness of their importance. Second, although there has long been training and education on specific topics such as leadership, negotiation, and time management, there has simultaneously been a cultural belief that general soft skills, such as motivation, preferences, attitudes, and personality traits are relatively enduring over the lifespan and therefore outside the realm of education, training, or of what ought to be trained. This situation has been changing as we document in the introduction to this chapter. But the third explanation might be that we simply have lacked good approaches for measuring soft skills. This holds back progress because if we cannot measure soft skills training system efficacy very well, then it is difficult to determine what works.

Soft skills are almost exclusively measured with rating scales (e.g., Likert scales) completed by the self or by others. Rating scales have well documented problems such as response style biases, reference group effects, and social desirability bias as well as halo and horn effects (with others' ratings). The ETS system relies on a behavioral rating system, FACETS, to measure students' competency levels. Although FACETS eliminates some of these biases (i.e., response style, social desirability) due to its multidimensional forced-choice design (Naemi et al., 2014), it is nevertheless fundamentally a self-evaluation method rather than a performance method. There are performance measure components to the ETS soft skills training system, such as declarative knowledge fact checks which assess one's knowledge of the competency. There are also performance tasks, such as the mock interviews, role plays, and critiques, accompanied by rubrics for evaluating these. These are useful assessments, and it would be worthwhile pursuing how such assessments might be automatically scored. Chen et al. (2016) provide preliminary evidence of progress along these lines.

Another measurement challenge is the vagueness of the items measuring the constructs being evaluated. For example, the competency *Initiative and Perseverance* can be self-assessed with agreement self-ratings to statements such as "I get chores done right away," "I am exacting in my work," or "I make plans and stick to them." (items taken from the International Personality Item Pool, Goldberg et al., 2006). These are typical behavioral statements used in soft skills assessments. But these are vague and subject to idiosyncratic interpretations. Situational judgment tests (McDaniel & Nguyen, 2001) are one attempt to avoid the vagueness of these kinds of statements, by providing more context. There are performance tasks for other domains as well, such as real-effort tasks for conscientiousness constructs (Charness et al., 2018) and negotiation (Martin-Raugh, Kyllonen et al., 2020) and collaborative problem-solving (Hao et al., 2019) for social skills. These are research endeavors now but may soon be used routinely in operational soft skills assessment contexts.

Finally, there is a question of the importance of adaptation in instruction generally, but in soft skills instruction particularly. A critique of intelligent tutoring is that its usefulness is observed primarily in well-defined domains, but soft skills training would be considered an ill-defined domain (Fournier-Viger et al., 2010). This makes for challenges for implementing soft skills training using intelligent tutoring. Currently, the ETS system is designed for human instructor use, and the human instructor makes pacing decisions. But the benefits of individual tutoring are well documented, as are the benefits of computer tutoring (Kulik & Fletcher, 2015; van Lehn, 2011). Thus, it should be possible to improve on classroom soft-skills instruction by adopting intelligent tutoring frameworks. At this point, it seems that constructs and measures must be developed further to make them more amenable to intelligent tutoring, and there is probably a great need for additional measurement opportunities and tasks to elicit evidence for possession of the target competencies.

Recommendations and Future Research

First, given the dearth of soft skills training in the world (Martin-Raugh, Williams et al., 2020) and its complete absence in the world of intelligent tutoring, combined with the expressed need for it, it would be useful to initiate a project that attempted to implement soft skills training using an intelligent tutoring framework. This could be done in GIFT according to our analysis here. It is often only through the action of attempting to build something that the possibilities and limitations become clear. Quoting Danny Hillis, "you can cut through a lot of philosophy with a few demonstrations" (Sanger, 1987, p. D2). It may be that the most useful recommendations for potential new GIFT features would emerge as a result of implementing soft skills training in GIFT.

However, we have identified measurement as a key limitation of current soft skills training. There is a need to develop additional performance measures of soft skills, which include situational judgment tests, but can also include real-effort tasks, collaborative problem-solving tasks, time-, risk-, and social-preference tasks, and other new measures. GIFT easily accommodates basic multiple-choice tasks and rating scales, but it may be useful to expand GIFT capabilities towards compliance with the Question and Test Interoperability Version 3.0 (QTI v3) specification (IMS Global, 2020) for storing and exchanging items and tests, deploying item banks, and reporting test results in a consistent manner.

Also, many of the assessments in the current ETS soft skills training, involving mock interviews, role plays, and character critiques, are ill-defined, subjective, and complex. There is a need for technology to help evaluate performances in these kinds of settings. Preliminary, basic research is underway, but there is a need for more investment in these kinds of realms. GIFT features that would facilitate these kinds of assessments would include a capability to video record a trainee's performance (e.g., interview, role play) following a prompt, which could be later retrieved by an evaluator for scoring using playback (e.g., pause, fast-forward) and annotation and input tools (e.g., for comments, ratings). Over time real-time automated scoring based on stimulus features could be added as in Hoque et al. (2013).

Finally, there is a need for further development of the constructs that are taught and assessed in soft skills training. Although steering this research agenda is not a GIFT activity, accommodating new developments will be enhanced with interoperability features, such as QTI v3 (IMS Global, 2020). It could be that with additional measures and with extensive additional data collection, better understanding of the nature of soft skills themselves will result. The causes and consequences of personality variation has long been of interest (Lee, 2012). Training personality, and training soft skills will shed light on these issues.

Conclusions

There has been increasing awareness of the importance of soft skills in education, training, and in the world of work. Although there has long been soft skills training for particular skills, such as time management or leadership, there has been a lack of general soft skills training targeting social skills or work ethic, for example. It would be useful to attempt to implement soft skills training on a platform such as GIFT, building on extant soft skills training designed for classroom use. Challenges faced and lessons learned would be useful for our understanding of soft skills generally as well as leading to a useful practical tool.

References

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228(4698),* 456-462. DOI: 10.1126/science.228.4698.456

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4(2),* 167–207.

Anderson, L. W., & Krathwohl, D. R. (2005). *A taxonomy for learning, teaching, and assessing.* London, England: Longman

ARL (2015a). Generalized Intelligent Framework for Tutoring (GIFT), Wiki, Overview. https://gifttutoring.org/projects/gift/wiki/Overview

ARL (2015b). Generalized Intelligent Framework for Tutoring (GIFT), Wiki, GIFT Glossary https://gifttutoring.org/projects/gift/wiki/GIFT_glossary

Arthur, W., Jr., Bennett, W., Jr., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88,* 234–245. https://doi.org/10.1037/0021-9010.88.2.23

Barr, A., Beard, M., & Atkinson, R. C. (1975). The computer as a tutorial laboratory: The Stanford BIP project. *International Journal of Man-Machine Studies, 8(5),* 567-582. https://doi.org/10.1016/S0020-7373(76)80021-1

Charness, G., Gneezy U, & Henderson A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization, 149*, 74-87. doi: 10.1016/j.jebo.2018.02.024. PubMed PMID: WOS:000438480800005.

Chen, L., Feng, G., Martin-Raugh, M. P., Leong, C. W., Kitchen, C., Yoon, S.-Y., et al. (2016). Automatic scoring of monologue video interviews using multimodal cues. *Interspeech 2016*. DOI: 10.21437/Interspeech.2016-1453

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.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49(4),* 219-243. DOI: 10.1080/00461520.2014.965823

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*, 1–444. https://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., Wood, D., Mottus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A.G.C., Ziegler, M., & Zimmermann, J. (2021). Bottom-up construction of a personality taxonomy, *European Journal of Psychological Assessment, 36,* 923-934. https://doi.org/10.1027/1015-5759/a000626.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adaptive Interaction, 4(4),* :253–278.

Corcoron, R. P., Cheung, A. C. K., Kim, E., & Xie, C. (2018). Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research. *Educational Research Review, 25,* 56-72.

D'Mello, S. K., Picard, R., & Graesser, A. C. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, *22*, 53–61.

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., & Maris, G. (2018). Learning meets assessment: On the relation between item response theory and Bayesian knowledge tracing. arXiv:1803.05926v2

Ding, X., & Larson, E. C. (2019). *Why deep knowledge tracing has less depth than anticipated*. In M. Desmarais, C. F. Lynch, A. Merceron, & R. N. Kambou (Eds.), The 12th International Conference on Educational Data Mining (pp. 282-287). Montreal, Canada.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical Report 1311). Ft. Belvoir, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Duncan, R., Washburn, I.J., Lewis, K.M., Bavarian, N., DuBois, D.L., Acock, A.C.…Flay, B.R. (2017). Can universal SEL programs benefit universally? Effects of the Positive Action Program on multiple trajectories of social-emotional and misconduct behaviors. *Prevention Science, 18,* 214–224.

Durlak, J.A., Weissberg, R.P., Dymnicki, A., Taylor, R.D., & Schellinger, K.B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82,* 405–432.

ETS (2018a). *WorkFORCE® Program for Career Development: Using Your Results.* Princeton, NJ: Educational Testing Service.

ETS (2018b). *WorkFORCE® Program for Career Development: Instructional Guide.* Princeton, NJ: Educational Testing Service.

ETS (2021). *Automated scoring and natural language processing*. https://www.ets.org/research/topics/as_nlp

Fournier-Viger, P., Nkambou R., & Nguifo, E. M. (2010). Building Intelligent Tutoring Systems for Ill-Defined Domains. In: Nkambou R., Bourdeau J., & Mizoguchi R. (Eds.) *Advances in Intelligent Tutoring Systems. Studies in Computational Intelligence, vol 308.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_5

Gagné, R. M., Briggs, L. J., & Wager, W. W. (1992). *Principles of instructional design (4th ed.).* Forth Worth, TX: Harcourt Brace Jovanovich College Publishers.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Graesser, A.C., Hu, X., Nye, B., Sottilare, R. (2016). Intelligent tutoring systems, serious games, and the Generalized Intelligent Framework for Tutoring (GIFT). In H.F. O'Neil, E.L. Baker, and R.S. Perez. (Eds.), *Using games and simulation for teaching and assessment* (pp. 58-79). Routledge: Abingdon, Oxon, UK.

Halderman, L. K., Finn, B., Lockwood, J. R., Long, N. M., & Kahana, M. J. (2021). *EEG correlates of engagement during assessment* (Research Report No. RR-21-01). Educational Testing Service. https://doi.org/10.1002/ets2.12312

Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. A. (2019). *Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving* (Research Report No. RR-19-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12276

Haviland, S.B., Robbins, S., Belur, V., Cherfrere, G., & Klieger, D. (2021). Improving Workforce Readiness Skills among Adult Learners through new Technologies: Lessons from Two Schools. *Metropolitan Universities, 32(1),* 35-53. DOI: 10.18060/23884

Heckman, J. J. & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19(4):* 451–464. doi:10.1016/j.labeco.2012.05.014.

Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). MACH: my automated conversation coach. In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp '13). ACM, New York, NY, USA, 697-706. ISBN 9781450317702.

IMS Global (2020). Question and Test Interoperability (QTI) Overview: IMS Candidate Final Public Version 3.0. https://www.imsglobal.org/spec/qti/v3p0/oview/

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (p. 102–138). Guilford Press.

Khajah, M. M., Huang, Y., Gonzalez-Brenes, J. P., Mozer, M. C., & Brusilovsky, P. (2014). *Integrating Knowledge Tracing and Item Response Theory: A Tale of Two Frameworks.* In M. Kravcik, O. C. Santos, & J. G. Boticario (Eds.) 4th International Workshop on Personalitzation Approaches in Learning Environments (PALE 2014), Aalborg, Denmark.

Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., & Shore, J. R. (2018). Development of the behaviorally anchored rating scales for the skills demonstration and progression guide. *ETS Research Report Series, 2018 (1),* 1-36. https://doi.org/10.1002/ets2.12210

Knowles, M. S., Holton, E. F., & Swanson, R. A. (2015). *The adult learner: The definitive classic in adult education and human resource development (8th ed.).* New York, NY: Routledge.

Kulik, J.A., & Fletcher, J.D. (2015). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 85,* 171-204.

Lee, J. J. (2012). Correlation and causation in the study of personality. *European Journal of Personality, 26(4*), 372-390. doi:10.1002/per.1863

Mahoney J. L., Durlak J. A., & Weissberg R. P. (2018). An update on social and emotional learning outcome research. *Phi Delta Kappan. 100(4),* 18-23. doi:10.1177/0031721718815668

Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., Yang, Z., Yan, F., & Barnwell, P. (2020). Negotiation as an interpersonal skill: Generalizability of negotiation outcomes and tactics across contexts at the individual and collective levels. *Computers in Human Behavior, 104,* 105966. https://doi.org/10.1016/j.chb.2019.03.030

Martin-Raugh, M. P., Williams, K. M., & Lentini, J. (2020). *The malleability of workplace-relevant noncognitive constructs: Empirical evidence from 39 meta-analyses and reviews* (Research Report No. RR-20-23). Educational Testing Service. https://doi.org/10.1002/ets2.12306

McCormack, M. H. (1984). *What they don't teach you at Harvard Business School: Notes from a street-smart executive.* New York: Bantam Books.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9(1-2),* 103–113. https://doi.org/10.1111/1468-2389.00167

Mendes, W. B. (2009). Assessing the autonomic nervous system. In: E. Harmon-Jones and J.S. Beer (Eds.) *Methods in social neuroscience* (pp.118-147). New York, NY: Guilford Press.

Merrill, M. D. (2012). *First Principles of Instruction.* Germany: Wiley.

Mota, S., & Picard, R. W. (2003). *Automated Posture Analysis for Detecting Learner's Interest Level.* Conference on Computer Vision and Pattern Recognition Workshop, 5, 49. http://dx.doi.org/10.1109/cvprw.2003.10047

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S.,Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality, 34,* 1175–1201. https://journals.sagepub.com/doi/full/10.1002/per.2311

Naemi, B. D., Seybert, J., Robbins, S. B., & Kyllonen, P. C. (2014). *Examining the WorkFORCE™ Assessment for Job Fit and Core Capabilities of FACETS™.* Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12040.

National Academies of Sciences, Engineering, and Medicine. (2017). *Supporting Students' College Success: The Role of Assessment of Intrapersonal and Interpersonal Competencies.* Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24697.

National Research Council (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: National Academies Press.

Ososky, S. (2017). *Generalized Intelligent Framework for Tutoring (GIFT) Cloud/Virtual Open Campus Quick-Start Guide (Revision 1)*. Technical Report ARL-CR-0816. Oak Ridge, TN: Oak Ridge Associated Universities. https://gifttutoring.org/documents/116

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In M. J. Jordan, Y. LeCun, & S. A. Solla (Eds.) *Advances in Neural Information Processing Systems*, pp. 505-513. MIT Press. ISBN: 9780262561457.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135(2),* 322–338. https://doi.org/10.1037/a0014996

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin, 143(2),* 117–141. https://doi.org/10.1037/bul0000088

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*, 1– 25.

Salgado, J.F., Anderson, N. & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 8,* 797-834. https://doi.org/10.1111/joop.12098

Salgado, J. F. & Táuriz, G. (2014) The Five-Factor Model, forced- choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies, *European Journal of Work and Organizational Psychology, 23(1),* 3-30, DOI: 10.1080/1359432X.2012.

Sanger, D. E. (1987, December 24). I.B.M. signals big shift in designing computers. *New York Times*, Section A, Page 1/Page D2, Column 4.

Saucier, G., & Iurino, K. (2020). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology, 119*, 1188-1219.

Shore, J. R., Lentini, J., Rikoon, S., Seybert, J., & Noeth, R. (2016). *The ETS Work Readiness Strength Assessment & Training System (WRSATS).* Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100(2),* 330–348. https://doi.org/10.1037/a0021717

Sottilare, R.A., Baker, R.S., Graesser, A.C., & Lester, J.C. (2018). Special issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a stable and flexible platform for innovation in AIED research. *International Journal of Artificial Intelligence in Education, 28,* 139-151.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: US Army Research Laboratory.

Sottilare, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012). *A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS).* In Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (pp. 1-13).

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29(3),* 184–203. https://doi.org/10.1177/0146621604273988

Van de Sande, B. (2013). Properties of the Bayesian knowledge tracing model. *Journal of Educational Data Mining, 5(2)*, 1-10.

Van der Linden, W. J. (2016a). *Handbook of item response theory: Volume One, Models*. Boca Raton, FL: CRC Press.

Van der Linden, W. J. (2016b). *Handbook of item response theory: Volume Two, Statistical Tools*. Boca Raton, FL: CRC Press.

Van der Linden, W. J. (2018). *Handbook of item response theory: Volume Three, Applications*. Boca Raton, FL: CRC Press.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist. 46 (4):* 197–221. doi:10.1080/00461520.2011.611369

Vanhove, A. J., Herian, M. N., Perez, A. L. U., Harms, P. D., & Lester, P. B. (2016). Can resilience be developed at work? A meta-analytic review of resilience-building programme effectiveness. *Journal of Occupational and Organizational Psychology, 89,* 278– 307. https://doi.org/10.1111/joop.12123

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics, 43(1),* 57-87. doi:10.3102/1076998617719727

Wilson, K. H., Xiong, X., Khajah, M., Lindsey, R. V, Zhao, S., Karklin, Y. et al (2016). *Estimating student proficiency: Deep learning is not a panacea.* Submission to the NIPS 2016 Workshop on Machine Learning for Education, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

SECTION II– SCENARIO BASED TRAINING FOR GROUPS AND TEAMS

Dr. Joan H. Johnson and Dr. Andrew J. Hampton, Eds.

Chapter 7 – Introduction to Scenario-Based Training for Groups and Teams

Joan H. Johnston[1] and Andrew J. Hampton[2]
U.S. Army DEVCOM Soldier Center[1], University of Memphis[2]

Core Ideas

This book section focuses on the critical importance of improving the capability of Intelligent Tutoring Systems (ITSs) to develop group and team competencies. As the focus of nearly three decades of military research, this aspect of ITS development has proven to be complex, difficult, and costly. Nevertheless, the steadily increasing demand for team tutors justifies its continued development. Chapters in this section present both broad and focused coverage of this topic to illustrate employing scenario-based training design principles to tailor competency development.

We begin this section with guidelines for designing competency-based scenarios and recommendations for GIFT design based on a team training study with U.S. Army squads (Johnston, Sottilare, Kalaf, & Goodwin). Next, Johnston, Patton, and Sinatra discuss results from the same study to propose a measurement framework for assessing development of individual and team resilience, and discuss implications and recommendations for the GIFT. The next two chapters delve into practical application and validation of intelligent agents for team and group tutors. Myers describes lessons learned for GIFT (Generalized Intelligent Framework for Tutoring) based on an assessment method they studied for evaluating synthetic teammate performance in a team triad. Johnson and Gratch present empirical findings from two experiments that employed a synthetic agent engaging with a human trainee to improve trainee negotiation skills. Together, these chapters provide working knowledge of the state of the art in scenario-based training for groups, as well as the theoretical background that informed efforts to expand and improve team and group tutors.

Individual Chapters

Johnston, Sottilare, Kalaf, and Goodwin present a discussion of the methodology employed to develop competency-based scenarios for the Squad Overmatch (SOvM) simulation and live training exercises. The Event Based Approach to Training (EBAT) method was used to design tactical combat casualty care (TC3) training scenarios for dismounted Army infantry squads that were each assigned a combat medic. Details about the EBAT method and scenario event development are described with a focus on specifying observable behaviors for advanced situation awareness, team development, stress management, and TC3 that could be assessed by subject matter experts. The authors discuss challenges encountered in conducting assessments, implications for GIFT, and provide future research recommendations to include defining and standardizing primitive verbal and non-verbal behavioral markers to support the interoperability of team assessment methods across adaptive instructional systems.

Johnston, Patton, and Sinatra apply a theoretical model developed by Bowers et al. (2017) to the Squad Overmatch (SOvM) use case described in the previous chapter to create an initial framework for measuring individual and team resilience. They discuss findings from the SOvM experiment in the context of the model, determining that SOvM Soldiers and squads demonstrated resilient cognitions and behaviors. Implications for using this approach in developing competency-based scenarios and GIFT design are discussed, and future research recommendations are provided.

Myers presents a discussion of previously published research conducted by the Air Force in which the performance of a synthetic teammate embedded with two novice human team members of a remotely piloted aerial system (RPAS) was compared with a novice RPAS human team and an RPAS human team that included an expert pilot. Researchers used ACT-R technology to create the synthetic teammate and designed competency based scenario events to test and compare the performance of the three study condition teams. They found the RPAS team with the expert pilot performed better than the other two conditions which performed at about the same level. Myers discusses the challenges in developing complex intelligent team behaviors for GIFT design, and presents future research recommendations.

Johnson and Gratch begin with an overview of cognitive tutors and diagnostic tutoring models. They then describe how general principles of negotiation informed behavioral models of an ideal negotiator and enabled the development of measureable behaviors with an automated tutor that could provide feedback to trainees. They demonstrate how providing automated, individualized feedback to a trainee about their use of negotiation principles improved their use of those principles later in the game and their final negotiation outcomes. The authors also report that they had successfully integrated one of their ITSs (IAGO) with GIFT to author training for negotiation that included personalized feedback and provided recommendations for improving the efficiency of GIFT to author training.

Chapter 8 – Training for Team Effectiveness Under Stress

Joan H. Johnston[1], Robert A. Sottilare[2], Mike Kalaf[3], and Greg Goodwin[1]
DEVCOM Soldier Center[1]; Soar Technology, Inc.[2]; Synaptic Sparks[3]

Introduction

Military research has demonstrated teamwork, advanced situation awareness, and stress management skills are necessary competencies for effective team decision making under stress (Johnston et al., 2019). Research has shown that competency-based scenario design that focuses on these three skill areas is the key to effective team training. The Event-Based Approach to Training, or EBAT method, was developed to enable trainers to create scenario events that elicit observable team member behaviors representing each of these competencies (Fowlkes et al., 1998; Johnston et al., 2018). The EBAT method ensures that enough critical events are created to give team members the opportunity to perform competency relevant behaviors. Behavioral checklists are used to assess competencies during scenario events and then to inform the post-scenario after action review (AAR) (Johnston et al., 2019). Teams can focus on targeted skill areas (e.g., teamwork) by reviewing and discussing the behavioral evidence associated with the events. Team performance will improve when the scenarios are adapted to focus on the skills needing improvement.

Recently, the EBAT method was used in a series of Squad Overmatch studies that culminated in a tactical combat casualty care training experiment involving dismounted Army infantry squads that were each assigned a combat medic (Johnston et al., 2019). Four control condition squads participated in one day of live training, while four experimental condition squads participated in a three and a half day integrated training approach that employed classroom, virtual, and live training exercises to develop skills in tactical combat casualty care, teamwork, advanced situation awareness, stress management and team self-correction in the AAR. The event-based virtual and live exercises emphasized performing tactical combat casualty care while continuing the tactical mission. Live exercises were conducted in an outdoor urban training site instrumented with human role-players and embedded simulations that included casualty mannequins, interactive virtual avatars, and non-explosive pyrotechnic sounds. Johnston et al. (2019) reported that compared to the control condition, squads trained with the integrated training approach displayed significantly more behavioral markers for tactical combat casualty care, team knowledge emergence (as represented by a combination of advanced situation awareness and teamwork), and team self-correction. In this chapter we describe how the EBAT scenarios were developed for the Squad Overmatch live training exercises, discuss challenges encountered in conducting assessments, and discuss implications for the Generalized Intelligent Framework for Tutoring (GIFT).

Event-Based Approach to Training

A research team comprised of research psychologists, training simulation experts, and military subject matter experts used the EBAT approach to develop two virtual training (B1 and B2) and three live training scenarios (M1, M2, and M3). Johnston et al. (2018) provide details describing each step in the process which we summarize here. First, mission tasks and objectives were defined in the context of the tactical combat casualty care mission. The Army mission essential task lists were used to define the squad tactical tasks, to include: apply troop leading procedures to plan, organize and prepare for missions; determine the pattern of life baseline; recognize changes in the pattern of life; assess changes in the pattern of life; use cues and indicators to make sense of tactical situations; interact with civilian populations; minimize casualties; and defeat the enemy. Next a visual flow diagram of key events was developed to establish a story narrative that logically linked the five scenarios together, with each scenario designed to take about 45 minutes to complete. Key events in the diagram enabled the research team to discuss and solve a multitude of interacting issues, such as event timing, stress levels, and placement of casualty events. The final product was a color-coded, network storyboard of sequenced events and indicators that included the squad's performance steps for each training event with an accompanying written description. This resulted in a training support package and a finalized scenario event checklist organized with a set of decision triggers that were numbered in chronological order. Materials from these data were used to produce the squad's operations order, fragmentary order, and an intelligence picture. The master scenario event lists were used to develop the performance objectives for each competency. Table 1 lists the event-based performance objectives identified for advanced situation awareness, Table 2 lists the teamwork objectives, and Table 3 lists the tactical combat casualty care performance objectives. Each objective is mapped to the 15 relevant critical scenario events in M3:

1. Establish a listening post/observation post
2. Key leader is spotted by squad
3. Conduct a key leader engagement
4. Key leader escorts squad into marketplace
5. Pregnant woman approaches squad for medical help
6. Tactical questioning of High Value Target 1
7. Report Intel about HVT to Platoon
8. Squad enters building with female civilian
9. Civilian receives amputation to lower arm
10. Tactical questioning of local female civilian
11. Soldier receives GSW
12. PL commands squad to enter building across street
13. Team pushes through building to clear targets
14. High value target 2 is killed
15. Two Soldiers receive gunshot wounds

Many performance objectives for each skill area were repeated across events to ensure the squad members had ample learning opportunities. Most events required assessment of at least two skill areas. Some events had many performance objectives to ensure the scenarios had sufficient levels of realism and stress. For example, event 15 was a complex, culminating event with four teamwork and nine tactical combat casualty care performance objectives. In event 15, sniper fire results in two Soldiers receiving gunshot wounds. Expected teamwork behaviors listed in Table 2 are to provide complete and accurate medical reports, provide guidance to each other, exchange information about the casualty, and provide tactical information up the chain of command. Expected tactical combat casualty behaviors listed in Table 3 were to return fire and lay suppressive fire as needed, provide the properly formatted "MANDOWN" report to the squad leader, treat the casualty appropriate to phase and wound, direct team members to suppress the enemy, request medical and / or tactical type information, provide advanced casualty care, direct team members to provide care to a specific casualty, provide medical updates to the squad leader, complete the MIST medical report (a military protocol for relaying critical details of a medical situation), and report the medical casualty via a 9-Line report up the chain of command.

Following performance objective creation, the research team produced tactical job aids that contained all mission essential information for the squad leader (e.g., target packages for each high value target and the pattern of tactical operations). Behavioral checklists were created from the performance objectives for real-time assessment of each skill area that subject matter experts used during both the virtual and live training exercises.

Table 1. Event-based performance objectives for advanced situation awareness in scenario M3.

		1	2	3	4	5	6	7	8	10
1	Squad divides into two separate forces for two Listening/Observation Post	x								
2	Communicate atmospheric details	x								
3	Townspeople avoiding Northeast end of town near Pavel's pad	x								
4	Communicates changes in baseline behaviors of town	x								
5	Positively identify key leader		x							
6	Sets security around key leader engagement in order to cover all avenues of approach			x						
7	Employs guardian angel /geometries of observation			x						
8	Communicates nonverbal behaviors of the key leader (rubbing hands, looking over shoulder)			x						
9	Communicates an assessment to include why s/he believes the validity, quantity of the information received (appears worried)			x						
10	Communicates deviations in baseline of behavior of key leader			x						
11	Identifies minor proxemics push from villagers away from squad				x					
12	Offers some medical care to local national (good shepherd)					x				
13	Employs guardian angel / geometries of observation.						x			
14	Communicates nonverbal behaviors of the high value target						x			
15	Communicates an assessment to include why s/he believes the validity, quantity of the information received						x			
16	Communicates anomalous antenna outside Pavel's pad							x		
17	Employs guardian angel/geometries of observation								x	
18	Employs guardian angel / geometries of observation									x
19	Communicates nonverbal behaviors of the local national									x
20	Communicates an assessment to include why s/he believes the validity, quantity of the information received.									**x**

Table 2. Event-based performance objectives for teamwork in scenario M3.

		1	3	6	7	8	9	10	11	12	13	15
1	Changes in priority are communicated to squad members	x										
2	Sources of information utilized during planning	x										
3	A situation update is provided up the chain of command		x									
4	Back up is provided to squad member engaging in interview		x									
5	Back up is provided to squad member engaging in interview			x								
6	Information is exchanged while posting security			x								
7	A situation update is provided up the chain of command				x							
8	Information is passed before being asked amongst squad members					x						
9	Squad members provide backup					x						
10	Complete medical updates/reports are provided (MIST report, and 9-Line if applicable)						x					
11	Squad members provide guidance to each other in further care of civilian casualty						x					
12	Back up is provided to squad member engaging in interview							x				
13	Information is exchanged while posting security							x				
14	A situation update is provided up the chain of command							x				
15	Changes in priority are communicated to squad members							x				

		1	3	6	7	8	9	10	11	12	13	15
16	Complete medical updates/reports are provided (MIST report, and 9-Line if applicable)								x			
17	Information exchanged between squad members about the casualty								x			
18	Information is exchanged between squad members about the sniper								x			
19	Direction is provided to squad members on how to provide casualty care								x			
20	Back up is provided								x			
21	Changes in priority are communicated to squad members									x		
22	Communications are clear										x	
23	Communications are brief										x	
24	Squad members provide guidance to each other										x	
25	Complete medical updates/reports are provided (e.g., MANDoWN, MIST, 9-Line)											x
26	Information exchanged between squad members about the casualty											x
27	Complete medical updates/reports are provided (MIST report, and 9-Line if applicable)											x
28	A situation update is provided up the chain of command											x

Table 3. Event-based performance objectives for tactical combat casualty care in scenario M3.

		9	13	15
1	Provides MANDOWN Report to SQUAD LEADER.	x		
2	Retrieves casualty (with cover as necessary)	x		
3	Requests medical and / or tactical SA type info.	x		
4	Return fire; lay suppressive fire as needed.		x	
5	Treats casualty appropriate to phase & wound.		x	
6	Communicates phase change in casualty care		x	
7	Establish /move casualty collection point assigning medical &/or tactical resources.		x	
8	Provides medical updates to squad leader.		x	
9	Provides MANDOWN Report to squad leader			x
10	Return fire; lay suppressive fire as needed			x
11	Treats casualty appropriate to phase & wound			x
12	Directs team members to suppress enemy			x
13	Requests medical and / or tactical SA type info			x
14	Provides advanced care			x
15	Directs team members to provide care to specific casualty			x
16	Provides medical updates to squad leader			x
17	Completes MIST medical report (if applicable), and 9-Line			x

Lessons Learned

Researchers and observers successfully used the checklists to assess team performance in real-time during the three live training exercises. During the live exercises, the outdoor facility was instrumented with live video feeds of squads moving through the village at key locations that could be seen by researchers, role players, and observers in the site's quiet viewing theater. In addition, all Soldiers were instrumented with open mics that could be heard via researcher headsets connected to a custom designed audio system in the same theater. Thus, researchers were able to ascertain specific Soldier communications by scenario event, and observers could more effectively conduct the AARs as they had a ready set of documented behaviors.

In contrast, the observational method was not successful with the two virtual training scenarios which took place in a classroom where the squad members were sitting side-by-side at their gaming computers. As the scenarios unfolded, Soldiers were able to speak into their headsets and to those sitting next to them which resulted in a noise level that mitigated researchers' ability to properly assess communication behaviors. Furthermore, researchers did not have gaming PCs that enabled them to easily view Soldier movements in the virtual world to ascertain when events happened and connect them to the communications.

In summary, Johnston et al. (2019) determined that logistics, time requirements, and current technological capabilities greatly limit supporting competency-based team assessments. The Squad Overmatch studies demonstrated that both training environments require automating the EBAT assessments to improve team effectiveness on specific competencies. To address this problem, the next section discusses implications and recommendations for adaptive training and GIFT.

Implications and Recommendations for Adaptive Training and GIFT

To automatically assess team progress toward each of the performance objectives listed in Table 1, GIFT would need to define and recognize a set of behavioral markers associated with each performance objective. The same would be required to support assessments in a virtual training exercise, but the methods to source and infer behaviors would likely be simpler because ground truth measures are available within the virtual simulation. Even so, detecting and recognizing the behaviors (verbal and non-verbal) poses a significant challenge.

A meta-analysis conducted by Sottilare et al. (2017) discovered significant antecedents to team performance and learning, and the identification of associated behavioral markers for several teamwork states. Though this work provides a significant insight to measures that support teamwork assessment, specific behaviors for team-based tasks must be defined and a method must be developed to recognize the behaviors associated with successful completion of each task. For example, the verbal behaviors require different sensors to source sound data while non-verbal behaviors require sensors to detect and recognize gestures.

To adapt to this approach, GIFT course authors would need to be able to define behavioral data sources (e.g., sensors) so that GIFT would recognize variables associated with measures of team assessment. This would involve the development of a compatible interface with the GIFT gateway and the generation of a JAVA condition class to publish the variable to the GIFT environment. With large numbers of variables, data sources, and behaviors this would be an extremely tedious task. Therefore, we have three recommendations:

- Define and reuse GIFT gateway specifications for commonly used sensors.
- Research methods to automatically generate JAVA condition classes for new GIFT variables.
- Define primitive behaviors (verbal and non-verbal) that can be reused in a variety of team training domains.

Furthermore, the integrated training approach incorporated the stress exposure training method which involves increasing exposure to common combat stressors across multiple scenarios. The goal is to "inoculate" Soldiers to stress as they practice and apply skills they are developing for stress management, decision making, and teamwork. Implementing EBAT exercises and collecting the required data would be needed to validate stress levels during team training. The methods described in this paper are based on the training from subject matter experts defining the required criteria and determining the effect of stress on the outcomes and performance. A recommendation that could complement this method of stress evaluation would be to introduce virtual reality into a GIFT environment and instrument each Soldier with sensors collecting pertinent biometric data. The data collected in a cross-referenced AAR could potentially increase the confidence of the training results and provide valuable feedback.

Future Research

Team training environments are highly complex and, as in all training, the context is a significant enabler of the effectiveness of architectural approaches. With this in mind we provide the following recommendations for future research:

- Research, define, and standardize primitive verbal and non-verbal behavioral markers to support the interoperability of team assessment methods across adaptive instructional systems.
- Research and develop methods to automatically detect and recognize primitive behaviors and build maps to recognize compound behaviors.
- Evaluate the effectiveness of adaptive instructional system interventions in team training and develop methods to automatically reinforce their decision-making to optimize team performance and learning.
- Evaluate implementing both virtual reality and sensors that would monitor team member vital biometric data at specific events and study how combining the biometric data and competency based data could potentially improve the impact of competency-based adaptive training.

References

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). *International Journal of Aviation Psychology, 8*(3), 209-221.

Johnston, J. H., Napier, S., Burford, C., Henry, S., Ross, B., & Patton, C. (2018). A test protocol for advancing behavioral modeling and simulation in the Army Soldier Systems Engineering Architecture. In *International Conference on Applied Human Factors and Ergonomics* (pp. 45–55). Springer, Cham.

Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., Patton, D. J., Cox, K. G. & Fitzhugh, S. M. (2019). A team training field research study: Extending a theory of team development. *Frontiers in Psychology*, *10*, 1480.

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2017). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education, 28*(1), 225–264.

Chapter 9– A Measurement Framework for Developing Resilience with Competency-Based Scenarios

Joan H. Johnston[1], Debra Patton[2], and Anne M. Sinatra[1]
DEVCOM Soldier Center[1], DEVCOM Data Analysis Center[2]

Introduction

The U.S. military has invested considerable resources in developing methods to enhance stress resilience in Warfighters. In the last decade, the U.S. Army began implementing a ready and resilient policy for Soldier support programs and research efforts focused on developing Soldier resilience (O'Keefe, 2014). A major focus of this program has been developing training interventions and technologies to accelerate resilience development (e.g., Patton et al., 2018a). Currently the Generalized Intelligent Framework for Tutoring (GIFT) researchers are developing technologies that reliably collect attitudinal, behavioral, and physiological data obtained from collective training simulations using competency-based scenarios for the Synthetic Training Environment (Goldberg et al., 2021). A major goal is to produce real-time feedback and after action reviews (AARs) that accelerate Soldier and team competencies to include stress management skills.

To support these efforts, we propose that a robust theoretical framework is needed to better define individual and team resilience performance indicators, specify how and when it develops, guide developing valid and reliable measures, and inform competency-based training interventions that optimize skills development. Bowers et al. (2017) proposed that team resilience is a complex, multi-level, dynamic construct because it is influenced by and mediates both individual and team behaviors. They developed a theoretical framework with individual, team and organizational inputs, processes, emergent behavioral states (such as team cohesion and efficacy), and outcomes (e.g., evidence of resilience). Emergent team states result from team interaction processes under stress that in turn represent a second order emergent factor of team resilience that influences individual, team and organizational outcomes. They define emergence of team states as "a dynamic process engaged in during significant adversity resulting in positive adaptation" (Bowers et al., 2017, p. 9). In this chapter we applied their model to a team training use case to create an initial framework for measuring individual and team resilience. We discuss implications for using this approach in developing competency-based scenarios and GIFT design, and provide future research recommendations.

Squad Overmatch Use Case

Starting in 2013, the Army's Squad Overmatch (SOvM) research program conducted a series of studies and experiments to improve dismounted Soldier tactical combat casualty care (TC3) (Johnston et al., 2019; Johnston et al., 2021, this book). The Stress Exposure Training (SET) method (Driskell et al., 2008) was used to design competency-based scenario events that would elicit advanced situation awareness, teamwork, and stress management behaviors from Soldiers in a squad formation with an embedded medic to improve performing the TC3 tasks (see Johnston et al., 2021, this book for further details). Five training scenarios (B1 and B2 in the Army's Virtual Battlespace 3 (VBS3) team training simulation and M1, M2, and M3 in an outdoor urban training facility) were developed with the SET method so that successive scenarios had more combat stress incidents than the previous one. For example, M3 had almost twice as many combat incidents as M2. Combat stressors included improvised explosive devices, a suicide vest explosion, sniper shootings, and Soldier and civilian injuries that were simulated in VBS3, and then implemented as realistic simulation devices and with role-players in the live exercises. The eight participating U.S. Army infantry squads were moderately to very experienced. Four squads participated in two days of classroom instruction with the two VBS3 scenarios, and then completed the three live training exercises over 1.5 days. During classroom instruction Soldiers learned how the three skill areas could improve their ability to perform TC3, and then focused on developing and applying these skills during the increasingly stressful scenarios. Squads were trained to use team self-correction in their AARs to encourage identifying and discussing skill areas needing improvement and set performance improvement goals for the next scenario. Measures of learning, self-reported stress, team perceptions, physiological responses, and observations of advanced situation awareness, teamwork, and TC3 behaviors were repeatedly collected throughout training. The same measures were collected while the four control condition squads received one day of conventional tactical training in the enhanced M2 and M3 scenarios. The two conditions were compared based on measures collected before, during, and after the M2 and M3 scenarios as reported in Johnston et al. (2019) and Patton et al. (2018b). It was expected that squads receiving SOvM training would perform better in the live scenarios than the control condition squads, and would improve their performance despite M3 having more stressful events. In the next sections we discuss research findings within the context of the Bowers et al (2017) framework.

Measurement Framework for Building Individual and Team Resilience

Individual Resilience

Table 1 lists the SOvM measures categorized as inputs, processes, emergent states, and outcomes for individuals.

Table 1. Measures of individual inputs, processes, emergent states and outcomes.

Inputs	Processes*	Emergent States	Outcomes
• Pre-training motivation • Ways of coping • Trait-based perceived stress • Physiological baseline measures for heart rate and heart rate variability • Soldier experience • Baseline self-reported skill levels • Baseline objective knowledge levels	• *Stress management* • *Controlled breathing methods* • *Social support* • *Mental simulation* • *Mindfulness*	• Situational Self Efficacy - Individual • Perceived stress • Self-reported cognitive workload • Physiological responses	• Change in self-reported skill levels • Change in objective knowledge levels

*Note: Not systematically measured in the SOvM study.

Inputs

Individual inputs are traits that mitigate the effects of stress and allow one to “bounce back” quickly following a stressor event. A positive adaptation to stress is seen as an improvement in stress responses following a stressful event, in contrast to recovery from stress or a return to the same baseline before the stressor (Raetze et al., 2021). Input measures we identified in the SOvM study were pre-training motivation, ways of coping, trait-based stress reactions, physiological measures for heart rate and heart rate variability, Soldier experience, baseline self-reported skills, and baseline objective knowledge. Patton et al. (2018b) found about the same high levels of pre-training motivation in both groups. They also found Soldiers that received the conventional tactical training scored higher on problem focused coping, had lower levels of anxiety and frustration, and expressed higher levels of risk taking.

Processes

Stress management, controlled breathing methods, social support, mental simulation, and mindfulness are thought of as effective, adaptive behavioral processes that when invoked during stressful experiences reduce negative stress responses (Bowers et al., 2017; Raetz et al., 2021). The SOvM study employed training strategies for these behaviors to enable Soldiers to adapt during the training scenarios. However, assessing observable changes in Soldier and team stress levels and responses presented significant challenges to providing feedback to squads on whether their overall stress resilience changed or improved. The first and second authors of this chapter were project leads on the SOvM study and witnessed stress and resilience subject matter experts (SMEs) reporting that while observing teams during the exercises they had heard some members trying to help manage other teammate stress responses by saying "how are you doing?", or "are you doing o.k.? or "take a knee", or "stay focused on what's important now." Generally, however, the SMEs reported they were not able to reliably observe Soldier stress responses during the exercises. During the AARs the Soldiers described how they implemented their own stress management skills, recalled and recounted their experienced stress reactions, as well as described how they managed their reactions, how other team members helped them manage, and then set personal stress management goals for the next exercise.

Individual Resilience as a second order emergent state

Bowers et al. (2017) proposed emergent team states as cognitive, motivational, and affective states that emerge from team member interactions, but did not specify individual emergent states. Nevertheless, drawing from the individual resilience research they cited, we adapted the model to include individual resilience as a second order emergent state. Perceived stress states, physiological responses, individual situational self-efficacy, and cognitive workload are typical measures of stress responses and we propose this cluster of constructs represent a second order factor of individual resilience. The SOvM findings indicated Soldiers in both conditions had a fairly resilient response to the most difficult scenarios. Patton et al. (2018b) found similar, unchanging high levels of individual self-confidence. All Soldiers experienced significantly higher levels of dysphoria (a combined measure of depression, hostility, and anxiety) and sensation seeking after both scenarios. Patton et al. (2018b) indicated these reactions were within the range of moderate stress levels based on the statistical norms established for these measures. All Soldiers reported low to moderate levels of cognitive workload after each scenario with a small, significant decrease in cognitive workload from M2 to M3. Patton et al. (2018b) reported physiological differences with the SOvM trained Soldiers experiencing significantly higher heart and respiration rates during M2, and then finding that heart rate was significantly lower and respiration rate significantly higher during M3.

Outcomes

Outcomes in the Bowers et al. (2017) model are positive indicators of resilience to include psychological and physical health, positive social interactions, and sustained cognitive ability. These appear to be distal rather than proximal measures that indicate resilience is sustained over some period of time. Measures were only collected from Soldiers after the last scenario. Change in knowledge levels were assessed with Johnston et al. (2019) finding the SOvM trained Soldiers reported significantly higher skill levels compared to their pre-training baseline and to squads in the conventional training condition, and demonstrating significantly higher knowledge levels on multiple choice tests. The squads receiving conventional training reported significantly higher skill levels, but no changes were found in their objective knowledge levels from the baseline.

Team Resilience

Table 2 lists the SOvM measures categorized as inputs, processes, emergent states, and outcomes for teams.

Table 2. Team measures as inputs, processes, emergent states and outcomes.

Inputs	Processes	Emergent States	Outcomes
• Team Tactical Expertise • Team familiarity	• Advanced Situation Awareness behaviors • Teamwork behaviors	• Situational Self Efficacy – Team • Perceived Team Efficacy • Perceived Team Cohesion • Perceived quality of team processes • Perceived quality of team performance • Perceived Shared Situation Awareness • AAR Climate (Culture) • Team Self-Correction behaviors during AAR • Team Knowledge Emergence	• Improved Tactical Performance

Inputs

Bowers et al. (2017) describe team inputs as stress mitigating factors at the team level to include trust, group norms, communication methods, membership stability, and psychological safety. Team tactical expertise and familiarity are listed as inputs assessed in the SOvM study that we considered as proxies for membership stability and group norms. Johnston et al. (2019) reported that nearly all participants in both conditions had served in their current position as an infantry Soldier an average of 7 months, with a range of having served in that position between zero and two and half years. About 75% to 80% of Soldiers reported some familiarity with other members of their squad and about two-thirds reported having had Combat Lifesaver (CLS) training.

Processes

The teamwork (initiative/leadership, backup, information exchange, proper communications) and advanced situation awareness behaviors (gathering and sharing information) assessed in the SOvM study are identified by Bowers et al. (2017) as team and organizational processes. Johnston et al. (2017) reported the SOvM training improved team processes, with the trained squads performing significantly more advanced situation awareness and teamwork tasks (up to 33% and 27% more on M3, respectively) than the control condition squads.

Team resilience as a second order emergent state

The emergent states measured in the SOvM study are a fairly good representation of what Bowers et al. (2017) had proposed. Johnston et al (2019) found Soldiers in both conditions reported similar, high levels of team situational self-efficacy, efficacy, cohesion, action processes, and performance. Situational self-efficacy was unchanged, but the other four measures showed a slight but significant increase between M2 and M3. Both conditions reported high levels of shared situation awareness that had a slight but significant increase between M2 and M3, with the experimental condition reporting significantly greater shared awareness than the control. The experimental condition teams demonstrated significantly more indicators of knowledge emergence across scenario events that increased from M2 to M3. Both groups reported high levels of positive reactions to the AAR, but the experimental condition squads exhibited significantly more team self-correction behaviors (up to 43% in the M3 AAR).

Outcomes

As noted above, the SOvM study was limited to collecting proximal outcomes from Soldiers which in this case was TC3 performance. Johnston et al. (2019) reported the experimental squads demonstrated significantly more TC3 behaviors (up to 41% more in M3) than the squads with conventional training indicating the SOvM training was effective.

Conclusions

Framing the SOvM study measures at both the individual and team levels created greater clarity in understanding measures of resilience at both levels of analysis. It was a unique opportunity to create a fairly robust "resilience readiness profile" of attitudes, cognitions, and behaviors for intact U.S. Army squads having substantial tactical experience and team familiarity. Taken together the individual and team resilience results indicated Soldiers and squads were fairly resilient. The findings can serve as normative data for comparison with other infantry squads. Additional analyses should be conducted to learn more about the relationship of baseline inputs at pre-training with the development of emergent states and outcomes; and factors in the model that distinguish Soldiers with higher levels of learned skills from those that do not. Future studies should include assessing resilience processes and post training proximal and distal outcome measures to assess how resilience is retained and lost.

Implications for GIFT and Research Recommendations

The resilience model and SOvM research findings can inform GIFT architecture developers. The following research questions should be explored:

- How should instructional strategies for developing resilience inform competency-based scenario design in GIFT?
- How should GIFT collect measures before, during, and after resilience training with competency-based scenarios?
- How does GIFT incorporate individual and team level measures of resilience to create resilience readiness profiles?

The distinction between individual and team measures identified in the SOvM study could be used as a worked example guide for developing team functionality in GIFT. GIFT is in the process of developing team tutoring functionality, which requires implementing changes and features in areas including but not limited to authoring tools, data processing, data extraction, and computer synchronization during tutoring. As team tutoring works best with real-time assessment of learner states, utilizing individual measures and understanding how they combine into/contribute to team level measures is a helpful process/exercise that can assist with real-time calculation in GIFT. By having real-time assessment of both individual and team processes in GIFT for resilience measures it can improve the system's ability to adapt in different situations. While all team scenarios are not the same, the lessons learned and ways that resilience measures are broken down can be used as a starting point for implementing both individual and team resilience measures in GIFT. There is also the possibility of standardizing outcome surveys and measures in GIFT that are in line with resilience measures at both the individual and team level. These implementations can assist in addressing the research questions above, in addition to providing tools for instructors to use as they develop their competency-based scenarios in GIFT.

References

Bowers, C., Kreutzer, C., Cannon-Bowers, J., & Lamb, J. (2017). Team resilience as a second-order emergent state: a theoretical model and research directions. *Frontiers in Psychology, 8*(1360). Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01360/full

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. 271-286). London: Ashgate.

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. In the annual Proceedings of the Interservice/Industry Training, Simulation, and Education Conference [CD-ROM], Orlando, FL. Arlington, VA: NTSA.

Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., Patton, D. J., Cox, K. G. & Fitzhugh, S. M. (2019). A team training field research study: Extending a theory of team development. *Frontiers in Psychology, 10*, 1480.

Johnston, J.H., Sottilare, R., Kalaf, M., & Goodwin, G. (2021). Training for team effectiveness under stress. In Sinatra, A.M., Graesser, A.C., Hu, X., Goldberg, B., Hampton, A.J., & Johnston, J.H. (Eds.). (2022). Design Recommendations for Intelligent Tutoring Systems: Volume 9 – Competency-Based Scenario Design. Orlando, FL: US Army Combat Capabilities Development Command – Soldier Center.

Johnston, J.H., Townsend, L.A., Gamble, K., Fitzhugh, S., Milham, L., Riddle, D., Patton, D., Phillips, H., Ross, B., Butler, P., & Wolf, R. (29 March 2017). Squad Overmatch: Phase 2 Final Report. Orlando, FL: Program Executive Simulation, Training and Instrumentation Command.

O'Keefe, G.B. (19 July 2014). Training Comprehensive Soldier and Family Fitness (Army Regulation 350-53). Arlington, VA: Department of the Army Headquarters.

Patton, D., Johnston, J., Gamble, K., Milham, L., Townsend, L., Riddle, D., & Phillips, H. (2018b). Training for Readiness and Resilience. In the Proceedings of the International Conference on Applied Human Factors and Ergonomics (pp. 292-302). Springer, Cham.

Patton, D., Townsend, L., Milham, L., Johnston, J., Riddle, D., Start, A. R., Adler, A.G., & Costello, K. (2018a). Optimizing team performance when resilience falters: An integrated training approach. In the Proceedings of the International Conference on Augmented Cognition (pp. 339-349). Springer, Cham.

Raetze, S., Duchek, S., Maynard, M. T., & Kirkman, B. L. (2021). Resilience in Organizations: An Integrative Multilevel Review and Editorial Introduction. Group & Organization Management, 46(4), 607-656.

Chapter 10 – Leveraging Lessons Learned From Synthetic Teammates for Intelligent Tutoring Systems

Christopher W. Myers
U.S. Air Force Research Laboratory

Introduction

Teams are a challenge to train, and as they get larger it becomes even more difficult. First, all team members have to be available at the time of training. Second, the difficulty in finding a time when all are available increases with team size. Indeed, in large teams it is unlikely that a date/time can be found that satisfies scheduling constraints for all participants, and thus there will be absences. When absences occur, subject matter experts or confederates can be hired to fill roles that were unable to attend, further driving up team training costs.

Synthetic teammates have been proposed as an approach to help facilitate team training, reduce associated costs, and tailor training to teams' deficiencies. Synthetic teammates, defined as intelligent systems capable of participating on a team, must work closely with team members to accomplish a common goal through deliberate coordination and have some ownership of a task, or tasks that are required to successfully complete the team's goal. Synthetic teammates may either closely approximate human cognitive processes (e.g., cognitive model), have little relationship to human cognitive processes, or some combination of the two. A potential benefit of synthetic teammates is the opportunity to use them to influence both individual and team skill acquisition (Cooke et al., 2013).

The purpose of the current chapter is to consider design recommendations and intelligent system requirements for a generalized intelligence framework for individual and team intelligent tutoring systems (ITSs). To this end, results from previously published research on the empirical evaluation of a synthetic teammate within a Remotely Piloted Aerial System (RPAS) are presented to frame design recommendations.

Method

The study in question, Myers et al. (2019) contextualizes the foundational capabilities within synthetic teammates to frame requirements and improvements in ITSs for teams. In the current section, the synthetic task environment used for conducting team experiments is described, followed by a description of a synthetic teammate capable of performing as a team member within the task environment. The section concludes with a description of a synthetic teammate evaluation study conducted to determine the efficacy of synthetic teammates on human teammate coordination and performance at the team and individual levels of analysis.

Synthetic Task Environment

The team task was an RPAS reconnaissance task that required three team members to work together to photograph ground-based targets. The RPAS–Synthetic Task Environment (RPAS-STE) has been widely used for the study of team cognition (Cooke et al., 2013; Cooke & Shope, 2004; Gorman et al., 2006;

McNeese et al., 2017; Myers et al., 2019). The RPAS-STE task was modeled after team task components from the United States Air Force Predator ground control station (Cooke & Shope, 2004). Within the RPAS-STE, three participants are assigned to the role of pilot, photographer, or navigator. Individuals are first trained on the tasks specific to their roles and then come together to work as a team to complete five 40-minute reconnaissance missions. The task requires teammates to communicate information necessary to successfully achieve the objective of photographing as many ground-based targets as possible.

Each participant is seated in front of two computer monitors that display unique task information and common vehicle information (heading, speed, altitude). Team member interaction occurs through text-based communications similar to instant messaging and email, enabling the recording of sender/receiver identities and timing. Team and individual measures have been designed, validated, and embedded in the task software (Cooke et al., 2013). To objectively determine team performance, a composite outcome score is computed for teams at the end of each 40-minute mission based on the number of targets successfully photographed and the duration of warnings and alarms incurred. Data have been collected from eight different experiments in the RPAS-STE leading to the development of a theory of interactive team cognition (Cooke et al., 2013). Consequently, the RPAS-STE provides a well-understood task for developing and objectively evaluating Autonomous Synthetic Teammates (ASTs).

Synthetic Teammate

A synthetic teammate was developed to perform the pilot's tasks and to participate as a team member (Ball et al., 2010; Rodgers et al., 2013). The evaluated synthetic teammate was developed to closely approximate humans and to determine if its human team members performed as well, better, or worse compared to all-human teams. Consequently, the synthetic teammate was developed using the Adaptive Control of Thought–Rational (ACT-R) computational cognitive architecture (Anderson, 2007)—a high-fidelity simulation of human cognitive capacities that can account for a broad range of human cognitive phenomena (see http://act-r.psy.cmu.edu/).

The ACT-R architecture is an example of a unifying computational cognitive architecture, where cognition revolves around the interaction between a central procedural system and multiple modules. There are modules for vision, audition, manual motor movement, declarative memory, and the model's current goal. Each module contains a buffer that can store one piece of information at a time. Modules are capable of massively parallel computation to obtain chunks.

The procedural memory system is a set of state-action rules, or production rules, that respond to the state of the buffer contents and act by manipulating buffer contents and/or the task environment. Only one production rule can fire at a time, and each rule within ACT-R takes 50 ms to fire. The central procedural system acts as a bottleneck, as all information passed between the buffers must go through the procedural memory system.

ACT-R is a hybrid cognitive architecture, containing symbolic and sub-symbolic systems. For example, chunks within declarative memory each represent a memory that has associated quantities calculated when a memory retrieval is attempted. One such quantity, *activation*, results from sub-symbolic computations regarding the recency and frequency of use of the chunk, its relevance to chunks in other buffers, and inherent processing noise within the memory system (Anderson, 2007). Activation values are used to determine if a chunk is retrieved from declarative memory, and if so, how long the retrieval takes, in milliseconds. The chunk with the highest activation above a predefined threshold is the one that is retrieved.

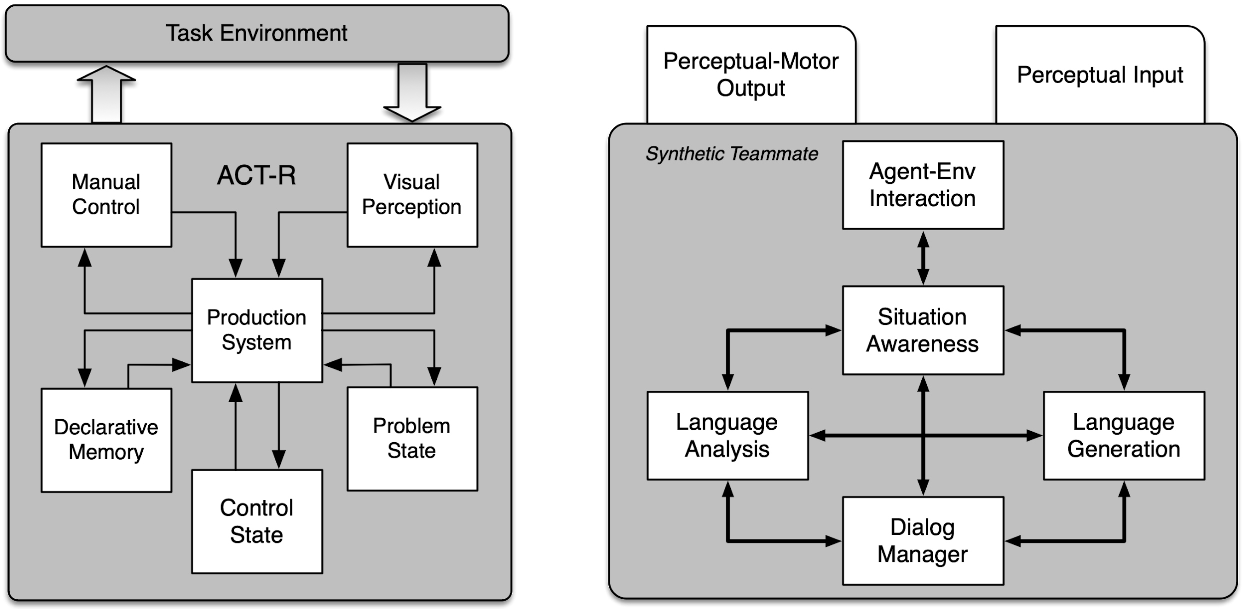


Figure 1. The ACT-R architecture and its modules (left) and the synthetic teammate capacities (right) developed using the ACT-R architecture.

The synthetic teammate was developed within the ACT-R architecture and contains a set of cognitive systems enabling it to operate as a teammate within the RPAS-STE. The integrated systems include language comprehension and generation, agent-environment interaction, situation awareness, and dialog management (Figure 1). Each system within the synthetic teammate uses ACT-R's procedural and declarative memory systems. In addition, the agent-environment interaction component uses ACT-R's perceptual-motor capabilities to interact with the synthetic task environment and passes visually encoded information on to the other components. For example, the agent-environment interaction component uses ACT-R's perceptual module to attend to visual objects in the RPAS-STE, its declarative memory to identify the type of object that has been attended, its production system to determine what action to take given the encoded object, and its motor system to produce an action in the RPAS-STE (e.g., keyboard input, mouse movement, etc.). Once the synthetic teammate reached a point in its development that it could complete multiple 40-minute missions, it was integrated as the RPAS-STE pilot to evaluate team and individual performance.

Empirical Evaluation

To determine if the synthetic teammate could provide enough task skill and communication capabilities to facilitate behavior at the team and individual levels of analysis, we manipulated team composition using three between subjects conditions: synthetic, experimenter, and control (Myers et al., 2019). The control condition was a task-naive all-human team. The experimenter condition had an expert human serve as the RPAS-STE pilot with task-naive human photographers and navigators. The synthetic condition had the synthetic teammate performing as the pilot, also with task-naive human photographers and navigators.

In the experimenter condition, the expert pilot focused on effective coordination of information within the team. The role and instruction given to the pilot in this condition were the same as the other two conditions, the only difference being they were experienced at coordinating task-specific information within and among the team. Specifically, the pilot in this condition would push and pull information among the team members if information was not given after a set amount of time, or if it was not forthcoming. To ensure that coordination occurred in a structured and routine manner across all teams, the pilot used a coordination script. This script consisted of *if–then* statements dependent on the task itself. For example: *If* the photographer does not request a certain airspeed or altitude of a reconnaissance target within one minute, *then* the pilot asks if the current speed and altitude are correct. The increased reliability of pushing and pulling information throughout the team in this condition is hypothesized to increase team performance. In addition, the experimenter condition provided a high-level benchmark for how an extremely high performing and *expert* AST should perform for training advanced teams. Where the AST pilot performance differs from a novice pilot provides an opportunity to better understand the weaknesses of the AST for focused improvement.

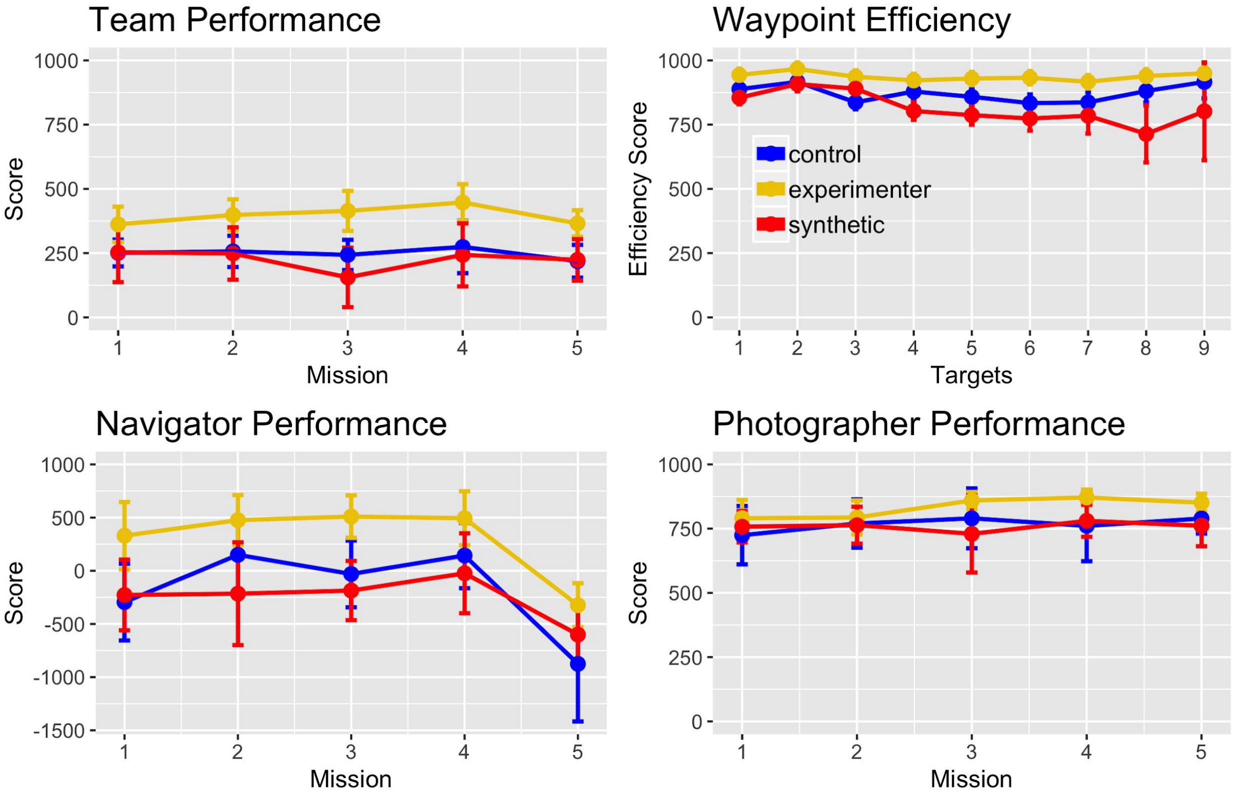


Figure 2. Synthetic Teammate empirical evaluation results at the team (i.e., team performance, waypoint efficiency) and task (i.e., navigator and photographer performance) levels of analysis.

Individuals were randomly assigned to form teams of three and then randomly assigned to each condition. Each team completed five unique missions, with the last mission being one of high cognitive workload (many more targets than missions 1-4). There were ten teams per team composition condition. To objectively determine the effects of the AST, performance was compared between conditions across individual and team reconnaissance tasks. We first present results from the team.

Results

Teams were evaluated on their team-level performance, the speed with which the team photographed waypoints upon arriving at them (i.e., waypoint efficiency), and their individual task performance (navigation and photography). While other analyses have been conducted and are certainly important, these analyses have been selected to demonstrate challenges with developing flexible, reliable, and trainable high-cognitive-fidelity instructors, tutors, and training teammates within complex systems and tasks.

There are three important takeaways from the empirical evaluation of the synthetic teammate. First, teams with the synthetic teammate performed as well as the control teams, yet not as well as the experimenter teams (see Figure 2, top-left plot). Second, teams with the synthetic teammate were not nearly as efficient as all-human teams considered together (control and experimenter) at obtaining photographs at waypoints (see Figure 2, top-right plot), decreasing in performance with increasing numbers of visited targets. Third, teammates that completed missions with the synthetic teammate (i.e., navigators and photographers) performed their tasks at a level of performance statistically indistinguishable from navigators and photographers who worked with naive human pilots (i.e., the control condition; see Figure 2, bottom plots). This is an important demonstration of a synthetic teammate supplanting a human operator and maintaining team and individual effectiveness. While this is a first and necessary step, we must next directly address the question of training by testing navigator and photographer in all-human teams after learning the task with an AST.

Discussion and Lessons Learned

The purpose of the research on synthetic teammates was to empirically determine their effects on human team members and what cognitive capacities were necessary for participating as a team member. While synthetic teammates have had some success (Myers et al., 2019), there remain challenges (McNeese et al., 2017). Results demonstrated that a synthetic teammate does not necessarily prevent its human counterparts from reaching the same level of performance as novice all-human teams (i.e., control). However, there is still much room for improvement if you compare teams with the synthetic teammate to teams with the experimenter participating as an expert pilot (synthetic vs. experimenter conditions in bottom plots of Figure 2). Indeed, identifying and correcting the gaps within the synthetic teammate will contribute to the development of a generalized intelligent framework for tutoring.

Based on the research, development, and evaluation of the synthetic teammate, there are three areas where their capabilities could be improved significantly. First, synthetic teammates must be capable of adapting their communication patterns to their team members to facilitate the establishment, maintenance, and repair of common ground as well as the adoption and use of novel terms and abbreviations introduced by team members (Clark & Wilkes-Gibbs, 1986). The ability to communicate information to team members in a settled on and expected structure has been theorized to facilitate information processing efficiency, and there appears to be a connection between the ability to adapt communications and the acquisition of skill at the team level of analysis (Bibyk et al., 2021).

Second, the development costs for synthetic teammates are too high to quickly develop and integrate synthetic teammates into complex tasks. The synthetic teammate discussed in this chapter was initiated in 2007 and its empirical evaluation completed in 2015, and was developed by a team of six technical experts and supported by dozens of others (e.g., empirical evaluation data collection, data analyses, etc.; Ball et al., 2010; McNeese et al., 2017; Myers et al., 2019; Rodgers et al., 2013). This is simply too costly in time and resources. While the achievement of the synthetic teammate demonstrates the art of the possible, we must now determine approaches that minimize the time for their development and reduce the requirements of technical expertise to improve development times from eight years to eight days.

Finally, synthetic teammates will never have all of the relevant knowledge they need to perform their specified tasks. Try as a development team might, some information important to completing the specified task or interacting with teammates will be overlooked. This is unavoidable and a consequence of the knowledge engineering bottleneck (Feigenbaum, 1980). To minimize the effects of absent information on task and team performance, the research and development of cognitive processes to detect and resolve gaps in requisite task and team interaction knowledge needs to be conducted (Bajaj et al., 2021; Schmidt, 2020). The integration of such metacognitive processes would facilitate independent knowledge acquisition capabilities and help to mitigate synthetic teammate frailties based on insufficient task and world knowledge.

Recommendations and Future Research

Based on the lessons learned, three areas for future research and capabilities within the Generalized Intelligent Framework for Tutoring (GIFT) are recommended. First, any ITS based on GIFT that must communicate with human trainees through natural language must be capable of establishing, maintaining, and repairing common ground between the ITS and the human trainee. This capability will facilitate task and team skill acquisition.

The second recommendation for research and GIFT capabilities revolves around reducing the development costs of ITSs. One potential approach is to develop GIFT-based ITSs through demonstration and instruction (Kirk & Laird, 2014; Laird et al., 2017). Another very similar approach is to provide GIFT-based ITSs with a set of written instructions on how to accomplish both taskwork and teamwork (Eberhart et al., 2020; Kupitz et al., 2021; Salvucci, 2021). The common goals across each of these approaches is a reduction in human knowledge engineering requirements, the large teams of experts for creating the synthetic teammates, as well as their development times. The ability for GIFT-based ITSs to learn new tasks through instruction for which they will be tutoring trainees will help to reduce their brittleness, improve their generality, and reduce their costs.

Finally, methods and processes for detecting and identifying gaps in a systems knowledge could be leveraged in GIFT-based ITSs to help identify such knowledge gaps in trainees' knowledge. Indeed, such abilities are already available in ITSs, but are mostly based on knowledge gap identification through observable behavior (actions applied to graphical user interfaces, eye movement sequences, etc.); however, tracking trainee knowledge gaps based on other sources of information, such as through ITS-trainee communications and observable behavior would improve the ability to identify potential causes of trainee failures and resolve the gaps in knowledge that led to them.

Conclusions

The current chapter covered empirical research on the development of synthetic teammates for team training. Through the results of the empirical research, weaknesses within the synthetic teammate were identified and presented. The proposed research to overcome these failures in synthetic teammates are also applicable to GIFT-based ITSs.

References

Anderson, J. R. (2007). How can the human mind exist in the physical universe? In F. E. Ritter (Series Ed.), *Oxford series on cognitive models and architectures*. Oxford University Press.

Bajaj, G., Current, S., Schmidt, D., Bandyopadhyay, B., Myers, C. W., & Parthasarathy, S. (2021). Knowledge gaps: A challenge for agent-based automatic task completion [Manuscript submitted for publication]. The Ohio State University.

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., & Rodgers, S. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, *16*(3), 271–299. https://doi.org/10.1007/s10588-010-9065-3

Bibyk, S. A., Blaha, L. M., & Myers, C. W. (2021). How packaging of information in conversation is impacted by communication medium and restrictions. *Frontiers in Psychology*, *12*(April), 1–19. https://doi.org/10.3389/fpsyg.2021.594255

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, l-39. Retrieved from https://pdfs.semanticscholar.org/dd2b/dd2c4df589cc3be1f4bfab6c42d8a9dc6609.pdf

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*, 255–285. https://doi.org/10.1111/cogs.12009

Cooke, N. J., & Shope, S. M. (2004). Designing a synthetic task environment. In S. G. Schiflett, L. R. Elliott, E.

Salas, & M. D. Coovert (Eds.), *Scaled worlds: Development, validation, and application* (pp. 263–278). Surrey, England: Ashgate

Eberhart, A., Shimizu, C., Stevens, C., Hitzler, P., Myers, C. W., & Maruyama, B. (2020). A domain ontology for task instructions. In B. Villazón-Terrazas, F. Ortiz-Rodríguezm, S. M. Tiwari, & S. K. Shandilya (Eds.), *Knowledge Graphs and Semantic Web. Second Iberoamerican Conference and First Indo-American Conference, KGSWC 2020* (pp. 1–13). Mérida, Mexico: Communications in Computer and Information Science, Vol. 1232.

Feigenbaum, E. A (1980). *Knowledge engineering: The applied side of artificial intelligence* (STAN-CS 80-812). Stanford University, CA: Computer Science Department.

Gorman, J. C., Cooke, N. J., Pedersen, H. K., Winner, J., Andrews, D., & Amazeen, P. G. (2006). Changes in team composition after a break: Building adaptive command-and-control teams. In *Proceedings of the Human Factors and Ergonomics Society* (Vol. 50, pp. 487–491). https://doi.org/10.1177/154193120605000358

Kirk, J. R., & Laird, J. E. (2014). Interactive task learning for simple games. *Advances in Cognitive Systems*, *3*, 13–30.

Kupitz, C., Eberhart, A., Schmidt, D., Stevens, C., Shimizu, C., Hitzler, P., … Myers, C. W. (2021). Toward undifferentiated cognitive models. In the *Proceedings of the International Conference on Cognitive Modeling & Society for Mathematical Psychology* (pp. 1–6). Virtual.

Laird, J. E., Gluck, K., Anderson, J., Forbus, K. D., Jenkins, O. C., Lebiere, C., … Kirk, J. R. (2017). Interactive task learning. *IEEE Intelligent Systems*, *32*(4), 6–21. https://doi.org/10.1109/MIS.2017.3121552

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2017). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60,* 262–273

Myers, C., Ball, J., Cooke, N., Freiman, M., Caisse, M., Rodgers, S., … Mcneese, N. (2019). Autonomous intelligent agents for team training: Making the case for synthetic teammates. *IEEE Intelligent Systems*, *34*, 3–14.

Rodgers, S., Myers, C., Ball, J., & Freiman, M. (2013). Toward a situation model in a cognitive architecture. *Computational and Mathematical Organization Theory*, *19*, 313–345.

Salvucci, D.D. (2021), Interactive grounding and inference in learning by instruction. *Topics in Cognitive Science*, *13,* 488-498. https://doi.org/10.1111/tops.12535

Schmidt, D. P. (2020). *Identifying knowledge gaps using a graph-based knowledge representation* [Master's Thesis, Wright-State University]. Retrieved from https://etd.ohiolink.edu/apexprod/rws_olink/r/1501/10?p10_etd_subid=185367&clear=10

Chapter 11– The Impact of Personalized Feedback on Negotiation Training

Emmanuel Johnson and Jonathan Gratch
Institute for Creative Technologies, University of Southern California

Introduction

Intelligent tutoring systems (ITSs) have made great strides in teaching cognitive skills, including math (Koedinger et al., 1997; Koedinger & Corbett, 2005; Koedinger & Corbett, 2006), reading (Mills-Tettey, et al., 2009; Wijekumar et al., 2005;) and computer literacy (Guo, 2015; Olney et al., 2017;). Recent research has begun to extend these techniques to interpersonal skills such as public speaking (Chollet et al., 2014), medical interviews (Pataki, 2012; Stevens, 2006), collaborative problem solving (Graesser et al., 2018) and negotiation (Gratch et al., 2016; Kim et al., 2009). An extensive body of research has documented the benefits of ITSs for cognitive skill development, but relative to this, research on ITSs for interpersonal skills is still in its infancy. This chapter highlights our efforts in adapting ITS techniques to teaching negotiation.

Negotiation is a crucial interpersonal skill. Students entering the modern workforce must successfully interview, negotiate their salaries and job responsibilities, work in teams, resolve conflicts and solve problems creatively and collaboratively. Such interpersonal skills are rarely assessed or taught in the classroom, and little research has explored the potential for automated tutoring of these foundational social skills. The US Academy of Sciences and the World Economic Forum identify negotiation as a foundational social skill essential for the future of work through its impact on organizational creativity and productivity (Forum, 2016; National Academies of Sciences, 2016). Deficits in negotiation ability contribute to the underrepresentation and lack of advancement of women and minorities in STEM fields (Goldman, 2012; Morela Hernandez, 2016).

Just as with cognitive skills like math, social skills are taught with a mixture of instruction, practice exercises and feedback. But as interpersonal skills are fundamentally social, practice typically involves pairing up and practicing skills with other students. Typically, negotiation skills are taught through a mixture of formal and experiential learning. After a lecture introducing some specific negotiation principles, students form pairs or small groups and practice through loosely-structured roleplaying exercises. As the students negotiate with one another in these exercises, the instructor walks around the room, observing and evaluating how well students apply the principles that were just introduced. Afterwards, an instructor might initiate general class discussions highlighting specific individuals' successes and failures. This use of negotiation exercises aligns with an experiential theory perspective, suggesting that learning occurs when students are able to practice and then reflect upon their performance (Kolb & Kolb, 2012), especially compared to best practices. An important limitation to this common method of instruction is that, although the lecture-then-exercise format encourages student practice, little emphasis is placed on personalized feedback. Instructors' attention is limited, and they do not have enough bandwidth to evaluate all students' use of the negotiation principles, especially in large classes. This is problematic as receiving constructive personalized feedback is integral to skill development (Hattie & Timperley, 2007; Johnson et al., 2017). Furthermore, when attempting to reflect on their negotiation skills without such personalized feedback, students might try to improve their skills through observing others, which may lead to imitating suboptimal strategies (i.e., the blind leading the blind). Finally, these courses are labor intensive and quite expensive. Two of the most well-known programs are the Harvard Program on Negotiation and Northwestern University's Conflict Resolution Institute training program. Programs such as these can range anywhere from $3000 to $10,000 for a two to four-week course, a sum the average American cannot afford (Durante et al., 2017).

ITSs have the potential to address these concerns, lower costs, and increase access to negotiation training, however the key challenge is understanding and assessing the rich spoken communication that students use during their peer-exercises. One solution, and the one we advocate in this chapter, is the development of "virtual role-players" that allow students to practice negotiation skills with a computer-controlled partner. These are computer-controlled agents that can serve as a credible negotiation opponent, allow students to exercise classroom concepts, assess their mastery of key concepts, and provide the basis for providing personalized feedback. Virtual role-players have several potential advantages over peer partners. Computer agents can provide a more consistent partner that is designed to evoke "teachable moments" (Kapur, 2008; VanLehn, 2003). They can be instrumented to capture important aspects of student behavior. They have also been shown to reduce some of the social anxiety that can serve as a barrier to enrolling in training (Gratch et al., 2016).

Mere practice with a virtual negotiation partner can improve skills (Gratch et al., 2016; Lin et al., 2009), but for systems to be most effective as learning tools, they must both allow negotiators to practice *and* provide personalized feedback. This feedback is important in helping participants to reflect on (and improve) their performance. Accordingly, in the current work, we extend one of these previous systems (Gratch et al., 2016) to provide personalized feedback based on whether students adhere to principles of good negotiation. We argue that this feedback should be grounded in well-established negotiation principles. Therefore, participants first negotiated with a virtual human negotiation partner, and then were provided personalized feedback using negotiation principles established by Kelley (Kelley, 1966). Importantly, these principles have been quantified through automated methods (Johnson et al., 2017). Here, we take the important next step: we empirically test the impact of providing students such automated feedback about their negotiation skills.

In this chapter, we describe an approach to negotiation which borrows methods from Cognitive Tutoring (Koedinger & Corbett, 2005; Koedinger & Corbett, 2006), an ITS method that utilizes a cognitive model to diagnose student actions and inform feedback on the correctness or incorrectness of their choices. We then outline two studies that demonstrate the potential of this approach. The first proof-of-concept study explores the viability of this approach in a spoken-language negotiation using a "puppeted" opponent (a virtual character controlled through a Wizard-of-Oz setup interface). The second study demonstrates the successes of the approach with a fully-automated web-based negotiation agent. Each study allows students to practice their skills on a conventional negotiation exercise, during which we can measure students' negotiation skills using established automated methods (Johnson et al., 2017). Using this design, we can empirically assess whether or not this feedback impacted their negotiation skills, and outcomes, in the subsequent negotiation. Specifically, we hypothesized that 1) personalized feedback can be used to cultivate key negotiation skills in subsequent negotiations, as measured by automatic metrics (Johnson et al., 2017), and 2) personalized feedback will help students to obtain better outcomes in negotiations after receiving it (compared to before). This chapter is structured as follows: We begin with an overview of cognitive tutors and model-based diagnosis. We next review some general principles of negotiation that inform our model of ideal negotiator behavior. From there we show how these principles can be quantified and measured by an automated system. Next we provide the study design and methods used to assess the impact of feedback. We then present our results followed by discussions and recommendations.

Model-Based Diagnosis of Negotiation Skills

Cognitive Tutors (Koedinger & Corbett, 2005; Koedinger & Corbett, 2006; Koedinger et al., 1997) are a common ITS approach used to teach "hard" skills like math or physics. The key idea is that the tutor has cognitive models of the skill to be learned. The tutor includes both models of the skill that is ideally executed by students, but also models of common mistakes. For example, in the domain of simple arithmetic, a cognitive tutor might include a model of the steps students should follow to perform column addition. But the tutor will also include models of comment "buggy" models, such as forgetting to carry a digit to the next column. Cognitive tutors perform model-based diagnosis over student solutions to identify these possible bugs. If students execute the correct procedure, they receive positive feedback. But if students return incorrect answers, cognitive tutors can identify the specific misconception and provide targeted and personalized feedback. Our work seeks to extend this basic idea to the realm of negotiation skills.

Research into negotiation has helped to identify a number of principles that can help individuals achieve better outcomes. . To date, a set of principles that guarantee successful negotiation outcome has not been established. However, the principles proposed (Galinsky & Mussweiler, 2001; Gratch et al., 2015; Kelley, 1966) seem to be robust enough to be used as good indicator of negotiation success. Kelley (1966) found that good negotiators do a number of things that correlated with positive negotiation outcomes; they avoid early commitment, make efficient concessions, induce their opponent to concede, shape their opponent perception of value, and gather information in advance of the negotiation. Johnson and colleagues showed that these principles can be quantified (Johnson et al., 2017) and measured in a negotiation. Here we focus on two broad sets of principles, value creating and value claiming. Value-creating tactics provide ways to "grow the pie" in ways that allows both parties to obtain more of what they want (i.e., to maximize the joint-value of the negotiated outcome). Thus, negotiation training provides a basis for avoiding the presumption that the pie is fixed (i.e., the fixed pie bias) and helping students learn ways not only to claim value but also to create value. Several biases undermine a student's ability to create value. Novices often assume their interests are directly opposed to the interests of their opponent and thus simply assume that they should split each issue down the middle. Negotiation courses teach several value-creating tactics to overcome fixed-pie perceptions. For example, students are encouraged to exchange information (e.g., asking their opponent about their preferences over different issues and revealing their own preferences in exchange). Students are also encouraged to make offers that explore tradeoffs across issues. For example, in a tactic known as logrolling, students are encouraged to claim more of their highest priority issue in exchange for concessions on less important issues. Negotiation courses also teach specific value-claiming tactics to help students feel comfortable with setting ambitious aspirations for the outcome. One common tactic is anchoring. Anchoring refers to the idea that a negotiator should start with a strong initial offer, and this tends to "anchor" concession-making around this point. Research shows that when negotiators make a high initial offer, they frequently obtain better final outcomes (Galinsky & Mussweiler, 2001). Other tactics include making full use of time (i.e., not conceding too early) and communicating a willingness to abandon the negotiation.

To have the greatest impact, automated assessment should operate in a way that is independent of the specific scenario or algorithm a student uses to practice. Also, this assessment should follow common approaches used in the classroom. Typical feedback in a negotiation classroom can be divided into two categories: value claiming and value creating. Value claiming feedback provides insight on the negotiator's ability to gain more value in a negotiation. Value creating feedback on the other hand, measures the extent to which a negotiator gathered information about an opponent and is able to use that to gain more value in the negotiation. To accomplish our goal of providing automated feedback, we mapped the metrics highlighted above onto the value claiming and value creating feedback framework. From this, we derive outcome measures (i.e., was the final deal successful at creating and claiming value), and also process measures that assess the extent to which students used tactics that create and claim value.

Value Claiming

A student's ability to claim value was assessed by measuring the individual points they obtained in the final deal. Another process measure was used to gain insight into why they may have failed to claim value.

Specifically, the point value of the student's initial offer was examined. The assessed metrics are input into a decision-tree that chooses feedback to provide to students. The feedback is based on instructor-crafted templates that include slots filled by the automatic assessments. When students achieve good outcomes or follow recommended tactics, this is positively reinforced (e.g., “The first offer you made would have gotten you about 76% of the points. Pretty good”) and the principle emphasized (“By claiming most of what you want early in the negotiation, you can manage your negotiation partner's expectations of what they will receive”).

Value Creation

A student's ability to create value was assessed by measuring the joint points achieved in the negotiated agreement (i.e., the points obtained by both the student and the agent). Several process measures were assessed to gain insight into why a student may have failed to create value. For example, we assessed student ability to employ the tactic of logrolling by the extent to which they made tradeoffs in their initial offer to the agent (specifically, the number of highest-value items they claimed minus the number of lowest-value items they offered), the amount of information exchanged between a negotiator and their opponent by the amount of questions asked and responded to and how well a negotiator understood their opponent’s preferences .The “inefficiency" of the student's final offer was measured by the offer's distance from the Pareto frontier (the set of deals in which neither the agent nor the participant could have done better without the other doing worse). This is essentially a measure of how much value is left on the table.

Methods

In order to teach value creation and value claiming, we ran two randomized between subject studies to contrast personalized feedback with generic and or no feedback (i.e., “mere practice”). In the first study students interacted with a Wizard-of-Oz system, the Conflict Resolution Agent. This system is depicted in Figure 1, which has an example of the agent. In the second study they interacted with an “off-the-shelf” negotiation agent (depicted in Figure 2). Students did an initial negotiation, received the experimental treatment (personalized v. generic v. no feedback) and then performed a second negotiation to assess any improvements.

Study One

Participants

We recruited 63 participants (34 females) through Craigslist; and they were compensated $30 for their participation. Technical failures resulted in unusable data for three participants, therefore the below analyses were conducted on data from the remaining 60 participants (30 per condition). In addition to base pay for participation, participants were incentivized to perform well in the negotiation by entering them into a $100 lottery based on how much they got in the negotiation.

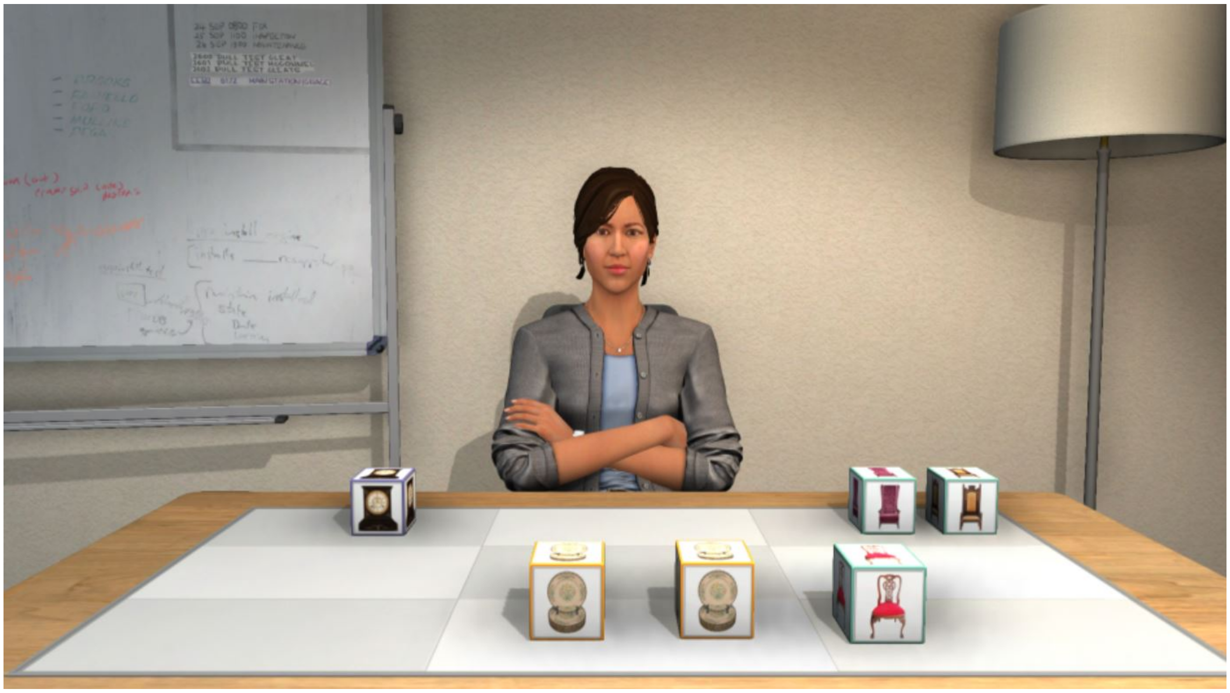


Figure 1. Conflict Resolution Agent

Study Design

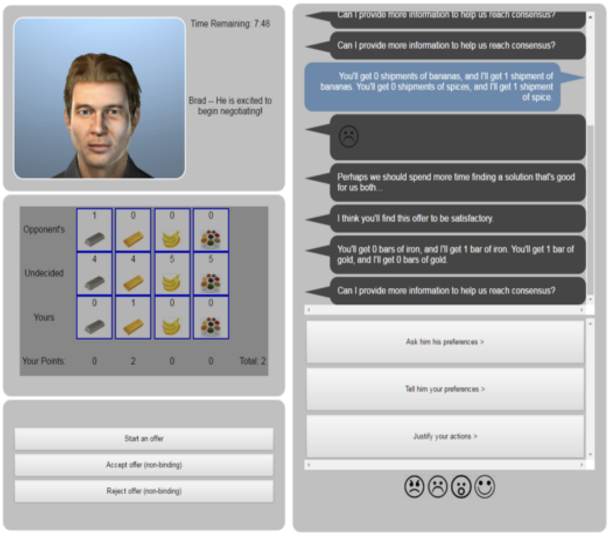
In this study, participants completed two negotiations with the Conflict Resolution Agent (CRA). They were randomly assigned to either the feedback (experimental) or control conditions. Participants in the experimental condition received personalized feedback on their first negotiation performance prior to the second negotiation beginning. Participants in the control condition did not receive feedback. Instead, after receiving their negotiation score following the first round, controls were told to just reflect on the negotiation for five minutes.

Negotiation Task

In each of the two negotiations, participants were asked to role-play as an antique salesperson participating in a multi-issue bargaining task. In this task, there were six items to negotiate within each negotiation, and they had up to ten minutes to negotiate with the agent over how to divide a collection of antique items between them. In the first negotiation, these items included three records, two lamps, and a painting. In the second negotiation the items were changed to chairs, plates, and a clock respectively to prevent the participant from knowing the agent's preferences for the items before the negotiation had begun. However, these three item types were direct analogs to the original items in terms of value. For simplicity, we will refer only to the original item types (records, lamps, and painting). The goal of each negotiator in this task was to reach an agreement that afforded them the highest total value in received items. Each type of item had a set value to the participant and agent. For both players, each of the records was worth 30 points and each of the lamps was worth 15 points. This was designed to be a distributive negotiation; thus, items were generally of equal value to both negotiators. The painting was the only item that held a different value to the participant: it was worth 5 points to the participant but had no value to the agent. Participants could thus discover that the painting could be claimed without consequence, as it had no value to the agent. Although all participants reached agreement, they were told that if they failed to reach agreement within the 10-minute limit (or chose to walk away from the negotiation), they would receive one of their highest priority items as an alternative to negotiated agreement. Thus, the negotiation outcome for the participant ranged from 30 to 125. Prior to each negotiation, participants were given a description of the task, which gave the relative value of each item. Specifically, they were told that each of the records was worth at least twice as much as one lamp, and that the painting was the least valuable item. They were told that they would earn tickets toward the lottery for $100 based on how many (and which) items they acquired in the negotiation. Participants were then given a short quiz to verify that they understood the negotiation task, their priorities for the different item types, and their real-world incentive to do well in the negotiation (entries into a lottery for $100). The negotiation began once the quiz was checked, and any misunderstandings resolved.

Study Two

One of the limitations of the previous study is that users did not interact with a fully automated system. Our next step was to create a task agnostic feedback system that was fully automatic and verify these promising findings still hold. Thus, we adapted our metrics and feedback system to work with an off the shelf automated negotiation platform seen in Figure 2 (the IAGO platform (Mell & Gratch, 2016)).


Time Remaining: 7:48
Brad -- He is excited to begin negotiating!
Opponent's
Undecided
Yours
Your Points:
Total: 2
Start an offer
Accept offer (non-binding)
Reject offer (non-binding)
Ask him his preferences >
Tell him your preferences >
Justify your actions >

Can I provide more information to help us reach consensus?

Can I provide more information to help us reach consensus?

You'll get 0 shipments of bananas, and I'll get 1 shipment of bananas. You'll get 0 shipments of spices, and I'll get 1 shipment of spice.

☹

Perhaps we should spend more time finding a solution that's good for us both...

I think you'll find this offer to be satisfactory.

You'll get 0 bars of iron, and I'll get 1 bar of iron. You'll get 1 bar of gold, and I'll get 0 bars of gold.

Can I provide more information to help us reach consensus?

Opponent's	1	0	0	0	
Undecided	4	4	5	5	
Yours	0	1	0	0	
Your Points:	0	2	0	0	Total: 2

Negotiation 1	Gold	Iron	Bananas	Spices
Agent Points	1	2	3	4
Student Points	4	3	2	1

Negotiation 2	Records	Clocks	Painting	Lamps
Agent Points	2	3	1	4
Student Points	3	2	4	1

Figure 2. The left image illustrates the IAGO agent interface. The tables on the right illustrate the issues and payoff for the two negotiations.

Participants

English speaking American participants (*n*=120) were recruited via Mechanical Turk following standard experimental practices. To motivate their performance, participants were paid $3/hour for their participation in the study and entered into a lottery to win a prize of $10. Of these participants, 19 were excluded from analysis (9 failed the attention check and 10 failed to reach an agreement or experienced software failure).

Study Design

Participants negotiated using the IAGO online negotiation platform (Mell & Gratch, 2016). This platform allows students to practice negotiation with a number of possible agents. IAGO is designed to support basic tactics that expert negotiators used to create and claim value. Negotiators can exchange offers but also information (do you like A more than B?) and send other messages such as threats. The platform also provides tools to customize agent behavior including the ability to incorporate common biases shown by negotiators (such as the fixed-pie bias). It has been used by a number of researchers to build human-like negotiating agents (Johnson et al., 2019; Roediger, 2018). Student behavior in a negotiation is heavily influenced by the skill of their opponent. To better simulate the experience of a novice student, we adopted an existing IAGO agent that incorporates several common biases found in novice negotiators. The agent incorporates some behaviors that were shown to undermine value creation. Specifically, it adopts a fixed-pie bias" (it assumes it is fighting over how to divide a fixed-sized pie) and is not motivated to exchange information unless the student initiates the exchange (but it will use preference information if the student provides it). It also incorporates behaviors known to undermine value creation. Specifically, the agent employs anchoring (it makes a strong initial offer). Finally, the agent adopts a fair concession strategy. After its initial anchor, it responds to user offers by adjusting them towards a fair split (based on whatever knowledge it has about the student's preferences).

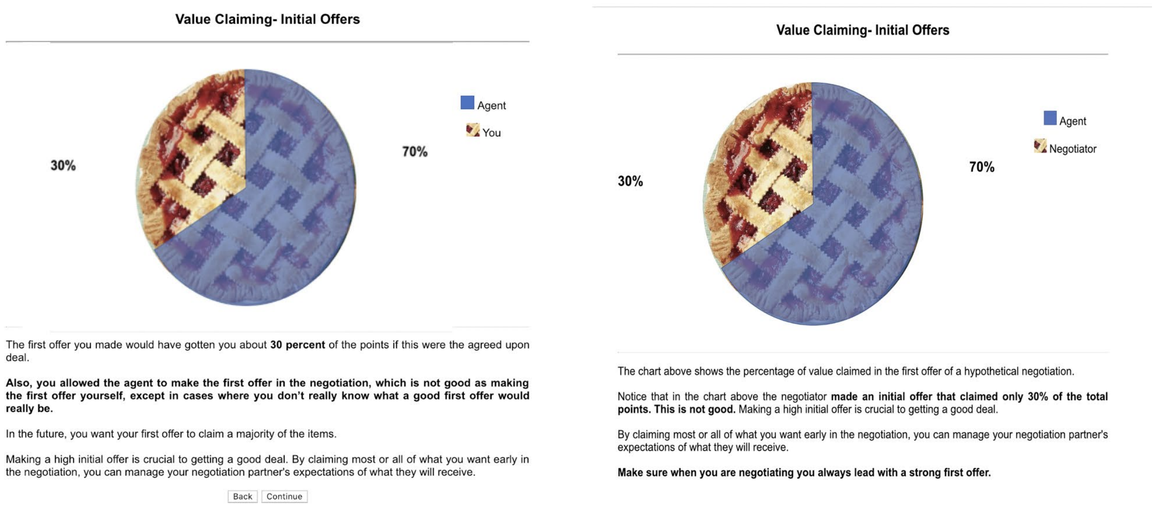
Negotiation Task

Participants were asked to engage in two negotiations. Each had the same mathematical structure (a 4-issue, 6-level multi-issue bargaining task) but used a different cover story and a different ordering of the issues to obscure this similarity. The tasks were framed as a negotiation between antique dealers on dividing the contents of an abandoned storage locker. The first negotiation involved splitting 5 bars of gold, 5 bars of iron, 5 shipments of spices, and 5 shipments of bananas. The second involved 5 clocks, 5 records, 5 paintings and 5 lamps. Both the agents and participants had distinct preferences across the items, and neither the agent nor the participant knew the other's preference. The structure of these points ensures that parties can create value by making tradeoffs between items (e.g., in the first negotiation, the player can create value by taking all the gold and iron and offering all the spices and bananas). Prior to each negotiation, participants were told how much each item was worth to them. In addition to the worth of items, participants were also told they would receive only 4 points if they failed to reach an agreement.

Experimental Manipulation

Participants were randomly assigned to one of three experimental conditions. In the Personalized Feedback condition participants were provided personalized feedback on their initial offer, understanding of their opponent's preferences and the overall value of their final claim. Prior to providing the participants with feedback, the system asked participants if they remembered their preference as well as their opponents and this information was logged and used to guide the personalized feedback. In the Generic Feedback condition participants received feedback using the same templates as with personalized feedback, but whereas slots were filled in with the student's own behavior for personalized feedback, in generic feedback they were filled using information from the same “generic” student and described as feedback on a hypothetical negotiation. For example, participants were shown the initial offer, final deal and information exchanged from a hypothetical negotiation. They were provided suggestions on how good that person did and how their results could have been improved. In the No Feedback condition participants were told the points they received but provided no other information.

Figure 3 shows an example of personalized value-claiming feedback. This is contrasted with generic feedback (feedback a student might receive if personalization was not available). In this example, both the personalized and generic feedback are provided for a participant who has made an initial offer which claims 30% of the total points. If this was indeed their initial offer, then the personalized feedback would indicate that. Regardless of their initial offer, the generic feedback uses this single example to illustrate a poor first offer. It does not take the users initial offer into account.



The first offer you made would have gotten you about **30 percent** of the points if this were the agreed upon deal.

Also, you allowed the agent to make the first offer in the negotiation, which is not good as making the first offer yourself, except in cases where you don’t really know what a good first offer would really be.

In the future, you want your first offer to claim a majority of the items.

Making a high initial offer is crucial to getting a good deal. By claiming most or all of what you want early in the negotiation, you can manage your negotiation partner's expectations of what they will receive.

The chart above shows the percentage of value claimed in the first offer of a hypothetical negotiation.

Notice that in the chart above the negotiator **made an initial offer that claimed only 30% of the total points. This is not good.** Making a high initial offer is crucial to getting a good deal.

By claiming most or all of what you want early in the negotiation, you can manage your negotiation partner's expectations of what they will receive.

Make sure when you are negotiating you always lead with a strong first offer.

Personalized Feedback

Generic Feedback

Figure 3. Example of Personalized (left) and Generic (right) Feedback

Results

Study One

Participants' negotiation metrics in the negotiation were analyzed by performing 2 (feedback: personalized feedback versus no feedback) x 2 (time: negotiation 1 versus negotiation 2) mixed ANOVAs. First, metrics around value claiming: strength of initial offer, use of available time, and total value claimed (total across negotiation and average for any given offer) were analyzed. Analysis of the initial offer revealed that a significant main effect of time (F (1,47) = 9.10, p =.004) was qualified by feedback condition (F (1,47) = 8.26, p = .006). As depicted in Figure 4 (left), participants who received personalized feedback made stronger initial offers in the second negotiation (M = 95.42, SE = 2.91) than the first (M = 78.96, SE = 2.97; F (1,23) = 15.07, p = .001), but there was no difference in the control condition (M = 80.60, SE = 2.91 vs. M = 80.20, SE = 2.91; F (1,24) = 0.01, p = .91). Next, value claimed was analyzed across the negotiation. Again, there was a main effect of time (F (1,48) = 7.37, p = .009), which was qualified by feedback condition (F (1,48) = 3.92, p = .05). As shown in Figure 4 (right), those who received personalized feedback tried to claim more total value in the second negotiation (M = 476.80, SE = 39.58) than the first (M = 322.00, SE = 31.25; F (1,24) = 8.89, p = .006), but no difference was found in the control group (M = 367.69, SE = 39.58 vs. M = 343.40, SE = 31.25; F (1,24) = 0.35, p = .56).

To test whether the effect found for total value claimed could be found at any given point during the negotiation, we analyzed average value claimed during each offer. This also revealed a main effect of time (F (1,47) = 24.79, p < .001), which was qualified by the feedback condition (F (1,47) = 7.14, p = .01). Figure 5 (left) shows only marginally more value was claimed by the control group in the second negotiation (M = 73.37, SE = 2.00) than the first (M=69.90, SE=1.65; F (1,24) = 3.38, p = .08) but these results were not significant. Participants who received feedback between negotiations made higher average claims in the second negotiation (M=82.42, SE=2.05) than the first (M = 70.88, SE = 1.68; F (1,23) = 23.81, p < .001). Finally, the ultimate outcome (final score) of the negotiation was then analyzed. As with the above metrics, the significant main effect of time (F (1,58) = 45.28, p < .001) was qualified by feedback condition (F (1,58) = 13.47, p = .001). As depicted in Figure 5 (right), while only marginally better outcomes were obtained for the control group in the second negotiation (M = 58.50, SE = 1.92) than the first (M = 54.33, SE = 2.19; F(1,29) = 3.92, p = .06) these results were not significant. Participants who received personalized feedback significantly improve in the second negotiation (M = 67.33, SE = 1.92) compared to the first (M = 53.17, SE = 2.19; F(1,29) = 67.05, p < .001). For the principle of value creating, participants did a better job of identifying the value of the items to the agent in the second negotiation (M = 6.27 object relationships found; SE = 0.49) than in the first (M = 5.24 object relationships found; SE = 0.42; F (1,56) = 3.42, p = .07), however, there was no interaction with condition (F (1,56) = 0.39, p = .54).

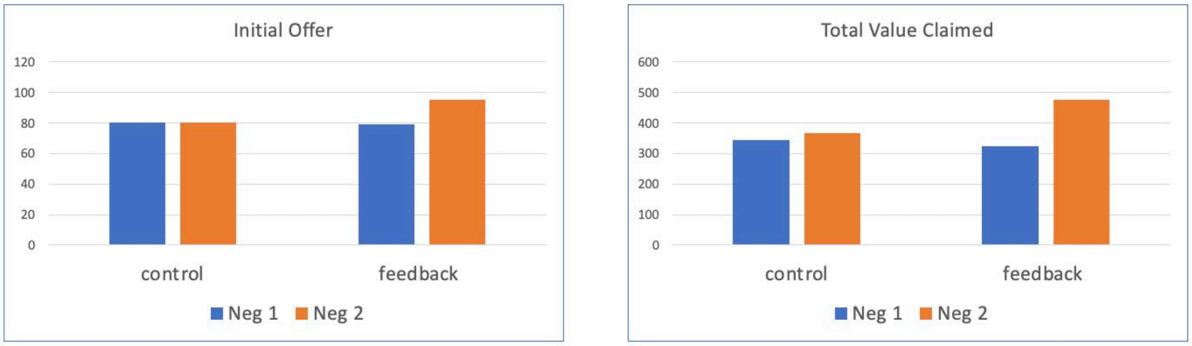


Figure 4. Effect of feedback condition on improvement in users' initial offers from negotiation 1 to 2 (left) and on improvement in users' total value claimed from negotiation 1 to 2 (right).

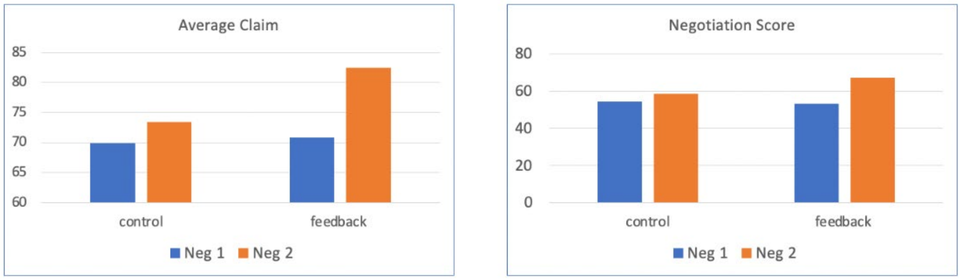


Figure 5. Effect of feedback condition on improvement in users' average claimed value from negotiation 1 to 2 (left) and improvement in final negotiation score from negotiation 1 to 2 (right).

Study Two

We evaluated the effects of practice and feedback with a 3 (feedback: none v. generic v. personalized) x 2 (time: negotiation 1 v. negotiation 2) mixed ANOVA. Mean values for each study and condition can be found in Table 1. For value claiming, students benefited from practice alone and this benefit was enhanced by feedback (both in tactics and final outcome). Students made stronger initial offers on the second negotiation (F (1, 98) = 33.47, p < .001) than the first, and the interaction with the type of feedback was not significant (F (2, 98) = 3.01, p =.054). Participants who received feedback (either personalized or generic) claimed more value. In terms of final outcome, we see a significant main effect of time (F (1,98) = 30.40, p < .001) and a significant interaction with the type of feedback (F (2, 98) = 3.808, p = .026). Participants obtained more points in the second negotiation and those who received personalized feedback gained the most points. For creating value, we found a clear benefit of practice and a strong effect of feedback for logrolling and joint points but not the questions asked. Concerning the final outcome, we find a significant benefit of practice on joint points as they created more value in the second negotiation than the first (F (1, 98) = 7.322, p = .008). Personalized feedback yielded the highest joint points, and the interaction was significant (F (2, 98) = 8.187, p = .001). Students engaged in logrolling more with practice (F (1, 98) =

37.495, *p* < .001) and there was a significant interaction with condition such that this improvement in logrolling from the first negotiation to the second was strengthened by personalized feedback (*F* (2, 98) = 4.930, *p* = .009). Students asked more questions over time (*F* (1, 98) = 24.461, *p* < .001) and asked the most with personalized feedback, though the interaction with condition was not significant (*F* (2, 98) = 1.711, *p* = .186).

Table 1. Study 2 Mean Value for Each Negotiation Outcome Metrics

Negotiation Principle	Outcomes			
	Metric	Condition	Negotiation 1	Negotiation 2
Value Claiming	Individual points	Personalized	25.72	31.21
		Generic	25.10	27.93
		No Feedback	25.72	27.30
	Initial Offer	Personalized	20.62	24.21
		Generic	17.69	24.62
		No Feedback	19.65	22.26
Value Creating	Joint Points	Personalized	57.48	63.45
		Generic	58.14	58.79
		No Feedback	59.70	58.91
	Questions Asked	Personalized	.59	1.69
		Generic	.52	1.14
		No Feedback	.74	1.21
	Logrolling	Personalized	1.72	5.14
		Generic	1.24	3.72
		No Feedback	2.37	3.16

Discussion and Conclusion

In this chapter, we examined the impact of personalized feedback on negotiation skills. Participants who received personalized feedback improved their outcomes in the second negotiation more than those who did not. Furthermore, personalized feedback did help students to improve their use of good negotiation principles. It increased learning by helping students to make more ambitious offers. Study 2 also emphasized that the behavior of the intelligent agent has a strong impact on student's behavior and these effects, depending on the agent, may not be well-aligned with the pedagogy. The results provided some support for both of our hypotheses. First, compared to those who did not receive any personalized feedback, providing students with *automated* personalized feedback about their use of negotiation principles helps them to improve their use of those principles. Furthermore, participants who received such personalized feedback also improved their outcomes in the negotiation more than those who did not. Specifically, in addition to participants in the personalized feedback condition showing greater improvement over time in their initial offer and value claimed (on average and in total), they also improved more at achieving good outcomes for themselves in the negotiation – compared to their counterparts who did not receive this feedback.

Some methodological choices seemed to undermine the benefits of personalized feedback. We used a fully integrative task (in which the agent and participant had complementary interests). This created a tension between value creating and value claiming (at least as we are measuring it in our studies) such that increased value claiming leads to less value creation in Study 2. Further, this task may be too simple, as students were quite good at creating value without feedback. A task that combines both fixed pie and logrolling issues would have allowed us to better tease apart value creation and value claiming. The choice of agent behavior also worked against our instruction. We found that the "fair" concession strategy of the agent punished students for making ambitious initial offers and helped students that were less ambitious. Future work should consider custom agent behavior that rewards students for following "best practices." Finally, anecdotal feedback suggests that we overloaded students by trying to teach both claiming and creation in the same exercise. Perhaps restructuring the exercise to teach one lesson at a time might be better.

GIFT Integration

The integration of IAGO and the Generalized Intelligent Framework for Tutoring (GIFT) offers a number of benefits to both researchers and ITS authors (see Figure 6). GIFT uses a service-oriented architecture which allows a number of modules to communicate with each other through asynchronous messages (Ragusa, Hoffman, & Leonard, 2013). This approach decouples the various components of the learning system making it easier to integrate third party applications. GIFT also provides tools that allow embedding domain knowledge, creating course flow and integrating different content into a course. This allows for easy integration of IAGO with GIFT and the creation of negotiation courses that are adaptable to an instructor's need. In addition to this, we are able to leverage other components of GIFT to enhance negotiation training. For example, the GIFT architecture permits the integration of a number of sensors which allow for affective sensing. As emotions are an important part of negotiation (Barry et al., 2004), these sensing modules can be leveraged to better understand a negotiator's affective state. In addition to the value GIFT offers for teaching negotiation through IAGO, the integration of the two is helpful for those who utilize GIFT as well.

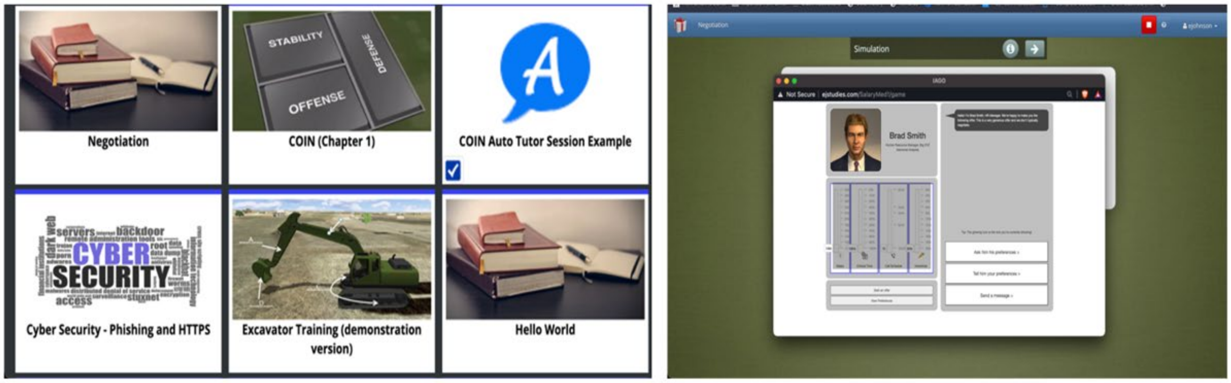

Negotiation
COIN (Chapter 1)
COIN Auto Tutor Session Example
Cyber Security - Phishing and HTTPS
Excavator Training (demonstration version)
Hello World
Negotiation
Simulation
IAGO
Brad Smith

Figure 6. On the left is an image of the negotiation module that was added to GIFT and on the right is an example of a simulated IAGO negotiation running in GIFT.

The integration of GIFT and IAGO also provides a test bed for others to study collaborate problem solving. Collaborative problem solving can be viewed as the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution (Herborn et al., 2020). Current GIFT users can leverage this integration with IAGO to run dyadic or multi party negotiation as a way to better understanding how individuals collaboratively problem solve. Towards this end, we completed an initial integration of IAGO with GIFT using the cloud-based authoring tool to create a negotiation course in GIFT. We utilized the gateway module to allow GIFT to launch a negotiation training course and simulation through IAGO. With this, we were able to create a course flow that includes video materials for students to learn how to negotiate, a survey to assess their negotiation abilities as well as their views on the negotiation progress and lastly a simulation with personalized feedback.

Recommendations and Future Research

This research established that automated personalized feedback can be used to improve participants' negotiation skills and outcomes; however, there are still several areas left for future work. First, although some work has shown that people view virtual negotiators similar to human negotiators (Gratch et al., 2015), future research is needed to determine whether getting to practice negotiating with (and receiving personalized feedback from) automated systems will transfer to negotiations with other humans. Interacting with a virtual human (instead of a real human negotiation partner) may be advantageous for novice negotiators who tend to feel anxiety or discomfort when first learning to negotiate (Gratch et al., 2016), but on the other hand because the negotiation partner is not a real human, the skills may not transfer as well as if they had practice with humans. To ensure that the benefits of automated negotiation partners and feedback can be fully realized, future work needs to systematically investigate how these virtual encounters affect, and hopefully promote, participants' ability to use negotiation skills in real-world interactions. Although this work is promising, our ultimate goal is to show that the benefits accrued through such automated practice, assessment and feedback will generalize outside these simulations. Future planned studies will examine if students improve in both computer-mediated and face-to-face negotiations with other students. There is additional work needed to generalize this approach to other soft skills as well as to expand our work to teaching more.

Our work illustrates the potential of using an ITS to assess and improve a human's negotiation abilities, and the critical role GIFT can play in this process. GIFT's modular architecture makes it possible to integrate standalone simulations and rapidly develop intelligent courses. However, GIFT does have some limitations. For one, in order to complete a course, individuals must have a GIFT account and this additional step makes it cumbersome for running experiments. A number of our studies are ran through Amazon Mechanical Turk where participants have a limited amount of time to complete a study. By requiring participants to create an account to have full access to the system, the additional steps may reduce participation. GIFT also offers an authoring tool for building courses, questionnaires and surveys. However, most researchers tend to use external surveying platforms like Qualtrics and Survey Monkey. It would be beneficial to have a way to easily integrate these external platforms. Lastly, GIFT was designed to support the rapid development of ITSs. As such, it lacks the support and suite of tools to allow users to rapidly load and configure various learning environments and character behaviors for social skills training simulations.

References

Barry, B., Fulmer, I. S., & Van Kleef, G. A. (2004). I laughed, I cried, I settled: The role of emotion in negotiation. In M.J. Gelfand & J.M. Brett (Eds.), The Handbook of Negotiation and Culture (pp. 71-94). Palo Alto, CA: Stanford University Press.

Chollet, M., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2014). An interactive virtual audience platform for public speaking training. In B. Ana, & M. Huhns (Ed.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (pp. 1657-1658). Paris, FR: IFAAMAS PRESS.

Durante, A., Larrimore, J., Park, C., & Tranfaglia, A. (2017). Report on the economic well-being of US households in 2016 (No. 2963). Washington, DC: Board of Governors of the Federal Reserve System (US).

Forum, W. E. (2016). The future of jobs: employment, skills and workforce strategy for the fourth industrial revolution. Global Challenge Insight Report. Geneva, Switzerland: World Economic Forum.

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: the role of perspective-taking and negotiator focus. Journal of personality and social psychology, 81(4), 657-669.

Goldman, E. G. (2012). Lipstick and labcoats: Undergraduate women's gender negotiation in STEM fields. NASPA Journal About Women in Higher Education, 5(2), 115-140.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. Psychological Science in the Public Interest, 19(2), 59-92.

Gratch, J., DeVault, D., & Lucas, G. (2016, 9). The benefits of virtual humans for teaching negotiation. In D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Schere, & A. Leuski (Eds.), Proceedings of the 16th International Conference on Intelligent Virtual Agents (Vol. 10011, pp. 283-294). Cham: Switzerland. Springer International Publishing. Retrieved from http://iva2016.ict.usc.edu/wp-content/uploads/Papers/100110276.pdf

Gratch, J., DeVault, D., Lucas, G. M., & Marsella, S. (2015). Negotiation as a challenge problem for virtual Humans. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), Proceedings of the 15th International Conference on Intelligent Virtual Agents (pp. 201-215). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-21996-7_21

Guo, P. J. (2015). Codeopticon: Real-time, one-to-many human tutoring for computer programming. In L. Celine (Ed.), Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (pp. 599-608). New York: NY: Association for Computing Machinery.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? Computers in Human Behavior, 104, paper 105624.

Johnson, E., DeVault, D., & Gratch, J. (2017). Towards An Autonomous Agent that Provides Automated Feedback on Students' Negotiation Skills. In K. Larson, & M. Winikoff (Ed.), In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (pp. 410-418). Liverpool, UK: International Foundation for Autonomous Agents and Multiagent Systems.

Johnson, E., Roediger, S., Lucas, G., & Gratch, J. (2019). Assessing common errors students make when negotiating. In P. Catherine, & M. Jean-Claude (Eds.), In Proceedings of the 19th ACM Intelligent Conference on Intelligent Virtual Agents (pp. 30-37). New York, NY: Association for Computing Machinery.

Kapur, M. (2008). Productive failure. Cognition and instruction, 26(3), 379-424.

Kelley, H.H. (1966). A classroom study of the dilemmas in interpersonal negotiations. In K. Archibald (Ed.), Strategic interaction and conflict (pp. 49–73). Berkeley, California: University of California, Institute of International Studies.

Kim, J. M., Hill, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., . . . Hart, J. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. International Journal of Artificial Intelligence in Education, 19(3), 289-308.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 30-43.

Koedinger, K. R., & Corbett, A. (2005). Cognitive tutors. In R. Sawyer (Ed.), The Cambridge Handbooks of the Learning Sciences (pp. 61-78). Cambridge, MA: Cambridge University Press.

Koedinger, K. R., & Corbett, A. (Eds.) (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. Cambridge, MA: Cambridge Univerity Press.

Kolb, A. Y., & Kolb, D. A. (2012). Experiential learning theory. In N. M. Seel (Ed.), Encyclopedia of the Sciences of Learning (pp. 1215-1219). Boston, MA: Springer.

Lin, R., Oshrat, Y., & Kraus, S. (2009). Investigating the benefits of automated negotiations in enhancing people's negotiation skills. In C. Sierra, & C. Castelfranchi (Eds.), Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (Vol. 1.1, pp. 345-352). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Mell, J., & Gratch, J. (2016). IAGO: Interactive Arbitration Guide Online. In C. M. Jonker, & S. Marsella (Eds.), Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (pp. 1510-1512). Liverpool, UK: International Foundation for Autonomous Agents and Multiagent Systems.

Mills-Tettey, G. A., Mostow, J., Dias, M. B., Sweet, T. M., Belousov, S. M., Dias, M. F., & Gong, H. (2009). Improving child literacy in Africa: Experiments with an automated reading tutor. In the Proceedings of the International Conference on Information and Communication Technologies and Development, (pp. 129-138). Piscataway,NJ: IEEE.

Morela Hernandez, D. R. (2016). Getting the Short End of the Stick: Racial Bias in Salary Negotiations. MIT, MA: MIT Sloan Management Review.

National Academies of Sciences, Engineering and Medicine (2016). Promising Practices for Strengthening the Regional STEM Workforce Development Ecosystem. Washington, DC: National Academies Press.

Olney, A., Bakhtiari, D., Greenberg, D., & Graesser, A. C. (2017). Assessing computer literacy of adults with low literacy skills. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 128-134). Wuhan, China: International Educational Data Mining Society.

Pataki, C. P. (2012). Virtual patients as novel teaching tools in psychiatry. Academic Psychiatry, 36(5), 398-400.

Ragusa, C., Hoffman, M., & Leonard, J. (2013). Unwrapping GIFT: A primer on developing with the generalized intelligent framework for tutoring. In E. Walker, & C.-K. Looi (Eds.), Workshop Proceedings of the 16th International Conference on Artificial Intelligence in Education - Recommendations for Authoring, Instructional Strategies and Analysis for Intelligent Tutoring Systems (ITS): Towards the Development of a Generalized Intelligent Framework for Tutoring (GIFT) (pp. 10-19), Memphis, TN.

Roediger, S. (2018). The effect of suspicion on emotional influence tactics in virtual human negotiation.[Master's Thesis, University of Twente]. Los Angeles, CA: University of Southern California/Institute for Creative Technologies.

Stevens, A. H. (2006). The use of virtual patients to teach medical students history taking and communication skills. The American Journal of Surgery, 191(6), 806-811.

VanLehn, K. S. (2003). Human tutoring: Why do only some events cause learning. Cognition and Instruction, 21(3), 209-249.

Wijekumar, K., Meyer, B., & Spielvogel, J. (2005). Web-based intelligent tutoring to improve reading comprehension in elementary and middle schools: Design, research, and preliminary findings. In G. Richards (Ed.), Proceedings of the E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 3206-3211). San Diego, CA: Association for the Advancement of Computing in Education.

Section III – Computational and Quantitative Models

Dr. Xiangen Hu, and Dr. Benjamin Goldberg, Eds.

CHAPTER 12 – INTRODUCTION TO COMPUTATIONAL AND QUANTITATIVE MODELS

Xiangen Hu[1] and Benjamin Goldberg[2]
University of Memphis[1]; U.S. Army Combat Capabilities Development Command - Soldier Center[2]

Core Ideas

One of the most important and necessary assumptions when implementing competency-based scenario design in any adaptive instructional system (AIS) is that the competencies (e.g., knowledge, skill, abilities) are measurable. The measurability of competencies and their constituent parts drive the requirements for learner modeling (Sottilare et al., 2013) and efficient assessment (Goldberg et al., 2021). The four chapters in this section focus on the measurability of competency in three levels: How to build mathematical models of competencies; How to implement measurable competencies in working Intelligent Tutoring Systems (ITSs); How to record/store measurable competencies in the form of experiential learning data standards.

Individual Chapters

The chapter by ***Robson, Hu, and Graesser*** presents a formal model for competencies. In this chapter, competencies such as knowledge, skills, abilities, attitudes, traits, etc. are defined in a mathematical framework (i.e., competency framework). Within the framework, other concepts, such as states, levels, and assertions are all defined mathematically. This mathematical framework for competencies can be used as an implementation guide for competence-based scenario design in AISs.

The next two chapters, by ***Maniktala, Barnes, Chi, Hampton, and Hu***, and ***Abdelshiheed, Maniktala, Barnes, and Chi*** give two example ITS applications in which measurable competencies are used. ***Maniktala et al.*** share exciting work on data-driven classifiers linked to the effectiveness and efficiency of learners' problem-solving behaviors, with a focus on determining optimal feedback to support the learning process as a whole. An overview on the development and application of their HelpNeed agent is provided, followed by the results of an experiment comparing performance gains between an adaptive feedback condition and an associated control when applied to a homework assignment in a collegiate level math course. In the other chapter by ***Abdelshiheed et al.,*** competencies such as metacognitive skills and motivation are quantified and examine their joint impact on learning (preparation for future learning (PFL)).

The last chapter by ***Florian Tolk*** provides an example showing how one would leverage an internationally accepted learner experience data standard, xAPI. This involves guidelines on system requirements for recording relevant learning activities with the purpose of building quantitative measures of competence.

References

Goldberg, B., & Owens, K. (2021). GIFT in a Blended Learning and Competency Development Continuum. *Proceedings of the Ninth Annual GIFT Users Symposium (GIFTsym9)*, 89.

Sottilare, R., Graesser, A., Hu, X., & Holden, H. (2013). *Design Recommendations for Intelligent Tutoring Systems: Volume 1 - Learner Modeling*. U.S. Army Research Laboratory.

Chapter 13 – Mathematical Models to Determine Competencies

Robby Robson[1], Xiangen Hu[2,3], Elliot Robson[1], and Arthur C. Graesser[2]
Eduworks Corporation.[1]; University of Memphis[2]; Central China Normal University[3]

Introduction

In this chapter, *entity* refers to a person, group of people or organization whereas *competencies* refer to the knowledge, skills, abilities, attitudes, traits, and capabilities that define what an entity knows and can accomplish. The purpose of this chapter is to survey models for determining whether (or to what extent) an entity E has a set of competencies in a *competency framework* $\boldsymbol{F}$ related to a job, task, or domain. We call this the *state* of E with respect to $\boldsymbol{F}$. In our formulation, the state at a given time t is defined by the values of functions (see Section 2) that can be estimated from data. Our goal is to present a variety of mathematical models, called *state models,* for computing this state. This state can play a role in predicting whether E will successfully perform a task and in determining how to best train or teach a competency. This chapter focuses only on mathematical models for computing competence.

The Term "Competency"

There is no general agreement on the meaning of the term *competency*, or even whether the term should be *competency* or *competence* (Markus et al., 2005; Teodorescu, 2006). Many communities of practice have narrower definitions of terms such as knowledge, skills, abilities, capability, and competency, whereas some use community-specific terms such as "standard" in US K-12 education (Porter et al., 2011). Tasks (which are not competencies) are often used as stand-ins for the ability to perform the tasks. Credentials (also not competencies) are often equated with the competencies they imply belong to the credential-holder. We use the single term *competency* for all of these related concepts.

Levels and Conditions

Many competencies can be scaled and categorized at different *levels*, such as novice, intermediate, advanced, and expert. We treat different levels as *separate competencies* and denote the levels of C by $\boldsymbol{L}(C)$. This allows different levels of C to have different sub-competencies, different related competencies, and different performance criteria. Similarly, competencies can be assessed or demonstrated, or applied under varied conditions and can have different degrees of difficulty. Just as with levels, the same competency under different conditions is treated as different competencies. Thus *navigating in daylight* is a different competency than *navigating in darkness.* Being an expert at navigating in daylight may require basic map reading skills, whereas being an expert at navigating in darkness may require expert map reading skills. This treatment is necessary for developing more sophisticated computational models and is consistent with instructional design and workforce development practices that include evaluation criteria and conditions in learning objectives and competencies.

Competency States

Let $\boldsymbol{F}$ be a set of competencies in a framework and E be an entity. In our formulation, the *competency state* (or simply *state*) $\boldsymbol{S}$ of E with respect to a competency $C \in \boldsymbol{F}$ at time t is described by the values of four functions:

- s: $\boldsymbol{F} \rightarrow \{T,F,U\}$, a discrete function that determines whether E possesses the competency C at time t, with T = True, F = False, and U = Unknown.
- p: $\boldsymbol{F} \rightarrow [0,1]$, a function that is often interpreted as the probability that E possesses C at time t.
- r: $\boldsymbol{F} \rightarrow [0,\infty)]^n$, a *repetition* function that represents the degree to which E has repeatedly practiced or trained on competency C. This function tracks quantities such as flight time and training sessions and is an important component for competencies that require experience to acquire. Although n quantities can be tracked, it is often practical to combine them into one ($n = 1$).
- e: $\boldsymbol{F} \rightarrow (-\infty,\infty)$, an *evidence* function that measures the strength of the evidence that E possesses C at time t. Negative values indicate evidence that E does not possess C. This function can be thought of as an assessment score for a competency.

State models often contain just one of s or p and do not explicitly include r or e. Nonetheless, all of these values are useful. For example, an adaptive instructional system (AIS) may use the Boolean state of pre-requisite competencies to determine which competencies to target and use p, r, and e to determine how the instruction should be provided.

Levels

As mentioned above, we treat levels of a given competency as separate competencies. If we are interested only in the Boolean state of a competency, we can introduce a function l: $C \rightarrow \boldsymbol{L}(C)$ for each $C \in \boldsymbol{F}$ that indicates the level at which C is possessed. If the levels $\boldsymbol{L}(C)$ are hierarchical, which is often the case, then $l(C)$ is the highest level $L \in \boldsymbol{L}(C)$ for which $s(C) = T$. The function l is not part of the state itself but can be computed from it.

Assertions

Our goal is to estimate the Boolean or probabilistic state of an entity with respect to competencies in $\boldsymbol{F}$. This should be done based on observations, measurements, and other data, which in our approach are all converted into *assertions.* An assertion a has the following required parameters:

- An **Agent** making the assertion. This is the person, organization, or system that is asserting competency or lack thereof.
- A **Source** of evidence or data used by the Agent. This could be the agent itself but is more often an assessment instrument, a training exercise, a credentialing organization, a talent management system, or another system that provides data used to draw conclusions about competency.
- An **Entity** and a **Competency** about which the assertion is made.
- A **time** at which the assertion is made.
- A **verb** which is "has" "does not have" or "attempted." The first two are used when the Agent uses the evidence or data provided by the Source as a basis for asserting that the Entity has or does not have a competency. "Attempted" is used to assert that the Entity attempted or practiced a competency without drawing any conclusion about whether the Entity possesses it.

An assertion may also have the following optional parameters

- The **Evidence** used by the agent to make the assertion. Evidence can be assessment results, data from a training system, notes from an instructor, or other evidence.
- A **confidence in [0,1]** that describes the confidence the agent has in the assertion being made.
- An **expiration time** at which the assertion expires.
- A **decay function** that describes how this confidence diminishes over time. This is a monotonically decreasing function with domain $[t_{0,}t_1]$ where t_0 is the timestamp of the assertion and t_1 is the expiration time of the assertion.

In formulas, we use a parameter □ known as the *polarity* of the assertion with □ = 1 if the verb is "holds," □ = 0 if the verb is "attempted," □ = -1 if the verb is "does not hold."

Note that the source of an assertion is not necessarily the same as the agent. For example, an exercise in a simulator might provide evidence of competency that an instructor interprets, in which case the simulator is the source, the results of the exercises are the evidence, and the instructor is the agent.

Creating Assertions

Assertions can originate from many types of data. Training systems can report activities and results in standardized formats (Bakhouyi et al., 2017). Credentials can attest to the possession of competencies (Klein-Collins, 2012). Instructors can report that a learner has demonstrated a competency. Individuals can self-assert competency (Forsman et al., 2020). Assertions can be extracted from performance records (Chen et al., 2014) and reviews. Workplace data running from sales results to operational data can be analyzed to create assertions. In most cases the raw data will not be in the form of an assertion but can be transformed into one. In the following sections we assume this has been done, noting that the confidence parameters may be missing.

State Models

A *state model* is a method for determining the state of an entity from a set of assertions about that entity. Such models can range from models that set the state of a competency based only on the latest assertion made to sophisticated machine-learned models and everything in between. In this section we review a few of the most common such models.

Sequential Models

The simplest and most common way to determine whether an entity *E* holds a competency *C* is to look at the sequence of assertions made about *E* with respect to *C*. Instructional systems do this when they assess competencies based on a test. In this model the only variables computed are (1) *e*, which might be a test score, and (2) *s*, which indicates whether the learner has or does not have the competency. In most cases, *s(C)* is determined by the latest test result, i.e. by the most recent assertion, although there is no reason that results cannot be averaged in some way.

Bayesian Knowledge Tracing (BKT) is another sequential model. In BKT the goal is to estimate *p(C)* based on a series of assessments of *C*. In our formulation, each assessment generates an assertion where the evidence is the result of the assessment and the verb could be "holds" or "does not hold." For each new assertion, we compute the new *p(C)* based on the prior *p(C)*, the probabilities of a Guess or Slip (assumed to be constant), and the probability of transitioning from *s(C)* = F to *s(C)* =T due to whatever learning occurred between assessments, which is also assumed to be constant (van De Sande, 2013).

Rollup Rules

The Sequential Models discussed above treat each competency independently, but there are often relations among competencies that can be used to infer information about states. The rules used to make these inferences are often called *rollup rules* because they determine the state of a competency by "rolling up" the state of its sub-competencies. To formulate rollup rules, it is necessary to identify relations among competencies in ***F***. Data standards for frameworks explicitly express a variety of relations, such as *A requires B*, *A enables B, A is a prerequisite for B, A broadens B,* and *A is similar to B* (Doignon & Falmagne, 2015). Authors of frameworks often use tables and number schemes to express hierarchies among competencies. When computing state models, we map these relations onto a parent/child relation and assume that ***F*** becomes a directed acyclic graph (DAG). Theoretically, this may require collapsing cycles formed through similarity or equivalence relations, but it is rare for such relations to exist within a single framework.

With the assumption that ***F*** is a DAG, a rollup rule is a rule that computes the new state of a competency *C* from the current state of *C* and its children. The assumption that ***F*** is a DAG allows rollup rules to be computed by starting with the leaf nodes of ***F*** and moving up the DAG, and in practice, most rollup rules involve only the immediate children of *C*. As an example of a rollup rule, if *A* and *B* are children of *C*, a rollup rule for the Boolean state of *C* might say that *s(C) = T* if and only if *s(A) = T* and *r(C) > 20* and either *p(A) >.7* or *s(B) = T*. Rollup rules can reflect "*C* requires *A"* by making it so that the s*(C)* cannot be T unless the *sS(A) = T,* and can reflect various enabling relations by setting thresholds for evidence scores on the children of *C*.

Models that use Rollup Rules

An early example of a rollup rule model is described in Robson and Poltrack (2017). This model, which was implemented in the Competency and Skills System (CaSS) developed by Eduworks Corporation (*CaSS Authoring Tool Final Report*, n.d., *Competency and Skills System (CaSS)*, n.d.)), implemented the following logic:

- If there are both positive and negative assertions about an entity *E* with respect to a competency *C*, then the state of *E* with respect to *C* is unknown or is determined using a conflict resolution rule. Such rules might include:
    - Use the latest assertion to determine *s(C).*
    - Set *s(C) = T* if any unexpired assertion is positive.
    - Set *s(C)=T* if there is at least one positive assertion and the number of negative assertions is less than the number of positive assertions.
    - Let $\{a_1,...,a_m\}$ be the set of assertions about *E* with respect to *C*, let c_i be the confidence for each a_i, and let $\square_i$ be the polarity of a_i. Define $e(C) = \sum_{i=1}^{m} c_i\lambda_i$. Then set *s(C)=T* if *e(C) > d*, where *d* is a threshold that *e* must exceed.
- If *D* is a child competency of *C*, then it assumes that *C* requires *D*, so if *s(D) = F* then *s(C) = F.*

Another example of a model that uses rollup rules is that used by the *Generalized Intelligent Framework for Tutoring* (GIFT) *Domain Knowledge File* (DKF) (Sottilare et al., 2017). In GIFT's DKF, tasks can be broken into *concepts*, and performance on concepts is determined by *condition classes* that measure actions (Gilbert et al., 2018). Concepts can have subconcepts, creating a hierarchical structure. Condition classes can be used by multiple concepts, causing the hierarchical structure to form a DAG. Performance on concepts is reported at, above, or below expectations.

GIFT itself does not maintain a competency state model *per se*. In current work funded by the US Army Research Development and Engineering Command, a framework ***F*** and a state model are stored in CaSS.

The structure of GIFT's DKF parallels that of ***F***, and GIFT reports training events and outcomes via xAPI statements that are translated into assertions. These assertions are used to compute the functions *r* and *e*, which apply a *forgetting function* (Averell & Heathcote, 2011) and a *spaced repetition function* (Kang, 2016) to decrease the value of older evidence and to count repetitions less if they occur closely together. The assertions from each source, which in this case means from each different type of training environment, are collated and weighted to produce *r* and *e*. Rollup rules are of three basic types: (1) Rules that reflect when a sub-competency is required, (2) rules that require sufficient repetition (as measured by *r*) and (3) rules that require a sufficiently high *e*-score. These rules are combined via Boolean combinations.

Applications of Models

The state models described in Section 4 draw conclusions about the state of *E*, including estimates of the probability that *E* has a competency *C* at a particular time. This suffices if the only question of interest is whether or not *E* has *C*. However, end users usually want to predict performance or recommend training, and possessing a competency does not imply that it will be performed correctly every time. For example, we might say that a baseball player is an excellent hitter, which is a statement about his/her competency, without meaning to imply that the player will get a hit every time. Naive ways to turn a state model into a performance prediction include equating the level at which a competency is held with a predetermined probability of success, or equating the probability of *having C* with the probability of *performing C*, but in practice more complex approaches, including using machine-learned predictive models, are warranted.

Similarly, knowing which competencies are held is important in determining the best training pathway, but further models must be developed to create recommendations. Many AISs do this today in one form or another, although typically they are driven by observations of learner performance without taking the extra step of computing a competency state model. As pointed out in the work on AIS interoperability being performed in the IEEE Learning Technology Standards Committee (*Working Groups and Study Groups — IEEE Learning Technology Standards Committee*, n.d.), using internal measures will not lead to the ability for AISs to share learner profile data and therefore will hinder the applicability of state models. Exchanging state models, however, is very feasible and will allow different AISs to share data about learners while continuing to operate at arm's length.

References

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35.

Bakhouyi, A., Dehbi, R., Talea, M., & Hajoui, O. (2017). Evolution of standardization and interoperability on E-learning systems: An overview. *2017 16th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 1–8.

CaSS Authoring Tool Final Report. (n.d.). https://apps.dtic.mil/sti/pdfs/AD1125206.pdf

Chen, Y., Wrenn, J., Xu, H., Spickard, A., 3rd, Habermann, R., Powers, J., & Denny, J. C. (2014). Automated assessment of medical students' clinical exposures according to AAMC geriatric competencies. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2014*, 375–384.

Competency and Skills System (CaSS). (n.d.). Retrieved September 29, 2021, from https://adlnet.gov/projects/cass/

Doignon, J.-P., & Falmagne, J.-C. (2015). Knowledge Spaces and Learning Spaces. In *arXiv [math.CO]*. arXiv. http://arxiv.org/abs/1511.06757

Forsman, H., Jansson, I., Leksell, J., Lepp, M., Sundin Andersson, C., Engström, M., & Nilsson, J. (2020). Clusters of competence: Relationship between self-reported professional competence and achievement on a national examination among graduating nursing students. *Journal of Advanced Nursing*, *76*(1), 199–208.

Gilbert, S. B., Slavina, A., Dorneich, M. C., Sinatra, A. M., Bonner, D., Johnston, J., Holub, J., MacAllister, A., & Winer, E. (2018). Creating a Team Tutor Using GIFT. *International Journal of Artificial Intelligence in Education*, *28*(2), 286–313.

Kang, S. H. K. (2016). Spaced Repetition Promotes Efficient and Effective Learning: Policy Implications for Instruction. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 12–19.

Klein-Collins, R. (2012). Competency-based degree programs in the U.s.: Postsecondary credentials for measurable student learning and performance. *Council for Adult and Experiential Learning*. http://files.eric.ed.gov/fulltext/ED547416.pdf

Markus, L., Thomas, H. C., & Allpress, K. (2005). Confounded by competencies? An evaluation of the evolution and use of competency models. *New Zealand Journal of Psychology*, *34*(2), 117.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core standards. *Educational Researcher* , *40*(3), 103–116.

Robson, R., & Poltrack, J. (2017). Using competencies to map performance across multiple activities. *Proceedings of the I/ITSEC*. https://adlnet.gov/assets/uploads/2017%20---%20(IITSEC)%20Robson,%20Poltrack%20-%20Competencies.pdf

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*. https://www.gifttutoring.org/attachments/download/2076/Updated%20Concept%20for%20the%20Generalized%20Intelligent%20Framework%20for%20Tutoring_9%20May%202017.pdf

Teodorescu, T. (2006). Competence versus competency: What is the difference? *Performance Improvement Advisor*, *45*(10), 27–30.

van De Sande, B. (2013). Properties of the Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining*, *5*(2), 1–10.

Working Groups and Study Groups — IEEE Learning Technology Standards Committee. (n.d.). Retrieved September 29, 2021, from https://www.ieeeltsc.org/

Chapter 14 – What's the Value of a Step? Using Data to Solve the Assistance Dilemma for Adaptive Problem Solving Help

Mehak Maniktala[1], Tiffany Barnes[1], Min Chi[1], Andrew J. Hampton[2] and Xiangen Hu[2,3]
North Carolina State University[1]; University of Memphis[2]; Central China Normal University[3]

Introduction

Research suggests that students often exhibit poor help-seeking behavior, where some abuse hints to expedite problem completion while some avoid seeking help when in need (Aleven et al., 2006; Aleven & Koedinger, 2000; Price et al., 2017). Several intelligent tutoring systems (ITSs) provide unsolicited assistance to prevent the negative consequences of help avoidance (Arroyo et al., 2001; Murray & VanLehn, 2006). However, it is difficult to determine when and whether to provide proactive assistance, i.e., unsolicited help in anticipation of future struggle. Providing more assistance than needed can lead to shallow learning, while providing less assistance than needed can lead to frustration and wasted time. This challenge in the domain of ITSs is called the assistance dilemma (Koedinger & Aleven, 2007). Addressing this dilemma is even more important yet challenging for complex, open-ended learning tasks that involve several learner activities with differing goals (Kalyuga Singh, 2016). For example, learning in well-structured domains like math and biology can be complex based on open-ended problems (i.e. those with a high number of interacting elements and many solutions), introducing challenges to working memory (i.e. recognizing which rule to apply), while also requiring self-regulated learning skills (i.e. setting problem-solving goals and monitoring progress). While researchers have recently explored ways to address the assistance dilemma, most work has been conducted on simpler tasks where problems have a single solution (Ueno & Miyazawa, 2017) or do not generalize because of domain-specific models (Klahr, 2009; McLaren et al., 2014).

In this chapter, we present data-driven methods to determine the value of a step in well-structured complex learning domains and showcase how we can use them to address the assistance dilemma in a logic tutor. Solving logic proofs has long been used as a technique to engage students in critical thinking, but this complex learning task induces considerable cognitive load, and there may be times when students need support to productively continue. Our approach builds upon the Hint Factory (Barnes et al., 2011), a method that leverages the Bellman equation for value iteration on structured prior student data to generate adaptive hints. We present a principled HelpNeed classification to determine when students are likely to need help -- which could be used as a competency measure that reflects the level of expertise a student demonstrates during problem solving. We also present a controlled study that examines the impact of providing proactive assistance upon prediction of HelpNeed. Our results suggest that students who receive this adaptive support exhibited higher training productivity, a lower chance of help avoidance, and significantly better posttest performance than those who did not receive the adaptive interventions. Finally, we conclude with recommendations for GIFT and future intelligent tutoring systems.

Method

In this section, we first define the HelpNeed classification, our method for determining unproductive steps and later present our HelpNeed predictor that detects the need for help at the start of each step (Maniktala, Cody, Isvik et al., 2020).

HelpNeed Classification

Prior literature suggests that learning is reflected in both correctness and time in problem-solving (Beck & Gong, 2013; Corbett & Anderson,1994; Kai et al., 2018). Our HelpNeed classification is based on a problem-solving step's duration and efficiency. Step duration can be *Long* if it is carried out in a time greater than 75th percentile of student step time for that problem and *Quick* otherwise. Efficiency is our proxy for eventual correctness and optimality in multi-step problem-solving, where the final solution is needed to gauge these values. We present more details on efficiency later in this section.

Maniktala et al. define the following step behaviors in the HelpNeed model, ranging from most expert to least expert (Maniktala, Cody, Isvik et al., 2020). We categorize efficient steps as not needing help irrespective of duration, with short times classified as *Expert-like* and longer times classified as *Strategic*. Next, we consider a single quick inefficient step as a plausible guess, and categorize it as an *Opportunistic* step. We do not classify Opportunistic steps as HelpNeed because research suggests that students should be allowed to learn through plausible guesses in semi open-ended domains (Capraro et al., 2012; Polya, 2004) as long as they learn from their guesses. However, a prolonged guess-and-check strategy needs intervention. We define *Far Off* step as a sequence of quick inefficient steps that demonstrate a lack of strategy, and needing help. Finally, we classify a step to be *Futile* when a student spends significant time without making progress, and needs help. This category represents "wheel-spinning" or unproductive struggle, where students exhibit a lack of mastery in a timely manner (Beck & Gong, 2013; Kai et al., 2018).

Next, we describe our method for determining step efficiency (Maniktala, Cody, Isvik et al., 2020). We use 72,560 unique states in the prior student data for 35 problems and 796 students. Step efficiency is our novel extension of the Hint Factory (Barnes et al., 2011) for assessing whether a step contributes toward an efficient solution. While some solution paths may not lead to solutions, some may lead to solutions that can be highly inefficient. Further, while a student can be in a state that leads to an efficient solution, there may be low probability for the student to select that path. Our method takes these aspects into account while determining the value of each step. First, an interaction network is generated using prior student data, where each node is a problem-solving state, each edge is a step (state transition), and the probability of each state-transition is recorded. A state transition occurs upon deriving a new logic statement or deleting one from the proof. Next, we carry out the Bellman backup for value iteration (used in reinforcement learning) on the states to determine their *quality* values. This involves assigning large rewards to solution states, large penalties to states that never lead to solutions, and small penalty for each step (to penalize longer solutions). Note that the Bellmann backup also considers the probability of transition between states, estimated using frequency, while assigning state quality values. We define two types of state quality metrics: *local*, which is the same as the values computed in the Hint Factory, and *global*, which we define using the following modified Bellman equation.

$$GQV(s) := GR(s) + \gamma \sum_{s'} P_a(s'|s) GQV(s') \tag{1}$$

Equation 1 sums *GQV(s')* over all states *s'* reachable from *s*, weighted by $P_a(s'|s)$, taking into account all future actions *a* from a current state, rather than just the one with the maximum expected value. The global rewards *GR* are identical to the ones defined in the Hint Factory, except that we assign higher rewards for more optimal (shorter) solutions. Proof of convergence for the modified value iteration Equation 1 and an example problem illustrating this concept is given in Maniktala, Cody, Isvik et al. (2020).

The next step toward defining step efficiency is defining student progress. Since each problem can have a different range of state quality values, we need a reference state, or basis for comparison, from which we measure progress. There can be two reference states - the previous state, in which case we say that we are measuring *relative progress*, or the start state, in which case we measure the *absolute progress*. A step is called efficient if the progress, using either the local or global quality metric, is a non-negative number. Since there are two ways to measure quality, and two ways to measure progress, we investigated all four combinations to define step efficiency. Each combination of quality and progress captures a different perspective on step efficiency. We use the *global quality* and *absolute progress* measure of step efficiency to define HelpNeed because our prior analysis shows that it has the highest correlation with posttest performance (Maniktala & Barnes, 2020).

HelpNeed Predictor

With the aim to predict and prevent unproductive HelpNeed steps, we built our HelpNeed predictor with two classes: 1 for predicting HelpNeed, and 0 otherwise. Further, we developed two types of classifiers: *state-based* and *state-free*. The state-based classifier is used when a student's problem-solving state can be matched with prior student data, and the state-free classifier is used otherwise. We engineered 63 step features and trained a variety of predictive models on student data from two semesters ($N = 437$) with an objective to maximize both recall (proportion of HelpNeed steps correctly predicted) and area under the ROC curve (AUC, the ability of a model to distinguish between HelpNeed and non-HelpNeed steps) assessed using cross validation. Random forest models had the best performance for both state-based ($Recall = .9$, $AUC = .83$) and state-free predictions ($Recall = .91$, $AUC = .62$). We erred toward high recall at the expense of AUC to ensure that we increase the chance of predicting HelpNeed when help is needed. More details about the training process including feature selection, model selection, and feature importance can be found in (Maniktala, Cody, Isvik et al., 2020). Note, the top three features for the state-based classifier are (1) Global-Absolute Progress (33.5%), (2) current state's Global Quality (22%), and (3) Local-Absolute Progress (13.1%). The top three features for the state-free classifier are (1) problem time (10.7%), (2) total clicks performed in a problem (8.5 %), and (3) incorrect logic rule applications in the problem (7.4%). The HelpNeed predictor was, therefore, defined as the combined state-based and state-free random forest classifiers that predict the next step's HelpNeed classification as in Table 1 at the start of that step using the global-absolute step efficiency metric.

Results

The experiment was conducted in the Fall 2019 semester, where the tutor was given as a homework assignment to 123 undergraduate students in a discrete math course at North Carolina State University. Our controlled study had two conditions: in the *Adaptive* condition, students received proactive hints upon predictions of HelpNeed, and in the *Control*, no such interventions were given. Students in both conditions could request help on-demand. Our proactive hints use the interface of *assertions* that are partially worked steps hinting what logic statement to derive next (Maniktala, Cody, Barnes et al., 2020).

Procedure

The tutor is divided into four sections: introduction, pretest, training, and posttest. The introduction presents two worked examples to familiarize students with the tutor interface. Next, students solve two problems in a *pretest*, which is used to determine students' incoming competence. Students are assigned a condition randomly using stratified sampling on pretest performance, resulting in 70 in *Adaptive* and 53 in *Control*. Note that we set a larger sample size for the Adaptive condition to gather more data on how the adaptive policy was carried out. Next, the tutor guides students through the *training* section with fifteen problems of varying difficulty. Finally, students take a more difficult *posttest* with five problems.

Note that students can only receive hints in training, but the tutor is designed to provide immediate feedback whenever students make errors applying logic rules in all the sections. Among these participants, 111 (66 in Adaptive and 45 in Control) completed the tutor. We used a chi-squared test to assess the impact of tutor completion rates on the group sizes, and found that the impact was not significant ($\chi^2_{(1, N = 123)} = 0.16$, $p = .69$).

Posttest Performance

We hypothesized that students in the Adaptive condition would have better posttest performance than those in the Control condition, as measured by solution optimality and time. *Optimality* is an exponential decay function on normalized steps e^{-steps} to account for the small variance in the number of steps and normalization is done via robust scaling. We found that on posttest solution optimality, the Adaptive group (*Mean* = .71, *SD* = .27) performed significantly better ($t(110) = 1.74$, $p = .04$) than the Control (*Mean* = .59, *SD* = .33), with a moderate effect size (*Cohen's d* = 0.4). Next, on the total posttest time, we found a significant difference between the two groups using Welch's t-test and a large effect size, $t(110) = 3.99$, $p < .01$, *Cohen's d* = 0.8, with students in the Adaptive condition (*Mean* = 18 min, *SD* = 12 min) spending significantly less time on the posttest than those in the Control (*Mean* = 29 min, *SD* = 17 min). These results confirm our hypothesis about posttest performance. Note, we did not hypothesize or assess differences in the rule application errors because the tutor is designed to provide immediate feedback on incorrect rule applications without penalties, even in the pre- and post-tests.

Training Behavior

We also hypothesize that students in the Adaptive condition will exhibit better training behaviors, with (a) fewer HelpNeed steps, and (b) lower possible help avoidance, and higher possible help appropriateness (a higher chance of receiving help when it was likely to be needed), as measured using the HelpNeed classifier, when compared to the Control.

We first compare the training HelpNeed between the two conditions. Overall, the Adaptive condition took marginally significantly fewer steps than the Control: *Means: Adaptive: 121, Control: 133;* $t(110) = 1.29$, $p = .10$. However, we found a significant difference between the two conditions in their quick inefficient steps, where the Adaptive condition significantly outperformed the Control for both *Opportunistic* (Means: Adaptive: 5, Control: 7,;$p < .01$), and *Far Off* steps (*Means: Adaptive: 16, Control: 25; p = .02*), but there were no significant differences in *Futile* steps between the two conditions (*Means: Adaptive: 13, Control: 12, p = .47*). Since we only observed differences in the *Far Off* steps but not the *Futile* steps, these results only partially confirm our hypothesis on reduced training HelpNeed steps for the Adaptive condition. Our results suggest that compared with the Control condition, the Adaptive condition avoided unnecessary Opportunistic and Far Off steps that might distract them away from efficient solutions. The significantly higher Opportunistic and Far Off steps in the Control condition may be a result of help avoidance because students may not know when to seek help (Azevedo & Cromley, 2004; Peña et al., 2011).

Next, we present a comparison of help avoidance, abuse, and appropriateness between the two conditions. We use the help behaviors defined by Maniktala et al. using the HelpNeed classification and predictor (Maniktala, Cody, Isvik et al., 2020). *Possible help avoidance* is the percentage of total training steps that were observed to be HelpNeed but hints (on-demand or proactive) were neither requested nor proactively provided. The Adaptive condition (*Mean* = 12.5%, *SD* = 3.5%) has significantly lower possible help avoidance than the Control (*Mean* = 26.6%, *SD* = 5.2%) using a Mann-Whitney U test: ($U = 138$, $p < .01$). Next, *possible help appropriateness* is defined as the percentage of training steps predicted to need help and a hint was either requested or provided proactively. The Adaptive condition (*Mean* = 22.8%, *SD* = 4.8%) has significantly higher possible help appropriateness during training than the Control (*Mean* =

4.7%, *SD* = 2.2%): *U* = 155, *p* < .01. These results confirm our second hypothesis that students in the Adaptive condition had lower possible help avoidance and higher possible appropriate help than students in the Control.

Finally, *possible help abuse* is the percentage of training steps with neither predicted nor observed HelpNeed but students requested hints, indicating either that the prediction was wrong and help was needed and effectively used, or it was right and help was abused. While we did not hypothesize any differences between the groups for this metric, we found a significant difference between the two conditions (*U* = 365, *p* < .01), with the Adaptive condition (*Mean* = 1.3%, *SD* = 0.4%) having lower possible help abuse than the Control (*Mean* = 5.8%, *SD* = 2.4%).

Discussion

Prior literature defining unproductive behavior either deals with metrics of problem-completeness (Beck & Gong, 2013) or domain-specific definitions that require expert knowledge (McLaren et al., 2014). However, the HelpNeed classification is established on step-level metrics that are domain-agnostic: efficiency and time. The success of the HelpNeed classification is attributed to its roots in educational literature and our novel data-driven efficiency metrics, especially the global quality value. We defined it as an extension of the Hint Factory—by modifying the Bellman equation to be more representative of student actions and varying solution rewards in proportion to their optimality.

The next piece of the puzzle in solving the assistance dilemma was delivering assertions (partially worked steps) as hints upon predictions of HelpNeed. We used the interface of assertions because they have been shown to foster productive persistence among students with low prior knowledge (Maniktala, Cody, Barnes et al., 2020). Our post-hoc evaluation of the experiment shows that the Adaptive condition has fewer Opportunistic and Far Off steps in training, and better posttest performance (time and proof optimality) than the Control. While one can argue that the increased number of total hints could have improved the Adaptive condition's posttest performance, our research suggests that simply receiving more proactive hints at random times can be harmful, so it is important to identify when help is needed (Maniktala, Cody, Isvik et al., 2020).

Researchers have investigated several ways to address poor help-seeking behaviors such as help avoidance and help abuse (Aleven & Koedinger, 2000; Aleven et al., 2006; Price et al., 2017). Our evaluation of students' help behavior using our HelpNeed classification and predictor shows that the Adaptive condition had significantly lower help avoidance and abuse with significantly higher help appropriateness. Thus, our HelpNeed model effectively addresses the assistance dilemma.

Recommendations and Future Research

A tutor can incorporate the HelpNeed model if the following three aspects can be defined: (1) states, (2) state-transitions or steps, and (3) a scoring method for solutions. This lends itself readily to multi-step well-structured complex learning tasks in domains such as programming, math, physics, and statistics. Researchers have already experimented with adopting the Hint Factory to programming tutors (Peddycord III et al., 2014; Price et al., 2016; Rivers & Koedinger, 2013). Adopting the HelpNeed approach in these tutors can be relatively straightforward. Other domains are likely amenable, but have not been tested. One crucial step toward generalization is defining Opportunistic (inefficient but not HelpNeed) and Far Off (inefficient and HelpNeed) steps. For other problem-solving domains, these categories should be defined based on how long a student should be allowed to work without intervention.

The HelpNeed approach can enhance the pedagogical module of the Generalized Intelligent Framework for Tutoring (GIFT) for instructional management. HelpNeed could be incorporated into the Domain Module of GIFT, both to improve performance of the module and to validate HelpNeed's generalizability. The Domain Module interprets problem-solving states from various forms of input to inform the Learner Module (Sottilare & Brawner, 2018). This in turn adjusts the approach of the Pedagogical Module (i.e., adjusts difficulty, task parameters, feedback, etc.). Incorporating HelpNeed into the Domain module, would allow for the development of specific condition classes to provide context on the timing and quality of steps enacted while completing a problem. These conditions could provide an at-, below-, or above-expectation label linked to how fast each step was completed, and the efficiency of each step leading to the outcomes and HelpNeed classifications, thus allowing for optimization within a given domain and application.

Furthermore, future research could investigate the use of sequential HelpNeed classifications to better determine a performance plateau that might account to a competency level not inherently captured in problem by problem outcome data. Such sequences of expert-like, strategic, and opportunistic behaviors in complex learning tasks may reflect the development of expert skills and strategies. Typically, detecting expert skill development would require a complex domain model, but the HelpNeed approach relies in data typically available in transactional tutor logs, potentially enabling further optimization of GIFT and other tutors for competence modeling.

Conclusions

This chapter showcases how we can leverage the reinforcement learning technique of value iteration to define the value of a step in well-structured complex learning tasks. It also presents a HelpNeed classification that integrates step efficiency metrics with educational literature to determine steps where students need help, and use it to define an adaptive hint policy. The chapter presents empirical evidence to support that this adaptive hint policy effectively addresses the assistance dilemma in the logic tutor. Students who received adaptive assistance had fewer inefficient steps in training, and performed significantly better in the posttest (on time and optimality) than their control peers.

The HelpNeed model may be generalized to other intelligent tutoring systems such as GIFT, but its effectiveness should be studied across domains and learning tasks. Incorporation into the GIFT Domain Module can facilitate this testing and add a valuable resource to GIFT's toolkit.

References

Aleven, V. & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In Intelligent tutoring systems (pp. 292–303). Springer Berlin Heidelberg.

Aleven, V., McLaren, B., Roll, I. & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education,16*(2), 101–128.

Arroyo, I., Beck, J. E., Beal, C. R., Wing, R. & Woolf, B. P. (2001). Analyzing students' response to help provision in an elementary mathematics intelligent tutoring system. Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments, 34–46.

Azevedo, R. & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523–535.

Barnes, T., Stamper, J. & Croy, M. (2011). Using Markov decision processes for automatic hint generation. Handbook of Educational Data Mining, 467-480.

Beck, J. E., & Gong, Y. (2013, July). Wheel-spinning: Students who fail to master a skill. In International conference on artificial intelligence in education (pp. 431-440). Springer, Berlin, Heidelberg.

Capraro, M. M., An, S. A., Ma, T., Rangel-Chavez, A. F. & Harbaugh, A. (2012). An investigation of preservice teachers' use of guess and check in solving a semi open-ended mathematics problem. *The Journal of Mathematical Behavior, 31*(1), 105–116.

Corbett, A. T. & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C. & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining, 10*(1), 36–71.

Kalyuga, S., Singh, A.M. Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review, 28*, 831–852 (2016). https://doi.org/10.1007/s10648-015-9352-0

Klahr, D. (2009). ''To every thing there is a season, and a time to every purpose under the heavens": What about direct instruction? Constructivist Instruction: Success or failure?, 291–310.

Koedinger, K. R. & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264.

Maniktala, M., & Barnes, T. (2020). Extending the hint factory: Towards modelling productivity for open-ended problem-solving. Proceedings of the 13th International Conference on Educational Data Mining. International Educational Data Mining Society (IEDMS).

Maniktala, M., Cody, C., Barnes, T. & Chi, M. (2020). Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *International Journal of Artificial Intelligence in Education, 30*(4), 637–667.

Maniktala, M., Cody, C., Isvik, A., Lytle, N., Chi, M., & Barnes, T. (2020). Extending the Hint Factory for the assistance dilemma: A novel, data-driven HelpNeed Predictor for proactive problem-solving help. *Journal of Educational Data Mining, 12*(4), 24-65.

McLaren, B. M., Timms, M., Weihnacht, D., Brenner, D., Luttgen, K., Grillo-Hill, A. & Brown, D. H. (2014). A web-based system to support inquiry learning. Proceedings of the 6th International Conference on Computer Supported Education-Volume 1, 43–52.

Murray, R. C. & VanLehn, K. (2006). A comparison of decision-theoretic, fixed-policy and random tutorial action selection. Proceedings of the 8th international conference on Intelligent Tutoring Systems (pp. 114–123).Springer.

Peddycord III, B., Hicks, A. & Barnes, T. (2014). Generating hints for programming problems using intermediate output. Proceedings of the 7th International Conference on Educational Data Mining (pp. 92–98). International Educational Data Mining Society (IEDMS).

Peña, A., Kayashima, M., Mizoguchi, R. & Dominguez, R. (2011). Improving students' meta-cognitive skills within intelligent educational systems: A review. International Conference on Foundations of Augmented Cognition, 442–451.

Polya, G. (2004). How to solve it: A new aspect of mathematical method (Expanded Princeton Library Edition). Princeton University Press.

Price, T. W., Dong, Y. & Barnes, T. (2016). Generating data-driven hints for open-ended programming, Proceedings of the 9th International Conference on Educational Data Mining (pp. 191–198). International Educational Data Mining Society (IEDMS).

Price, T. W., Zhi, R., & Barnes, T. (2017). Hint generation under uncertainty: The effect of hint quality on help-seeking behavior. Proceedings of the 18th International Conference on Artificial Intelligence in Education (pp. 311-322). Springer, Cham.

Rivers, K., & Koedinger, K. R. (2013, June). Automatic generation of programming feedback: A data-driven approach. Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Part 9: The First Workshop on AI-supported Education for Computer Science (AIEDCS) (p. 50).

Sottilare, R., & Brawner, K. (2018). Component interaction within the Generalized Intelligent Framework for Tutoring (GIFT) as a model for adaptive instructional system standards. The Adaptive Instructional System (AIS) Standards Workshop of the 14th International Conference of the Intelligent Tutoring Systems (ITS) Conference, Montreal, Quebec, Canada.

Ueno, M. & Miyazawa, Y. (2017). IRT-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies, 11*(4), 415–428.

CHAPTER 15 – ASSESSING COMPETENCY USING METACOGNITION AND MOTIVATION: THE ROLE OF TIME-AWARENESS PREPARATION FOR FUTURE LEARNING

Mark Abdelshiheed, Mehak Maniktala, Tiffany Barnes, and Min Chi
North Carolina State University

Introduction

One fundamental goal of learning is preparation for future learning (PFL) (Bransford & Schwartz, 1999) and being able to extend acquired skills and problem-solving strategies to different domains and environments. While substantial research has shown that PFL can be accelerated by obtaining metacognitive skills (Chi & VanLehn, 2010; Zepeda et al., 2015) or influenced by the individual's motivation (Belenky & Nokes-Malach, 2013; Nokes & Belenky, 2011), no prior work investigated whether the interaction of the two factors could assess students' competency for PFL. In this chapter, we tackle this research question in one type of highly interactive e-learning environment, Intelligent Tutoring Systems (ITSs). More specifically, we investigate whether the combination of metacognitive skills and motivation would assess students' learning abilities in logic, and their competence to extend these abilities to a subsequent domain, probability.

We focus on two types of metacognitive skills related to problem-solving strategies: strategy-awareness and time-awareness, that is respectively, how and when to use each strategy. Moreover, we track the accuracy collected from the online traces from both ITSs to measure students' motivation levels. By doing so, we hypothesize that high-motivated students who additionally know how and when to use each strategy will yield the highest learning outcomes on a logic tutor and transfer their acquired skills to a subsequent probability tutor. Both tutors have been extensively evaluated in the past decade and a series of papers have been published to show their effectiveness independently (Barnes et al., 2008; Chi & Vanlehn, 2007; Chi & VanLehn, 2010).

In deductive task domains such as logic and probability, solving a problem often requires producing a proof, argument, or derivation consisting of one or more inference steps, and each step is the result of applying a domain principle, rule, or operator. Prior work has shown that students often use a mixture of problem-solving strategies such as forward chaining (FC) and backward chaining (BC) during their problem solving (Newell & Simon, 1972; Priest & Lindsay, 1992; Simon & Simon, 1978). Many prior studies investigated the impact of teaching students an explicit problem-solving strategy on their learning gains (Chi & VanLehn, 2007; Zepeda et al., 2015) or compared students who were taught different types of strategies (Boden et al., 2018; Chi & VanLehn, 2010). In this chapter, we argue that time-awareness should be considered as an independent type of metacognitive skill apart from problem-solving strategies, and we investigate:

1. How students' knowledge about how to use a problem-solving strategy (strategy-awareness) and when to use it (time-awareness) would impact their learning.
2. How such impact would change when we consider the student motivation.
3. How would the interactions between the two types of metacognitive skills and motivation impact students' learning in a new domain, probability.

Background

Bransford and Schwartz (1999) proposed the theory of PFL that states that students need to continue to learn, and investigates whether they are prepared to do so. Similar to prior work (Chi & VanLehn, 2010), we bring PFL into the ITS context, where it is possible to directly observe behaviors associated with PFL. In this chapter, we evaluate students' choices of how and when to select a problem-solving strategy. Based on Winne and Azevedo (2014), mastering how to use each strategy is a cognitive skill, but when incorporated with awareness about when such strategy should be used, it becomes a metacognitive skill. Therefore, we consider strategy-awareness and time-awareness to be two different types of metacognitive skills. Specifically, we investigate how the interactions of the two types of metacognitive skills and motivation could impact PFL.

The Impact of Metacognitive Skills on PFL

Metacognition indicates one's realization of their self-cognition as well as being able to regulate and understand it (Chambres et al., 2002; Roberts & Erdos, 1993). It is the act of exercising and monitoring control of cognitive skills (Efklides, 2011). Hence, we consider that a metacognitive skill consists of a cognitive skill and a regulator for controlling this skill. Many studies have shown that metacognitive skills have positive impacts on learning (Zepeda et al., 2019), on students' learning behaviors (Belenky & Nokes, 2009), and on PFL across ITSs (Zepeda et al., 2015; Chi & VanLehn, 2010). Several approaches have been used to evaluate metacognitive skills, such as strategy selection (Chi & VanLehn, 2010; Roberts & Erdos, 1993), tutoring prompts (Zepeda et al., 2015; Belenky & Nokes, 2009), and reading comprehension and memory recall (Chambres et al., 2002).

Zepeda et al. (2015) demonstrated that metacognitive instruction could influence student metacognitive skills, motivation, and transfer learning. Students who were taught planning, monitoring, and evaluating, made better metacognitive judgments and showed higher motivation levels (e.g. self-efficacy and task value) than those who were not taught these skills. As an example of PFL, the former also performed better on a novel self-guided learning task than the latter.

In our prior work, Chi and VanLehn (2010) investigated the transfer of metacognitive skills from a probability tutor to a physics tutor. We showed that an ITS teaching domain-independent metacognitive skills could close the gap between high and low learners, not only in the domain where they were taught (probability), but also in a second domain where they were not taught (physics). In that study, the metacognitive skills included a problem-solving strategy and principle-emphasis instructions. We found that it was the principle-emphasis skill that is transferred across the two domains and that closed the gap between the high and low learners. In this chapter, we investigate how students' own metacognitive skills (strategy- and time-awareness) would impact their learning and also prepare them for future learning in a new domain with a new ITS.

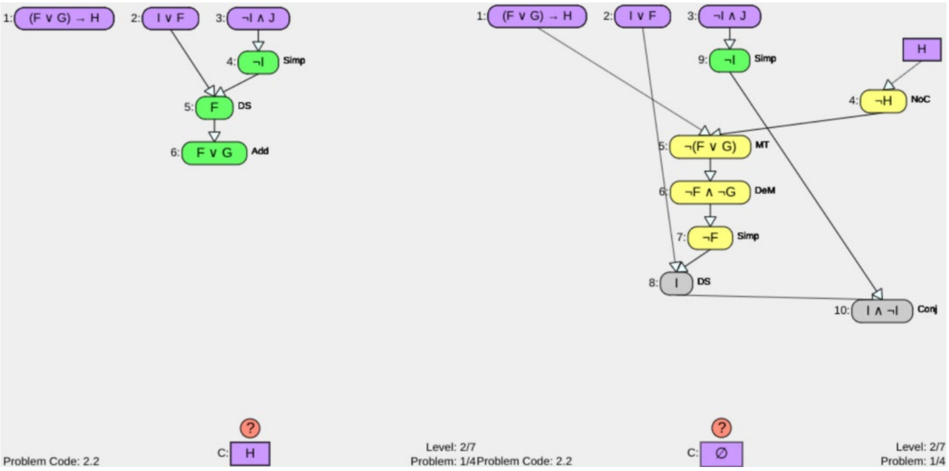
The Impact of Motivation on PFL

Substantial work has shown the impact of motivation on PFL (Belenky & Nokes-Malach, 2012, 2013; Nokes & Belenky, 2011). For instance, Belenky and Nokes-Malach (2012) found that PFL is influenced by the interaction of students' motivation and different forms of instruction. They found that students who had high mastery-approach goal orientation showed signs of transfer, irrespective of the instruction type. Furthermore, students who were allowed to innovate new strategies developed higher motivation aspects, compared to those who followed direct instruction. Later, the same innovative students showed strong evidence of PFL when given a final word problem. Nokes and Belenky (2011) incorporated students' achievement goals into a PFL framework that accounts for transfer success or failure. The framework represents a loop of goal generation, environment interpretation, knowledge & goal representation, solution generation, and solution evaluation. The last step decides whether the loop will be repeated or not. After testing this framework, they reported that mastery-approach goal-oriented students were more likely to succeed in knowledge transfer.

One of the crucial questions for research on motivation is how to define and measure motivation. Eccles (1983) defined motivation to be the individual's perception of three factors: expectations for success, subjective task value, and intrinsic interest. The three factors respectively represent how much success one would expect from a task, what value it has and why there would be an interest to accomplish it. Touré-Tillery and Fishbach (2014) classified motivation into two dimensions: process-focused 'doing it right' and outcome-focused 'getting it done'. To measure motivation, prior research explored self-efficacy (Boden et al., 2018; Kalender et al., 2019; Zepeda et al., 2015), goal orientation (Otieno et al., 2013; Belenky & Nokes-Malach, 2012, 2013), and accuracy (Touré-Tillery & Fishbach, 2014). The majority of these studies used surveys to measure these aspects. For instance, Kalender et al. (2019) used a survey to measure three motivational aspects based on the achievement goals: self-efficacy, interest, and sense of belonging. In recent years, digital technologies such as ITSs made it possible to measure motivation using students' online traces in ITSs (Otieno et al., 2013). Otieno et al. (2013) used the online use of hints and glossaries in a geometry tutor as a motivation measure and found that the online measures differ from the motivation measures using questionnaire data, and the former was more predictive of posttest scores than the latter. Therefore, in this chapter, we use students' online traces to measure their motivation levels. More specifically, we extract the initial accuracy from each tutor to measure students' motivation.

Experiment

Our data were collected from an undergraduate computer science course at North Carolina State University across three semesters. Students were trained on the logic tutor first, and then on the probability tutor six weeks later. The tutors were assigned as one of their regular homework assignments and the completion of both tutors was required for full credit. Students were told that the assignment would be graded based on the demonstrated effort, not performance. A total of 495 students finished both tutors with the following distribution: $N = 151$ for Fall 2017, $N = 128$ for Spring 2018, and $N = 216$ for Fall 2018.



(a) Direct Proof (b) Indirect Proof

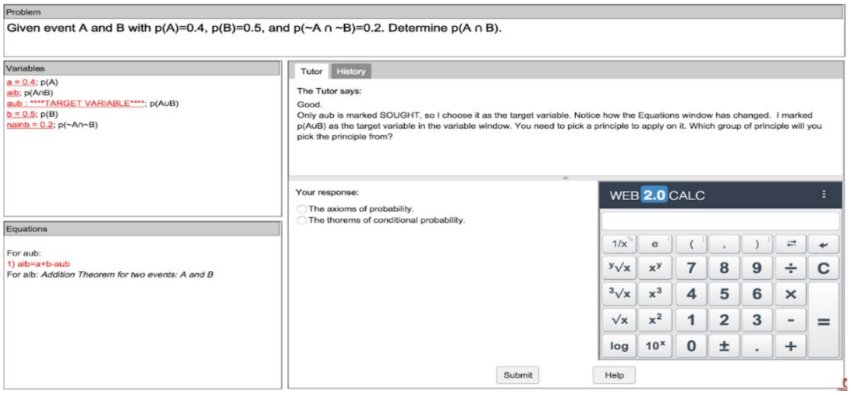
Figure 1. Logic Tutor Problem-Solving Strategies

Methods

The Logic and Probability tutors

Students went through a standard pretest-training-posttest procedure on each tutor. The logic tutor teaches students propositional logic proofs. A student can solve a problem in one of two strategies: direct or indirect. Figure 1a shows that for direct proofs, a student needs to derive the conclusion node at the bottom from the givens at the top; while Figure 1b shows that for indirect proofs, a student derives a contradiction from the givens and the negation of the conclusion. Both logic pre- and post-test have two problems and their scores are a function of time and accuracy. The training on the logic tutor includes 20 problems.

Figure 2 shows the graphical user interface (GUI) for the probability tutor that teaches students how to apply principles to solve probability problems. The pre- and post-test sections have 14 and 20 problems, respectively. These problems require students to derive an answer by writing and solving one or more equations. The training includes 12 problems.



Problem

Given event A and B with p(A)=0.4, p(B)=0.5, and p(~A ∩ ~B)=0.2. Determine p(A ∩ B).

Variables

a = 0.4; p(A)
aib; p(A∩B)
aub : ****TARGET VARIABLE****; p(A∪B)
b = 0.5; p(B)
nainb = 0.2; p(~A∩~B)

Equations

For aub:
1) aib=a+b-aub
For aib: *Addition Theorem for two events: A and B*

Tutor | History

The Tutor says:

Good.

Only aub is marked SOUGHT, so I choose it as the target variable. Notice how the Equations window has changed. I marked p(A∪B) as the target variable in the variable window. You need to pick a principle to apply on it. Which group of principle will you pick the principle from?

Your response:

○ The axioms of probability.
○ The thorems of conditional probability.

Figure 2. Probability Tutor Interface

There are two major differences between the two ITSs:

1. In the probability tutor, the pre- and post-test scores are based on accuracy. Both tests were graded in a double-blind manner by experienced graders using a partial-credit rubric. In the logic tutor, they are based on both accuracy and efficiency. Since there are only two questions in each test, the class instructor believes that it is as important for students to solve them accurately as for them to solve them quickly. For comparison purposes, all test scores are normalized to the range of [0, 100]. Note that in both tutors, the posttest is much harder than the pretest.
2. In the logic tutor, students can select FC-like direct proofs (the default) or choose to switch to BC-like indirect proofs. Conversely, in the probability tutor, students can only use BC during training. In both tutors, students can employ any strategy during the pre- and post-test.

Metacognitive Skills

Students can choose to switch problem-solving strategies only when training on the logic tutor. Hence, we measure students' metacognitive skills based on their interactions with the logic tutor alone. The training section in the logic tutor is organized into five ordered levels with an incremental degree of difficulty and each level consists of four problems. Each problem can be solved by either following the default FC (direct) or switching to BC (indirect). However, most advanced problems (in higher levels) can be solved much more efficiently by BC. Therefore, we expect that effective problem solvers should switch their strategy on these problems, and more importantly, they should switch it early when solving them. Thus, our metacognitive skill measurement is a combination of strategy-awareness: using the default direct proof or switching to indirect proof (Chambres et al., 2002; Chi & VanLehn, 2010; Roberts & Erdos, 1993), and time-awareness: when such switch happens (Winne & Azevedo, 2014). We consider two factors in time-awareness: one is that a student should switch in later levels (harder training problems) where the savings will be big and the other is that students should switch early (when convenient) during solving a problem. On average, students take 210 actions to solve a problem, and the median number of actions that a student takes before switching is 30. Therefore, we calculated the metacognitive score (MetaScore) for a student **i** as:

$$MetaScore_i = \sum_{L=1}^{5} \left[\sum_{p=1}^{4} \left[L * SAware_{ip} * TAware_{ip} \right] \right]$$

where strategy-awareness $SAware_{ip} = 1$ indicates that student i switched strategy when solving problem p at level L, while 0 means no switch. For time-awareness, $TAware_{ip} = 1$ indicates that the student i switched early on problem p (≤ 30 actions) while $TAware_{ip} =$-1 is for a late switch (> 30 actions). As stated before, the training levels have an incremental degree of difficulty, as each level introduces a new logic rule. Since the rate of change of rules per level is constant, the difficulty of the tutor was assumed to be linear, and therefore, we weighted the strategy- and time-awareness by the corresponding level number. Based on this formula, $MetaScore_i > 0$ suggests that student i is both strategy-aware and time-aware; if $MetaScore_i < 0$, it shows that student i is strategy-aware but not good at knowing when to switch (time-unaware). Finally, if $MetaScore_i = 0$, we do not have enough evidence on the student's metacognitive skills in that he may simply follow the default FC settings. Based on MetaScores, students are classified into three groups: those who show both strategy- and time-awareness (MetaScore > 0) are referred to as the '**Str_Time**' group ($N = 145$); those who showed strategy awareness only (MetaScore < 0) as '**Str_Only**' ($N = 166$); and the default students (MetaScore = 0) as '**Default**' ($N = 184$).

Motivation

Inspired by prior research (Touré-Tillery & Fishbach, 2014), we measured students' motivation by tracking the accuracy of their online traces. By doing so, we acknowledge that students often have different motivations: some are more process-focused for learning the domain as much as possible and some are more outcome-focused for better grades. Similar to prior work (Rheinberg et al., 2000; Vollmeyer & Rheinberg, 2006), we define students' motivation based on their initial interactions in the early stages of each tutor. More specifically, we use the percentage of correct rule applications in the first two problem-solving questions as our motivation indicators. In other words, our measured students' initial motivation levels do not consider that students' motivation levels may change over time. Students are divided into high- and low-motivation groups through a median split. For Logic: $HM_{Logic}(N = 248)$ and $LM_{Logic}(N = 247)$, and for Probability: HM_{Prob} $(N = 249)$ and $LM_{Prob}(N = 246)$. A chi-square test showed no significant evidence on students staying in the same motivation level from the logic tutor to the probability one: $\chi^2(1, N = 495) = 1.26, p = .26$. In other words, students' motivation levels may change over a course of a semester or change due to the subjective domains. Additionally, our motivation definition differs from students' incoming competence in that one-way ANOVA showed no significant difference in the pretest scores between the high- and low-motivation students: $F(1, 493) = 0.7, p = .17$ for Logic and $F(1, 493) = 0.001, p = .98$ for Probability.

Results

We will examine the impact of 1) metacognitive skills alone, 2) motivation alone, and 3) the interactions of the two on students' learning across both tutors. For each tutor, students' learning performance is measured using their corresponding pre- and post-test scores, together with their normalized learning gain (NLG) defined as: ($NLG = \frac{Post - Pre}{\sqrt{100 - Pre}}$) (Zhou et al., 2019), where 100 is the maximum posttest score. For reporting convenience, we normalize the pre and post scores to the range of [0, 100].

Table 1. Comparing the three Metacognitive Groups in each tutor

		Logic Tutor			Probability Tutor		
Group	Size	Pre	Post	NLG	Pre	Post	NLG
Str_Time	145	78.4 (3.2)	75.8 (1.7)	0.94 (.395)	72.3 (2.8)	75.5 (3)	0.02 (.06)
Str_Only	166	74.9 (3)	68.2 (1.67)	-0.46 (.39)	72.1 (2.5)	74 (2.8)	0.01 (.05)
Default	184	75.5 (2.8)	70.9 (1.68)	0.19 (.393)	71.8 (2.6)	73.4 (2.6)	-0.007 (.05)

Metacognitive Skills

Table 1 above compares the three metacognitive groups' learning performances on the logic and probability tutors. It shows the mean and SD of the pretest scores (Pre), the posttest scores (Post), and the NLGs. For the logic tutor, while we found no significant difference among the three groups on Pre, a one-way ANCOVA analysis with the metacognitive group as a factor and the pretest score as a covariate showed a significant difference in their posttest scores: $F(2, 491) = 17.3, p < .001, \eta = 0.3$. Subsequent contrast analyses showed that Str_Time scored significantly higher than both Str_Only: $t(309) = 5.8, p < .0001, d = 4.5$and Default: $t(327) = 3.8, p < .001, d = 2.9$. Additionally, Default scored significantly higher than Str_Only: $t(348) = 2.2, p = .03, d = 1.6$. For NLG, while a one-way ANOVA showed no significant difference among the three groups on the logic NLG, subsequent contrast analyses showed that Str_Time scored significantly higher than Str_Only: $t(309) = 2.4, p = .02, d = 3.6$. For the probability tutor, however, no significant results were found among the three metacognitive groups on either Pre, Post, or NLG.

To summarize, these results suggest that strategy-awareness alone cannot lead students to learn better in logic; students need to be time-aware as well. Additionally, while Str_Time group learned significantly better than Str_Only and Default in logic, this was not observed in probability. For Str_Only students, it seems they were negatively affected by their lack of time-awareness, given the aforementioned note that the posttest is much harder than the pretest.

Motivation Level

Table 2 compares the learning performance of the high- and low-motivation groups on the logic and probability tutors. As mentioned before, no significant difference was found between the high- and low-motivation groups on the pretest on either tutor. As expected, a one-way ANCOVA analysis using motivation as a factor and pretest as a covariate showed that on both tutors, high-motivation students scored significantly higher than their low peers on the corresponding posttest: $F(1, 492) = 15.8, p < .001, \eta =$

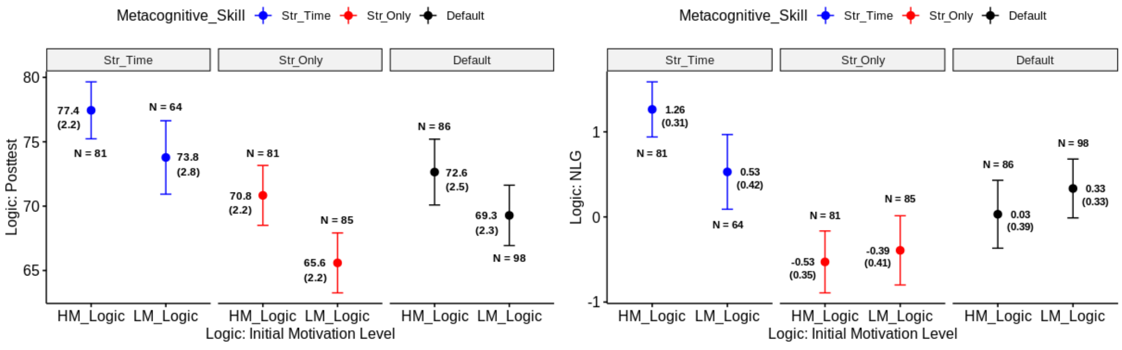
0.17 for logic and $F(1,492) = 24.5, p < .001, \eta = 0.17$ for probability. For the NLG, while no significant difference was found between the two groups' logic NLG, one-way ANOVA showed that highly motivated students had significantly higher probability NLG than their low peers: $F(1,493) = 7.6, p < .01, \eta = 0.12$. In short, this suggests that our motivation measure is reasonable in that: the highly motivated students indeed significantly outperformed their low-motivated peers on the posttest on both the logic and probability tutors. They also had significantly higher NLG on the probability tutor.

Table 2. Comparing the Motivation Level in each tutor

		Logic Tutor		
Group	Size	Pre	Post	NLG
HM_{Logic}	248	78.9 (5.3)	73.6 (1.4)	0.25 (.06)
LM_{Logic}	247	73.4 (5.5)	69.2 (1.4)	0.14 (.07)
		Probability Tutor		
Group	Size	Pre	Post	NLG
HM_{Prob}	249	81.7 (4.2)	79 (1.8)	0.05 (.04)
LM_{Prob}	246	77 (4.4)	69 (2.5)	-0.03 (.04)

Interaction Between Metacognition and Motivation

Logic Tutor



(a) Logic Posttest

(b) Logic NLG

Figure 3. Logic Performance: Metacognition and Motivation

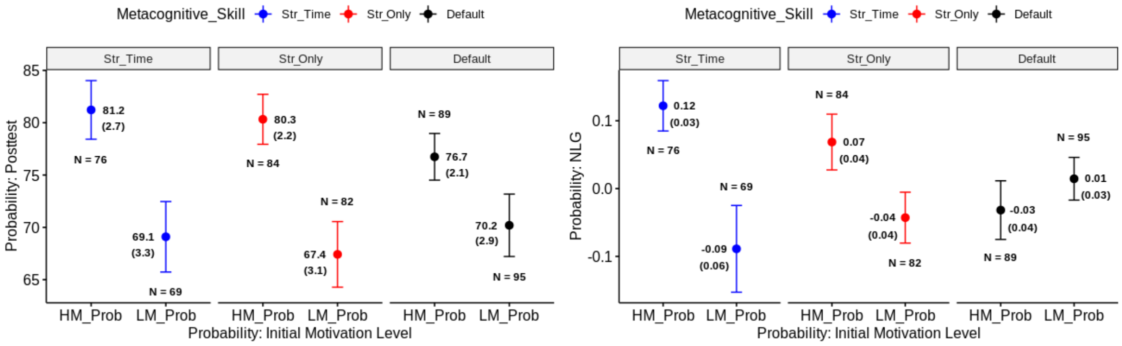
Combining the three metacognitive groups (Str_Time, Str_Only, and Default) with the two motivation levels (HM_{Logic} and LM_{Logic}) resulted in six groups. A chi-square test showed no significant difference in the distribution of motivation level across the metacognitive groups: $\chi^2(2, N = 495) = 2.87, p = .24$.

Additionally, no significant difference was found among the six groups on the logic pretest: $F(2, 489) = 0.69, p = .49$.

Figure 3 compares the six groups' performance on the logic tutor. Regarding the logic posttest (**Fig. 3a**), a two-way ANCOVA using the metacognitive groups and motivation levels as factors and pretest as a covariate showed no significant interaction effect. However, there was a main effect of metacognitive groups: $F(2, 488) = 16.6, p < .0001$, and a main effect of motivation level: $F(1, 488) = 16.7$, $p < .0001$. More specifically, within each metacognitive group, HM_{Logic} significantly outperformed the corresponding LM_{Logic}group: $t(143) = 2, p = .04, d = 1.4$ for Str_Time, $t(164) = 3.1, p < .01, d = 2.4$ for Str_Only and $t(182) = 2.1, p = .03, d = 1.4$ for Default. Among the three HM_{Logic}groups, the high-motivation Str_Time students scored significantly higher than their peers: $t(160) = 3.8, p < .001, d = 3$ against the high-motivation Str_Only peers and $t(165) = 2.8, p < .01, d = 2.1$ against the high-motivation Default ones.

For the NLG (**Fig. 3b**), a two-way ANOVA using the same two factors found no significant interaction effect nor any main effect. However, among the three HM_{Logic}groups, the high-motivation Str_Time scored significantly higher than both high-motivation Str_Only: $t(160) = 2.3, p = .03, d = 5.4$ and high-motivation Default: $t(165) = 2.2, p = .03, d = 3.5$. No significant difference was found among the three LM_{Logic}groups. Additionally, only within the two Str_Time groups, the high-motivation students scored significantly higher than their low-motivation peers. No significant difference between the high- and low-motivation groups was found within Default and Str_Only. In short, our results suggest that the high-motivation Str_Time group performs the best among the six groups in terms of both posttest and NLG scores on the logic tutor.

Probability Tutor



(a) Probability Posttest (b) Probability NLG

Figure 4. Probability Performance: Metacognition and Motivation

Similarly, we combined the three metacognitive groups with the two motivation levels defined based on students' interactions on the probability tutor: $HM_{Prob}$and LM_{Prob}, resulting in six groups. A chi-square test showed students' motivation level on the probability tutor did not differ significantly across the three metacognitive groups: $\chi^2(2, N = 495) = 0.53, p = .76$. Moreover, we found no significant difference between the six groups on the probability pretest: $F(2, 489) = 0.5, p = .63$.

Figure 4 depicts the performance of the six groups on the probability tutor. For the posttest (**Fig. 4a**), a two-way ANCOVA using metacognitive skills and motivation as factors and pretest scores as a covariate showed a significant interaction effect: $F(2, 488) = 3.8, p = .02, \eta = 0.09$. Additionally, there was a main effect of motivation in that high-motivation students scored significantly higher than their low peers: $F(1, 488) = 24.4, p < .0001$. Among the three highly motivated groups, both Str_Time and Str_Only scored significantly higher than Default: $t(163) = 2.4, p = .02, d = 1.9$ and $t(171) = 2.4, p = .02, d = 1.7$, respectively. However, no such difference was found among the three low-motivation groups.

Similarly, as shown in **Figure 4b**, a two-way ANOVA using metacognitive skills and motivation as factors showed a significant interaction effect on the probability NLG: $F(2, 489) = 6.4, p < .01, \eta = 0.16$ and there was also a main effect of motivation: $F(1, 489) = 7.8, p < .01$. Subsequent contrast analyses showed that high-motivation Str_Time students scored significantly higher than their low peers: $t(143) = 3.8, p < .001, d = 4.4$. The same pattern was observed between the two Str_Only groups: $t(164) = 2.2, p = .03, d = 2.9$. Across the three high-motivation groups, both Str_Time and Str_Only scored significantly higher than their Default peers: $t(163) = 3, p < .01, d = 4.2$ and $t(171) = 2, p = .04, d = 2.5$, respectively. In short, on our probability tutor, the high-motivation Str_Time group performs the best among the six groups, on both posttest scores and NLGs.

Conclusions and Discussions

In this chapter, we investigated how two factors, metacognitive skills and motivation, would impact student learning across two domains: logic and then probability. Our results from analyzing 495 students' performance on two tutors show that when considering each factor alone, no consistent robust pattern is found. However, when we combine the two factors, we find that students who are highly motivated, strategy-aware, and time-aware consistently outperform their peers across both domains.

Firstly, and most importantly, our analyses confirm the importance of motivation in that across both tutors, the impacts of metacognitive skills on student learning are only observed among the highly motivated student groups. For low motivated students, no significant difference was found among the three metacognitive groups in either tutor. In other words, our results reveal an aptitude-treatment interaction (ATI) effect (Kanfer & Ackerman, 1989) in that some students may be insensitive to learning unless the presented material matches their aptitude. While such findings are not surprising, they suggest that it is crucial to further understand why certain students lack motivation, and to explore how to motivate them. Moreover, our findings indicate that our choice of using students' online accuracy traces on the first two questions is a reasonable way to measure their motivation levels.

Secondly, while problem-solving strategies have been extensively explored in prior research, as far as we know this is the first work that investigates students' metacognitive skills from both strategy-aware and time-aware aspects. Our results suggest that these two skills are indeed different in that while both Str_Time and Str_Only groups know about problem-solving strategies, only the former knows when to apply them. More importantly, it is essential to consider the time-aware aspect when assessing students' metacognitive skills in that when highly motivated, Str_Time consistently outperforms their Str_Only and Default peers on both tutors.

Thirdly, our results show that Str_Only can benefit greatly by training on an ITS that explicitly teaches and follows problem-solving strategies. While the high-motivation Str_Only performed worse than their high-motivation peers on the logic tutor, they performed as well as the high-motivation Str_Time and both outperformed their Default peers on the probability tutor. One potential explanation is that the time-aware aspect of the skills is not needed when training on the probability tutor, since it follows the same explicit problem-solving strategy on all problems.

Finally, we emphasize the importance of mastering different problem-solving strategies for highly motivated students, and its role on PFL. We found that only across the highly motivated groups, both Str_Only and Str_Time had significantly higher probability scores than the Default group. This finding suggests evidence for metacognitive skill transfer for highly motivated students who are also aware of switching strategies. To sum up, while time awareness could be a decisive factor for consistency, strategy awareness might identify students who are prepared for future learning.

Despite these findings, it is important to note that there are at least two caveats in our analyses. First, we measured students' motivation using the first two problems on each tutor and we did not consider that students' motivation levels may vary during the training. Also, the probability tutor supports only one problem-solving strategy. A more convincing testbed would be to use any ITS that supports different types of strategies, so we can investigate whether students can properly use them.

Recommendations and Future Research

This chapter reinforces the significance of understanding how and when to apply each problem-solving strategy. This is consistent with prior findings that within multi-strategy domains, it is insufficient to only learn what each strategy is. Rather, it is equally important to learn when to use each. Therefore, the Generalized Intelligent Framework for Tutoring (GIFT) can benefit from this by prioritizing the importance of mastering **how** and **when** to use each strategy. For example, GIFT can ensure that individuals understand the methodology, context, timing, and reasoning beyond a given strategy.

For future work, we will investigate whether explicitly teaching different problem-solving strategies would boost the performance of students who lack strategy-awareness or time-awareness so that they can catch up with their peers. Additionally, we will explore different ways of motivating students and meeting their expectations about the content and interface of the tutors.

References

Barnes, T., Stamper, J. C., Lehmann, L., & Croy, M. J. (2008). A pilot study on logic proof tutoring using hints generated from historical student data. In *Educational Data Mining* (pp. 197-201).

Belenky, D. M., & Nokes, T. J. (2009). Examining the role of manipulatives and metacognition on engagement, learning, and transfer. *The Journal of Problem Solving*, *2*(2), 102-129.

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences*, *21*(3), 399-432.

Belenky, D. M., & Nokes-Malach, T. J. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. *Learning and individual differences*, *25*, 21-34.

Boden, K., Kuo, E., Nokes-Malach, T., Wallace, T., & Menekse, M. (2018). What is the role of motivation in procedural and conceptual physics learning? An examination of self-efficacy and achievement goals. In *Proceedings of the 2017 Physics Education Research Conference*, Cincinnati, OH (p. 60).

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, *24*(1), 61-100.

Chambres, P. E., Izaute, M. E., & Marescaux, P. J. E. (2002). *Metacognition: Process, function and use*. Kluwer Academic Publishers.

Chi, M., & VanLehn, K. (2007). The impact of explicit strategy instruction on problem-solving behaviors across intelligent tutoring systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29, No. 29).

Chi, M., & VanLehn, K. (2010). Meta-cognitive strategy instruction in intelligent tutoring systems: how, when, and why. *Journal of Educational Technology & Society*, *13*(1), 25-39.

Eccles, J. (1983). Expectancies, values and academic behaviors. *Achievement and achievement motives*.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational psychologist*, *46*(1), 6-25.

Kalender, Z. Y., Marshman, E., Schunn, C. D., Nokes-Malach, T. J., & Singh, C. (2019). Gendered patterns in the construction of physics identity from motivational factors. *Physical Review Physics Education Research*, *15*(2), 020119.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of applied psychology*, *74*(4), 657.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

Nokes, T. J., & Belenky, D. M. (2011). Incorporating motivation into a theoretical framework for knowledge transfer. In *Psychology of learning and motivation* (Vol. 55, pp. 109-135). Academic Press.

Otieno, C., Schwonke, R., Salden, R., & Renkl, A. (2013). Can Help Seeking Behavior in Intelligent Tutoring Systems Be Used as Online Measure for Goal Orientation?. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35, No. 35).

Priest, A. G., & Lindsay, R. O. (1992). New light on novice—expert differences in physics problem solving. *British journal of Psychology*, *83*(3), 389-405.

Rheinberg, F., Vollmeyer, R., & Rollett, W. (2000). Motivation and action in self-regulated learning. In *Handbook of self-regulation* (pp. 503-529). Academic Press.

Roberts, M. J., & Erdos, G. (1993). Strategy selection and metacognition. *Educational Psychology*, *13*(3-4), 259-266.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems.

Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass*, *8*(7), 328-341.

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, *18*(3), 239-253.

Winne, P. H., & Azevedo, R. (2014). Metacognition. In *The cambridge handbook of the learning sciences* (pp. 63-87).

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology*, *107*(4), 954.

Zepeda, C. D., Hlutkowsky, C. O., Partika, A. C., & Nokes-Malach, T. J. (2019). Identifying teachers' supports of metacognition through classroom talk and its relation to growth in conceptual learning. *Journal of Educational Psychology*, *111*(3), 522.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2019). Hierarchical reinforcement learning for pedagogical policy induction. In *International conference on artificial intelligence in education* (pp. 544-556). Springer, Cham.

Chapter 16 – Maintaining Chains of Evidence with xAPI

Florian Tolk
ADL Initiative

Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) is a very powerful platform for training, with multiple tools to improve both the learner's experience as well as enhancing the performance of the teacher. Within this chapter, we focus on the capability of GIFT to leverage the Experience Application Program Interface (xAPI; Advanced Distributed Learning Initiative, 2013) specification. The ADL Initiative is currently working towards building a learning services ecosystem that leverages xAPI in a manner detailed in this chapter.

In addition to the xAPI specification, this chapter borrows heavily from the cmi5 (cmi5 Working Group, 2021) specification, which adds additional rules and guidelines when using xAPI. This chapter covers how to best leverage xAPI not only to track meaningful data, but to also provide chains of evidence that connect earned competencies back to the exercises that were used to earn them.

Methods

The xAPI specification developed by IEEE details how to track and store information on a learner's interaction with educational experiences through the use of JavaScript Object Notation (JSON; Ecma International, 2017). This spec breaks interactions down into 6 major sections:

1. Actor – The entity that is the subject of the statement (Required)
2. Verb – The interaction that is being tracked (Required)
3. Object – The object that is being interacted with (Required)
4. Authority – The reporter of this event (Optional)
5. Result – The outcome of the tracked event (Optional)
6. Context – Any additional required information about the event (Optional)

Breaking down all learning events into these six properties allows any information needed to be tracked in a consistent, yet flexible structure. While this flexibility is a major strength of the xAPI spec, it allows the use of proprietary terms and vocabulary to describe the content, which can cause inconsistency in how data is tracked. This challenge is addressed by xAPI Profile Specification, which allows the definition of specific vocabulary for a particular set of xAPI Statements (e.g., already established definitions supported by the community of interest). Of particular interest is that this capability also supports drawing out templates for full xAPI statements and expected chain of events. While all three sections of an xAPI Profile are important, this chapter focuses on the use of templates, patterns, and the context field to maintain chains of evidence for learning in an xAPI enabled ecosystem, embedded in GIFT.

The xAPI specification can be broken down into 3 things: the JavaScript Object Notation (JSON) specification, sending/receiving xAPI statements, and the storage of these statements. First of all, xAPI is an extension of the JSON specification. This is a way to display data in such a way that both a machine and a human can understand it. xAPI takes this specification and further narrows how the data can be displayed, breaking each JSON statement down into roughly 3 sections: Actor, Verb, and Object; making each xAPI statement read like a simple sentence. For example: “Learner A (Actor) completed (Verb) Course 7.”

Once an xAPI statement is generated, it is sent to a Learner Record Store (LRS). To do this, an LRS essentially provides a URL that systems can then send these xAPI statements to through a secure HTTP request, or that systems can pull statements from via a secure HTTP request. Security requirements are included in the xAPI specification, such as the requirement for some form of authorization being included in all requests.

Finally, once the statement has been sent to the LRS, it has to verify that this JSON statement complies with the xAPI specification. If it complies with the specification, the statement is then stored in the LRS, and it sends the system that sent the statement a confirmation that this statement is now stored within the LRS and can be retrieved by other systems. Once a statement is saved within an LRS, it may not be deleted, and may not be altered any more, providing a reliable piece evidence that learning has occurred.

The first step to having traceable chains of evidence in an xAPI data pool is to determine in what order events will happen. This, as well as a lot of additional metadata, can be defined by using an xAPI profile. When using an xAPI profile, the expected order of events is called a pattern. Patterns can also be labeled as primary patterns if they must be checked for matching sequences of statements. This label marks specific flows as ones that are preferably, though not exclusively, followed.

It is possible to send and receive statements in an order that has not been defined in an xAPI profile, but it is best to add these sequences as non-primary patterns to an xAPI profile. These patterns will not restrict how the xAPI spec is used but will help data analysts generate better views of the data once the profile has been implemented within the learning environment.

Once all of the patterns have been established, it is time to allow statements to be linked together. Natively xAPI does allow three ways to link xAPI statements together:

1. setting the statement’s object as a SubStatement,
2. setting the statement’s object as a Statement Reference (StatementRef), and
3. using the registration property.

SubStatements and Statement References function in nearly the same way, and directly point to another xAPI statement. A SubStatement is a fully defined xAPI statement that is nested inside of another, and a Statement Reference is the statement id of another already existing xAPI statement. This allows previous xAPI statements to be directly targeted by a statement. A situation where this might become useful would be when one user sends an assertion of competence. This user’s assertion would then be processed either by a system, or another trusted user to approve or deny the previous statement:

1. **Statement 1:** Learner1 *Asserted* Competency 3
2. **Statement 2:** Admin3 *Approved* **Statement 1**

This works very well when one event is being evaluated, but not as well when there is a long series of related events occurring one right after the other, like during an exercise where all of a learner’s actions are tracked in xAPI. At this point, the registration property and patterns become the best alternative.

When the registration property is used, it is important to track when an xAPI statement is generated using the timestamp property. In addition to correctly timestamping xAPI statements, it is important to define the patterns in which these statements are expected to occur within an xAPI profile. This allows the future consumers of these statements to quickly track the connected xAPI statements, as well as quickly spot any irregularities within the collected data.

Once patterns have been properly defined it is possible to associate multiple statements with each other through the use of the registration property. The xAPI specification applies the concept of registration more broadly than most Learning Management Systems (LMSs) and can be considered to be an attempt, a session, or span multiple activities. Normally the same registration is used for requests to both the Statement and State Resources relating to the same learning experience so that all data recorded for the experience is consistent.

A sample of the registration properly being used would run as follows:

1. **Statement1:** Learner2 *Launched* Lesson1
    1. Registration Code: ec531277-b57b-4c15-8d91-d292c5b2b8f7
2. **Statement2:** Learner2 *Completed* Lesson1
    1. Registration Code: ec531277-b57b-4c15-8d91-d292c5b2b8f7
3. **Statement3:** Learner2 *Passed* Lesson1
    1. Registration Code: ec531277-b57b-4c15-8d91-d292c5b2b8f7

When a learning record consumer wishes to track these connected statements, they can now pull all xAPI statements with this specific registration code and use the timestamps and the patterns in the profile to order them as they occurred. This is the approach currently used by the cmi5 specification to link xAPI statements and maintain chains of evidence.

In order to add more clarity to statement associations, it is recommended to also leverage the context activities property. Many Statements do not just involve one (Object) Activity that is the focus but relate to other contextually relevant Activities. Context activities allow these relations to be expressed within an xAPI statement. There are four categories that these relations can be expressed as: Parent context activities, Grouping context activities, Categories context activities, and Other context activities.

- The Parent context activity is an activity with a direct relation to the activity which is the object of the statement. In almost all cases there is only one sensible parent or none, not multiple, but it is not forbidden. For example: A statement about a quiz question would have the quiz as its parent Activity.
- The Grouping context activity is an activity with an indirect relation to the activity which is the object of the statement. For example: a course that is part of a qualification. The course has several classes. The course relates to a class as the parent, the qualification relates to the class as the grouping.
- The Category context activity is an activity used to categorize or “tag” the statement. Category should be used to indicate a profile of xAPI behaviors, as well as other categorizations. For example: When the learner attempts a biology exam, and the statement is tracked using the cmi5 profile. The statement's object refers to the exam, and the category is the cmi5 profile.
- Finally, the Other context activity acts as a catch-all for any other activities that may be related to a statement and not covered by the other three categories. When using this context activity to track relations, it is important to clearly define how these context activities are related to the statement in an xAPI profile.

The last step to fully defining a chain of evidence is to define the competency assertion and provide a link to the end of this evidentiary chain. The ADL Initiative does this with a context extension, called the evidence extension, to their assertion statements (Advanced Distributed Learning Initiative, 2020). This context extension is not included in the xAPI specification but leverages the rules of the spec to extend the capabilities of xAPI. This extension is an array of pointers to "completed" xAPI statements.

The following example demonstrates all the principles discussed in this chapter. Using the previously listed tools, the average chain of evidence for an assertion would trace back like this:

1. **Assertion:** Learner 1 *Asserted* Competency2
    1. Evidence: [Passed3]
    2. Category: ADLProfile
    3. Timestamp: 1031 June 06, 2020
2. **Passed3:** Learner1 *Passed* Course1
    1. Category: ADLProfile
    2. Timestamp: 1030 June 06, 2020
3. **Completed5:** Learner1 *Completed* test2
    1. Registration Code: ec531277-b57b-4c15-8d91-d292c5b2b8f7
    2. Grouping: Course1
    3. Category: ADLProfile
    4. Timestamp: 1015 June 06, 2020
4. **Launched5:** Learner1 *Launched* test2
    1. Registration Code: ec531277-b57b-4c15-8d91-d292c5b2b8f7
    2. Grouping: Course1
    3. Timestamp: 0900 June 06, 2020
5. **Completed3:** Learner1 *Completed* exercise1
    1. Registration Code: 54792e5c-7496-4a60-b818-5737fa38e071
    2. Grouping: Course1
    3. Category: ADLProfile
    4. Timestamp: 1100 June 05, 2020
6. **Launched3:** Learner1 *Launched* exercise1
    1. Registration Code: 54792e5c-7496-4a60-b818-5737fa38e071
    2. Grouping: Course1
    3. Category: ADLProfile
    4. Timestamp: 1030 June 05, 2020
7. **Completed1:** Learner1 *Completed* test1
    1. Registration Code: bd4047f4-eda3-4316-a967-48799b2c49a3
    2. Grouping: Course1
    3. Category: ADLProfile
    4. Timestamp: 1530 June 01, 2020
8. **Launched1:** Learner1 *Launched* test1
    1. Registration Code: bd4047f4-eda3-4316-a967-48799b2c49a3
    2. Grouping: Course1
    3. Category: ADLProfile
    4. Timestamp: 1400 June 01, 2020

Figure 1. *A Sample flow of statements as a user interacts with a MOM focuses ecosystem*

Results

When the assertion has to be reviewed, the course completion event can be pulled from the LRS. Then how the learner interacted with the course can be pulled searching for statements about this Learner1 with a Grouping of 1. The timestamp for each of these events is included, and can be used to order them, and the evidence can then be traced back to the activity level. In addition to this, the use of a properly defined xAPI profile makes "Launched" and the other verbs controlled terms with very strict definitions. Launched can mean two different things normally, such as saying "I launched the activity" versus "I launched the rocket." By defining the verb "Launched" in a profile, "Launched" can only mean launching an activity.

The proper use of unique registration codes and the grouping property for each xAPI statement also adds additional context to this evidentiary chain. The grouping property clearly defines that statements 3-8 of Figure 1 are all related to a learner attempting to complete the same course, Course 1, linking them is statement 2. The unique registration course then breaks these statements down into clear learning sessions, where individual tests and assignments are completed over the course of a few days.
By properly leveraging all of these properties within an xAPI statement, an assertion of competence can now be traced back to the individual activities that a learner had to complete to generate that assertion.

Final Discussion

Because GIFT supports the use of the xAPI specification, it can now leverage these practices and become more interoperable with other learning systems. Additionally, these logs can assist GIFT with its intelligence as now courses that do not train as effectively can now be found by following the evidence quoted in assertions for poorly performing learners. The same can be done to find the most effective courses by tracing back the chains of evidence of overperforming learners.

In conclusion, by following these best practices when leveraging the xAPI specification, GIFT can better integrate with a larger learning ecosystem, audit both learners and the available learning resources, and link a learner's competencies and badges all the way back to their performance during their training.

References

Advanced Distributed Learning Initiative. (2013). xAPI-Spec. Retrieved version 1.0.3 on April 11, 2021, from https://github.com/adlnet/xAPI-Spec.

cmi5 Working Group. (2021). cmi5 Spec. Retrieved Quartz Edition on April 30, 2021. From https://github.com/AICC/CMI-5_Spec_Current/blob/quartz/cmi5_spec.md

Ecma International. (2017). JSON spec. Retrieved version 1 on April 30, 2021, from https://www.iso.org/obp/ui/#iso:std:iso-iec:21778:ed-1:v1:en

Advanced Distributed Learning Initiative. (2020) Master Object Model. Retrieved version 0.15 on April 30, 2021, from https://github.com/adlnet/MasterObjectModel/blob/master/MOM_Spec.md

BIOGRAPHIES

Editors

Dr. Anne M. Sinatra is a Research Psychologist in the Learning in Intelligent Tutoring Environments (LITE) Lab within the U.S. Army DEVCOM Soldier Center Simulation and Training Technology Center. The focus of her research is in cognitive psychology, human factors psychology, and adaptive team tutoring. She has specific interest in how information relating to the self and about those that one is familiar with can aid in memory, recall, and tutoring. Her dissertation research evaluated the impact of using degraded speech and a familiar story on attention/recall in a dichotic listening task. Her post-doctoral work examined the self-reference effect and personalization in the context of computer-based tutoring. Her work has been published in journals including the Computers in Human Behavior, Journal of Artificial Intelligence in Education, and Interaction Studies. Her work has also been published in conference proceedings including the Human Factors and Ergonomics Society conference, and the Human Computer Interaction International conference. She additionally has served as an editor on five books (she was lead editor on three of them), and chaired three team tutoring workshops during the Artificial Intelligence in Education conferences in 2018, 2019, and 2021). Dr. Sinatra received her Ph.D. and M.A. in Applied Experimental and Human Factors Psychology, as well as her B.S. in Psychology from the University of Central Florida.

Dr. Arthur C. Graesser is a 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 include 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. Dr. Hu has received funding for the above research from the US National Science Foundation (NSF), US Institute of Education Sciences (IES), ADL of the US Department of Defense (DoD), US Army Medical Research Acquisition Activity (USAMRAA), US Army Research Laboratories (ARL), US Office of Naval Research (ONR), UofM, and CCNU.

Dr. Benjamin Goldberg is a member of the Army Futures Command - Combat Capabilities Development Command Simulation and Training Technology Center in Orlando, FL. He has been conducting research in the Modeling & Simulation community for the past eight years with a focus on adaptive learning in simulation-based environments and how to leverage Artificial Intelligence tools and methods to create personalized learning experiences. Currently, he is the LITE Lab's lead scientist on instructional management research within adaptive training environments and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is a Ph.D. graduate from the University of Central Florida in the program of Modeling & Simulation. His work has been published across several well-known conferences, with recent contributions to the Human Factors and Ergonomics Society (HFES), Artificial Intelligence in Education and Intelligent Tutoring Systems (ITS) proceedings. Dr. Goldberg has also recently contributed to the journal Computers in Human Behavior and to the Journal of Cognitive Technology.

Dr. Andrew J. Hampton is a Research Scientist Assistant Professor at the Institute for Intelligent Systems & Department of Psychology, within the University of Memphis. He serves as project manager on the pioneering hybrid tutor ElectronixTutor and development leader on a conversational AI meant to aid in career planning through education and qualification tracking, intelligent recommendation, and mitigation of personal issues. He is also editing a book on ethics in artificial intelligence specifically from a psychological perspective. Research interests include technologically mediated communication, psycholinguistics, semiotics, adaptive educational technology, artificial intelligence, and political psychology.

Dr. Joan H. Johnston is a Senior Research Psychologist with the U.S. Army Combat Capabilities Development Command, Soldier Center. She began her military research career in 1990 with the U.S. Navy. For her work on the Tactical Decision Making Under Stress program she was awarded the Office of Naval Research Dr. Arthur E. Bisson Prize for Naval Technology Achievement and the Society for Industrial and Organizational Psychology M. Scott Myers Award for Applied Research in the Workplace, and was made a NAVAIR Fellow. In 2012, she became the Orlando Unit Chief of the Army Research Institute. Then in 2014 she joined the Army Research Laboratory, Human Research Engineering Directorate, and was awarded the US Army Civilian Service

Achievement Medal for an innovative team training strategy to improve decision making under stress in dismounted Army squads. Dr. Johnston received her M.A. and Ph.D. in Industrial and Organizational Psychology from the University of South Florida.

Authors

Mark Abdelshiheed is a Ph.D. student in the Department of Computer Science at North Carolina State University. His research is on transfer learning in intelligent tutoring systems, with an emphasis on the metacognitive skills of different learners. His long-term research goal is to leverage his research to improve teaching skills and students' learning outcomes.

Dr. Eva Baker is a Distinguished Professor Emerita at UCLA. Dr. Baker researches design and validation of multipurpose training and assessments systems, recently focusing on games, simulations, and scenario-based assessments (workforce skills) for the US Navy and PBS (early learning). Her AI studies include benchmarking as well as evaluations of ITSs, games, interventions, and applying AI to assessment. She has served as Chair of the Board on Testing and Assessment, National Research Council, Co-Chair of the "Standards for Educational and Psychological Testing," and been President of both the American and World Educational Research Associations. Dr. Baker has evaluated national tests and reform in the US and abroad. A member of the National Academy of Education and a fellow in scholarly association, she has numerous awards in measurement and is widely published.

Dr. Tiffany Barnes is Professor of Computer Science at NC State University. She received the B.S. and M.S. degrees in Computer Science and Mathematics, and the Ph.D. degree in Computer Science from N.C. State. Dr. Barnes has served as chair or program chair for many conferences, including ACM SIGCSE, Educational Data Mining RESPECT, STARS Celebration, and Foundations of Digital Games. Tiffany Barnes has recently served on the ACM Special Interest Group on Computer Science Education Board (2010-2016), the Board of Directors for the International Educational Data Mining Society (2011-present), Chair of IEEE Computer STC Broadening Participation, and Associate Editor for IEEE Transactions on Learning Technologies (2016-present). Dr. Barnes received an NSF CAREER Award for her novel work in using data and educational data mining to add intelligence to STEM learning environments.

Dr. Min Chi is an Associate Professor in the Department of Computer Science at NC State University. She joined the department in August 2013 as a Chancellor's Faculty Excellence Program cluster hire in the Digital Transformation of Education. She has established a foundational R&D portfolio with impactful advancements across four major lines of research, including Reinforcement Learning (RL)-based policy induction. She has served as the PI and Co-PI for a series of federally funded grants from NSF, NIH, and DOE and has led multidisciplinary collaborations. She has received numerous awards for her research expertise and impact, including an NSF CAREER Award, an Alcoa Foundation Engineering Research Achievement Award, and a series of Best Paper, Best Student Paper, and Outstanding Paper Awards.

Dr. Kilchan Choi is Associate Director for Statistics and Methodology at CRESST. Dr. Choi has rich expertise in developing and applying advanced statistical methodologies in education assessment, psychometric modeling with large-scale gameplay data, multisite evaluation, growth modeling, value-added models, and school effectiveness and accountability research. He brings expertise and experience from his many years of leading IES statistical methodology projects to develop the proposed new statistical models, estimation methods, and statistical programs that are part of this effort. As CRESST's principal methodologist, he is responsible for research design, instrument validation, statistical modeling, and technical quality control for nearly all CRESST projects. He has most recently been leading the Navy Training Assessment Framework effort as Project Director. In that project, he has developed a novel psychometric approach to validate assessments under a small sample case, and incorporate qualitative information from subject matter experts into the scoring process. Furthermore, he currently leads the development of psychometric and statistical approaches to analysis of very large-scale gameplay process data to estimate game players' performance level, play patterns, and thinking processes.

Jody L. Cockroft, AA, BS, CCRP is a Research Specialist at the University of Memphis (UoM) in the Psychology Department with the Institute for Intelligent Systems where she has been for the past five-plus years. She has over thirty years of experience in scientific research and has worked on both the bench and on various clinical studies. She has been an integral part of the Army Research Laboratory Cooperative agreement, the Advanced Distributed Learning Initiative ADL-A project, the Learner Data Institute (LDI), and the Advanced Learning Theories, Technologies and Impacts (ALTTAI) Consortium for the past several years while at the UoM. She has authored or co-authored over twenty articles in peer-reviewed journals as well as countless abstracts and posters. She is the treasurer of the Adaptive Instructional Systems (AIS) Working Group IEEE P2247.1. Her research interests include the standardization of adaptive instructional systems and improving human learning.

Dr. Jonathan Gratch is a Research Full Professor of Computer Science. Psychology and Media Arts and Practice at the University of Southern California (USC) and Director for Virtual Human Research at USC's Institute for Creative Technologies. He completed his Ph.D. in Computer Science at the University of Illinois in Urbana-Champaign in 1995. Dr. Gratch's research focuses on computational models of human cognitive and social processes, especially emotion, and explores these models' role advancing psychological theory and in shaping human-machine interaction. He is the founding Editor-in-Chief (retired) of IEEE's *Transactions on Affective Computing*, founding Associate Editor of *Affective Science,* Associate Editor of *Emotion Review* and the *Journal of Autonomous Agents and Multiagent Systems,* and former President of the Association for the Advancement of Affective Computing (AAAC). He is a Fellow of AAAI, AAAC, and the Cognitive Science Society.

Dr. Sara Haviland is a Research Scientist in the Center for Education and Career Development at Educational Testing Service in Princeton, NJ. She studies the social, policy, organizational, and individual factors that affect work and careers, with a focus on educational interventions and training to improve career trajectories. She also examines issues of educational access and training for adult learners, and the policy implications of workforce development programs. Dr. Haviland has served as a research evaluator for workforce development and community college programs, and publishes and presents regularly on these topics. Her research at ETS has focused on career and technical education, improving career pathways in community colleges, and building student success through soft skills training. Dr. Haviland holds an M.A. and Ph.D. in sociology from The University of North Carolina at Chapel Hill, and a B.A. in social theory and ethics from Oglethorpe University.

Emmanuel Johnson is a PhD candidate in Computer Science at the University of Southern California advised by Jonathan Gratch. His research focuses on improving negotiation training by using AI to provide personalized feedback. He holds a MS in Robotics from the University of Birmingham, and a BS in Computer Engineering from North Carolina Agricultural and Technical State University.

Mike Kalaf has over 30 years of Modeling, Simulation and Training leading large scale efforts leveraging cutting edge technology. Mike has worked in the commercial and military aviation, training and simulation business. In his most recent efforts, he has been leading new opportunities applying front end modeling, simulation and analysis. Mike has led several programs integrating "state of the art" technology and delivering highly successful technology and business innovation. Mike has been collaborating with educational organizations and exploring conceptual frameworks, platforms and business models to transform our current system and elevate the performance and quality. He was involved with the University of Central Florida's College of Education on a unique system of teacher training via classroom simulators. These projects fit well to advance science, technology, engineering and mathematics learning to lay the groundwork for a new generation of engineers and scientists. Mike volunteers his time to numerous education organizations including serving as a board member for the Central Florida STEM council and the Seminole County Public Schools Foundation. Mike's formal education includes an earned Mechanical Engineering degree from Rochester Institute of Technology, RIT.

Dr. Patrick Kyllonen is Distinguished Presidential Appointee in the R&D Division of Educational Testing Service in Princeton, NJ. Dr. Kyllonen received a B.A. from St. Johns University, Ph.D. from Stanford University, and authored Generating Items for Cognitive Tests (with S. Irvine,2001); Learning and Individual Differences (with P. L. Ackerman &amp; R.D. Roberts, 1999); Extending Intelligence: Enhancement and New Constructs (with R. Roberts and L. Stankov, 2008); and Innovative Assessment of Collaboration (with A. von Davier and M. Zhu, 2017). He is a fellow of American Psychological Association and American Educational Research Association and has coauthored several National Academy of Sciences reports, Education for Life and Work: Developing Transferable Knowledge and Skills in the 21 st Century (2012), Measuring Human Capabilities (2015), and Supporting Students' College Success: The Role of Assessment of Intrapersonal and Interpersonal Competencies (2017). Dr. Kyllonen is a recipient of The Technical Cooperation Program Achievement Award for the "design, development, and evaluation of the Trait-Self Description (TSD) Personality Inventory." Dr. Kyllonen directed the Center for New Constructs (later, Center for Academic and Workforce Readiness and Success) at ETS for 15 years. The Center focused on identifying and measuring new constructs for applications in K-12, higher education, and the workforce. While directing the center Dr. Kyllonen also led NAEP questionnaire work, developing white papers and 4th, 8th, and 12th grade background questionnaires for mathematics, English language arts, science, social science, and the National Indian Education Study, and Socioeconomic Status study. He also led PISA 2012 questionnaire development, introducing many new item types to PISA including anchoring vignettes, situational judgment tests, and forced-choice approaches, and was an expert group advisor on OECD's recently completed The Survey on Social and Emotional Skills Study.

Mehak Maniktala is pursuing a Ph.D. in Computer Science at North Carolina State University. Her research focuses on applying machine learning and user experience to solve the assistance dilemma in an intelligent logic tutor. Her contributions include a new principled approach to determine when students need help in open-ended learning environments, and a study showcasing how this approach can save students time and improve their performance.

Dr. Harry O'Neil is a Professor of Educational Psychology and Technology at the University of Southern California's Rossier School of Education. His current research interests include the effectiveness of computer games and simulations, team training and assessment, and teaching and assessment of 21st century skills. He has recently co-edited the following books: *Teaching and Measuring Cognitive Readiness (AKA 21st Century Skills)* (2014), *Using Games and Simulations for Teaching and Assessment: Key Issues* (2016), *Theoretical Issues of Using Simulations and Games in Educational Assessment* (in press, Routledge/Taylor & Francis), and *Using Cognitive and Affective Metrics in Educational Simulations and Games: Applications in School and Workplace Contexts* (in press, Routledge/Taylor & Francis). He is a Fellow of the American Psychological Association (APA), the American Educational Research Association (AERA), and the Association for Psychological Sciences (APS). In each of these organizations, less than five percent of the membership has fellow status.

Kevin Owens is an Engineering Scientist at the Applied Research Laboratories, The University of Texas at Austin – a US Navy University Affiliated Research Center (UARC). After retiring from the US Navy, Mr. Owens earned a BS and MS in Instructional Design and has 19-years' experience in DoD learning engineering science and technology (S&T) and DoD operational research. Mr. Owens supports US Army PEO-STRI and the Synthetic Training Environment (STE) requirements process, the Squad Immersive Virtual Trainer (SIVT) and Integrated Visual Augmentation System (IVAS) engineering team; STE Experiential Learning for Readiness (STEELR) S&T applied research; and US Navy learning engineering applied research.

Debbie Patton has over 30 years of experience conducting military research. She is currently a senior research psychologist at the US Army Combat Capabilities Development Command Analysis Center where she leads research efforts on understanding human performance and training effectiveness as well as conducting human systems integration evaluations. Ms. Patton leads a group on the Army Human Behavior Representation working group, is a contributor to developing methods and metrics to the Army's Maximizing Human Performance effort, and a moderator for the DAU Human Systems Community of Practice. Prior to joining the Data and Analysis Center, Ms. Patton served 27 years of service under the Human Research and Engineering Directorate at the Army Research Laboratory leading research in soldier stress and performance in both live and simulated military training. Here she developed methods and metrics to measure stress and realism in military live and simulated environments which lead her to train national and international scientists on how to measure subjective and objective stress in military environments. She was an invited researcher to the Engineer Scientist Exchange Program and worked in Defense Science Technology, Australia for one year. She chairs sessions at international conferences and serves on international conference panels for several years. She served three years as the lead for the human stress and performance subgroup for the DoD Human Factors and Engineering Technical Advisory Group. Ms. Patton has 30+ publications. Ms. Patton received her master's degree in experimental psychology from Town University, Towson, MD.

Dr. Steven B. Robbins is Principal Research Scientist at Educational Testing Service in Princeton, NJ. Prior to ETS, Dr. Robbins was Vice President for Research at ACT. He also is a former professor and chair of the Psychology Department at Virginia Commonwealth University. He was a James Scholar at the University of Illinois where he received his B.A. in Psychology. He received his Ph.D. in an APA-accredited counseling psychology program at the University of Utah. He was elected Fellow of the American Psychological Association in 1992, and received the Division 17 early career scientist-practitioner award. Dr. Robbins is a leading social scientist in his field, publishing more than 145 refereed articles and technical reports, and has conducted workshops and presentations around the world. He is a leading student education success expert, co-authoring *Increasing Persistence: Research-based Strategies for College Student Success* (Wiley, 2012), with Wesley R. Habley and Jennifer Bloom. His research draws upon a psychological perspective on human and social capital when understanding education and work success. He promotes evidence-based assessment and intervention practices that help underserved learner's bridge education and work.

Elliot Robson is the General Manager and previous Director of Research for Eduworks Corporation. He has served as Principal Investigator on the Personalized eBooks for Learning (PeBL) project funded by the US Advanced Distributed Learning (ADL) and is currently Co-PI on the NSF Competency Catalyst project, as well as on the Competency and Skills System (CaSS) project. He has worked as a leader in education technology transformation since 2005 with organizations including the NYC Department of Education, UNESCO, and Amplify Learning. He has served as chair for IEEE standards on intelligent eBooks and is an active member of the Open Skills Network's technical working groups. He has numerous publications and presentations including on methods for competency computation and application.

Dr. Robby Robson received his Ph.D. from Stanford in 1981, joined the mathematics faculty at Oregon State University in 1984, was an Alexander von Humboldt fellow, and received tenure in 1989. In 1995 he co-created one of the first online learning systems and moved to industry as Saba Software's "standards evangelist" in 2000, where became director of Product Management. In 2001 he co-founded Eduworks Corporation, a company that applies AI to education, training, and workforce development, where he is CEO. Since 2000, Robby has actively participated in IEEE standards as a Standards Committee chair and member of the IEEE Standards Association Standards Board and Board of Governors. Most recently, he has been instrumental in launching IEEE-SA OPEN. Robby has over 100 publications in diverse areas of research and is currently Principal Investigator on two significant projects that are applying AI and competency-based approaches to talent management and experiential learning.

Dr. Vasile Rus is a Full Professor of Computer Science at The University of Memphis and the Lead Principal Investigator of the newly NSF-funded Learner Data Institute to lay the foundations of a Data Science Institute for learner data (www.learnerdatainstitute.org). Dr. Rus' research interests lie at the intersection of artificial intelligence, human and machine learning, and natural language processing with an emphasis on developing interactive intelligent systems such as intelligent tutoring systems and care-bots (healthcare bots). Dr. Rus has served in various roles on research projects funded by the National Science Foundation, Department of Defense, Department of Education, and private companies. Many of those projects involved the development of intelligent tutoring systems and medium-size (10-25 people) interdisciplinary teams. For instance, Dr. Rus has led the development of the DeepTutor system (www.deeptutor.org), a project funded by the Department of Education and is currently leading the development of an NSF-funded project to develop an intelligent tutoring system for source code comprehension, called DeepCode. Dr. Rus produced more than 150 peer-reviewed publications and received 6 Best Paper Award nominations of which 3 were Best Paper Awards. His team won the first two Question Answering competitions organized by the National Institute for Science and Technology (NIST) and recently his team won the English Semantic Similarity challenge organized by the leading forum on semantic evaluations – SemEval. Among other accomplishments, Dr. Rus was named Systems Testing Research Fellow of the FedEx Institute of Technology for his pioneering work in the area of software systems testing and is a member of the PI Millionaire club at The University of Memphis for his successful efforts to attract multi-million funds from federal agencies as Principal Investigator (PI).

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 a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT), an open source, AI-based adaptive instructional architecture. GIFT has over 2000 users in 76 countries. Dr. Sottilare has 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, 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 adjunct professor at the University of Central Florida where he teaches a graduate level course in ITS design. Dr. Sottilare has also been a frequent lecturer at the United States Military Academy (USMA) where he taught a senior level colloquium on adaptive training and ITS design. He has a long history of participation in international scientific fora including NATO and the Technical Cooperation Program. He has over 200 technical publications in the learning sciences field with over 1500 citations in the last 5 years. His doctorate is in Modeling & Simulation with a focus in Intelligent Systems from the University of Central Florida. Dr. Sottilare is a recipient of the US Army Meritorious Service Award (2018; 2nd highest civilian award), the US Army Achievement Medal for Civilian Service (2008; 5th highest civilian award), and two lifetime achievement awards in Modeling & Simulation: US Army RDECOM (2012; inaugural recipient) and National Training & Simulation Association (2015). He was also recognized by NTSA in 2019 for his contributions to adaptive instruction and the design and development of GIFT.

Florian Tolk is a SETA contractor and a Software Engineer with the Advanced Distributed Learning (ADL) Initiative. In this role, he provides technical knowledge in working groups for the development of technical specifications and standards, as well as designing internal systems for Training and Education research experiments that implement these standards. Prior to working for the ADL Initiative, he worked as a software developer for SimIS inc. and developed in a team environment training simulations such as the Automated Intelligent Mentoring System (AIMS).

Dr. Kevin M. Williams is a Research Scientist in the Center for Educational and Career Development at Educational Testing Service. He received his Ph.D. in Personality Psychology with a minor in Quantitative Psychology from The University of British Columbia in 2008. Dr. Williams' research includes investigations into the predictive validity and malleability of noncognitive constructs (e.g., personality) in workplace contexts, validity research for high school equivalency tests, psychometric evaluations of novel psychological assessments, law enforcement personnel selection, response bias in job performance evaluations, career technical education (CTE) pathways, educational and career experiences of underrepresented groups, and identifying skills for the new economy through employer expectations. Prior to joining ETS, Dr. Williams' professional experience involved the development and validation of psychological and licensure assessments for stakeholders in educational, workplace, clinical, law enforcement, and correctional fields. Dr. Williams has presented at numerous international conferences and published several influential articles, whose citations number in the thousands.

Design Recommendations for Intelligent Tutoring Systems

Volume 9
Competency-Based Scenario Design

Design Recommendations for Intelligent Tutoring Systems (ITSs) explores the impact of intelligent tutoring system design on education and training. Specifically, this volume examines "Competency-Based Scenario Design". The Design Recommendations book series examines tools and methods to reduce the time and skill required to develop Intelligent Tutoring Systems with the goal of improving the Generalized Intelligent Framework for Tutoring (GIFT). GIFT is a modular, service-oriented architecture developed to capture simplified authoring techniques, promote reuse and standardization of ITSs along with automated instructional techniques and effectiveness evaluation capabilities for adaptive tutoring tools and methods.

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 and works on the Generalized Intelligent Framework for Tutoring (GIFT).*

- ***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 visiting professor at Central China Normal University.*

- ***Dr. Benjamin Goldberg*** *is a senior researcher at U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center and is a co-creator of GIFT.*

- ***Dr. Andrew J. Hampton*** *is a research assistant professor in the Department of Psychology at the University of Memphis.*

- ***Dr. Joan H. Johnston*** *is a senior research psychologist at U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center*

A Volume in the Adaptive Tutoring Series

9 780997 725810